

Preserving Privacy and Security in Federated Learning

Truc Nguyen[✉] and My T. Thai[✉], *Fellow, IEEE*

Abstract—Federated learning is known to be vulnerable to both security and privacy issues. Existing research has focused either on preventing poisoning attacks from users or on concealing the local model updates from the server, but not both. However, integrating these two lines of research remains a crucial challenge since they often conflict with one another with respect to the threat model. In this work, we develop a principle framework that offers both privacy guarantees for users and detection against poisoning attacks from them. With a new threat model that includes both an honest-but-curious server and malicious users, we first propose a secure aggregation protocol using homomorphic encryption for the server to combine local model updates in a private manner. Then, a zero-knowledge proof protocol is leveraged to shift the task of detecting attacks in the local models from the server to the users. The key observation here is that the server no longer needs access to the local models for attack detection. Therefore, our framework enables the central server to identify poisoned model updates without violating the privacy guarantees of secure aggregation.

Index Terms—Federated learning, zero-knowledge proof, homomorphic encryption, model poisoning.

I. INTRODUCTION

FEDERATED learning is an engaging framework for large-scale distributed training of deep learning models with thousands to millions of users. In every round, a central server distributes the current global model to a random subset of users. Each of the users trains locally and submits a model update to the server. Then, the server averages the updates into a new global model. Federated learning has inspired many applications in many domains, especially training image classifiers and next-word predictors on users' smartphones [19]. To exploit a wide range of training data while maintaining users' privacy, federated learning by design has no visibility into users' local data and training.

Despite the great potential of federated learning in large-scale distributed training with thousands to millions of users, the current system is still vulnerable to certain privacy and security risks. First, although the training data of each user

Manuscript received 1 April 2022; revised 13 October 2022 and 22 May 2023; accepted 3 July 2023; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor Y. Chen. Date of publication 25 August 2023; date of current version 16 February 2024. This work was supported by the National Science Foundation under Grant CNS-1935923 and Grant CNS-2140477. (Corresponding author: My T. Thai.)

The authors are with the Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: truc.nguyen@ufl.edu; mythai@cise.ufl.edu).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TNET.2023.3302016>, provided by the authors.

Digital Object Identifier 10.1109/TNET.2023.3302016

is not disclosed to the server, the model update is. This poses a privacy threat, as it is suggested that a trained neural network's parameters enable a reconstruction and/or inference of the original training data [1], [16], [29]. Second, federated learning is generally vulnerable to model poisoning attacks that leverage the fact that federated learning enables adversarial users to directly influence the global model, thereby allowing considerably more powerful attacks [3]. Recent research on federated learning focuses on either improving the privacy of model updates [2], [8], [39] or preventing certain poisoning attacks from malicious users [3], [37], [38], but not both.

To preserve user privacy, secure aggregation protocols have been introduced into federated learning to devise a training framework that protects the local model updates [2], [8], [39]. These protocols enable the server to privately combine the local models in order to update the global model without learning any information about each individual local model. Specifically, the server can compute the sum of the parameters of the local models while not having access to the local models themselves. As a result, the local model updates are concealed from the server, thereby preventing the server from exploiting the updates of any user to infer their private training data.

On the other hand, previous studies also unveil a vulnerability of federated learning when it comes to certain adversarial attacks from malicious users, especially poisoning attacks [3], [37]. This vulnerability leverages the fact that federated learning gives users the freedom to train their local models. Specifically, any user can train in any way that benefits the attack, such as arbitrarily modifying the weights of their local models. To prevent such attacks, a conventional approach is to have the central server run some defense mechanisms to inspect each of the model updates, such as using anomaly detection algorithms to filter out the poisoned ones [3].

However, combining these two lines of research on privacy and security is not trivial since they contradict one another. In particular, by allowing the server to inspect local model updates from users to filter out the attacked ones, it violates the security model of secure aggregation. In fact, under a secure aggregation setting, the server should not be able to learn any information about individual model updates. Therefore, with secure aggregation, the server cannot run any defense mechanism directly on each model update to deduce whether it is attack-free or not. For that reason, existing FL systems with secure aggregation prevent the server from detecting poisoned local models.

To tackle this issue, in this paper, we propose a framework that integrates secure aggregation with defense mechanisms against poisoning attacks from users. To circumvent the aforementioned problem, we shift the task of running the defense mechanism onto the users, and then each user is given the ability to attest the execution of the defense mechanism with respect to their model update. Simply speaking, the users are the ones who run the defense mechanism, then they have to prove to the server that the mechanism was executed correctly. To achieve this, the framework leverages a zero-knowledge proof (ZKP) protocol that the users can use to prove the correctness of the execution of a defense mechanism. The ZKP protocol must also make it difficult to generate a valid proof if the defense mechanism was not run correctly, and the proof must not reveal any information about the local model update in order to retain the privacy guarantee of secure aggregation. Via our ZKP protocol, the server can verify if a defense mechanism was properly executed by the users.

To demonstrate the feasibility of the framework, we use the backdoor attack as a use case in which we construct a ZKP protocol for a specific backdoor defense mechanism [25]. Furthermore, we also propose a secure aggregation protocol in federated learning to address the limitations of previous work. Specifically, we show that our proposed aggregation protocol is robust against malicious users in a way that it can still maintain privacy and liveness despite the fact that some users may not follow the protocol honestly. Our framework can then combine both the ZKP and the secure aggregation protocols to tackle the above-mentioned privacy and security risks in federated learning.

A. Contributions

Our main contributions are as follows:

- We establish a framework integrating both the secure aggregation protocol and defense mechanisms against poisoning attacks without violating any privacy guarantees.
- We propose a new secure aggregation protocol using homomorphic encryption for federated learning that can tolerate malicious users and a semi-honest server while maintaining both privacy and liveness.
- We construct a ZKP protocol for a backdoor defense to demonstrate the feasibility of our framework.
- Finally, we analyze the computation and communication cost of the framework and provide some benchmarks regarding its performance.

B. Organization

The rest of the paper is structured as follows. Section II provides the system and security models. The main framework for combining both the secure aggregation protocol and defense mechanisms against poisoning attacks is shown in Section III. Section IV presents our proposed secure aggregation protocol for federated learning. Section V gives the ZKP protocol for a backdoor defense mechanism. We evaluate the complexity and benchmark our solution in Section VI. We discuss some related work in Section VII and provide some concluding remarks in Section VIII.

II. SYSTEM AND SECURITY MODEL

Current systems proposed to address either the poisoning attacks from users or the privacy of model updates are designed using security models that conflict with each other. In fact, when preserving the privacy of model updates, the central server and the users are all considered as honest-but-curious (or semi-honest). However, when defending against poisoning attacks from users, the server is considered honest while the users are malicious.

Due to that conflict, this section establishes a more general security model combining the ones proposed in previous work on the security and privacy of federated learning. From the security model, we also define some design goals.

A. System Model

We define two entities in our protocol:

- 1) A set of n users \mathcal{U} : Each user has a local training dataset, computes a model update based on this dataset, and collaborates with each other to securely aggregate the model updates.
- 2) A single server S : Receives the aggregated model updates from the users and computes a global model that can later be downloaded by the users.

We denote the model update of each user $u \in \mathcal{U}$ as an input vector x_u of dimension m . For simplicity, all values in both x_u and $\sum_{u \in \mathcal{U}} x_u$ are assumed to be integers in the range $[0, N]$ for some known N . The protocol is correct if S can learn $\bar{x} = \sum_{u \in \mathcal{U}'} x_u$ for some subset of users $\mathcal{U}' \subseteq \mathcal{U}$. We assume a federated learning process that follows the FedAvg framework [27] which uses a horizontal setting for FL. We denote \mathcal{U}' as a random subset of \mathcal{U} (the server selects a random subset of users for each training round).

In addition to securely computing \bar{x} , each user $u \in \mathcal{U}$ computes a zero-knowledge proof π_u proving that x_u is a clean model and not leaking any information about x_u . This proof π_u is then sent to and validated by the server S . The server then verifies that $\{\pi_u\}_{u \in \mathcal{U}}$ is consistent with the obtained \bar{x} .

B. Security Model

1) *Threat Model*: We take into account a computationally bounded adversary that can corrupt the server or a subset of users in the following manners. The server is considered honest-but-curious in a way that it behaves honestly according to the training protocol of federated learning but also attempts to learn as much as possible from the data received from the users. The server can also collude with corrupted users in \mathcal{U} to learn the inputs of the remaining users. Moreover, some users try to inject unclean model updates to poison the global model and they can also arbitrarily deviate from the protocol.

As regards the security of the key generation and distribution processes, there has been extensive research in constructing multi-party computation (MPC) protocols where a public key is computed collectively and every user computes private keys [11], [13], [30]. Such a protocol can be used in this architecture to securely generate necessary keys for the users and the server. Hence, this paper does not focus

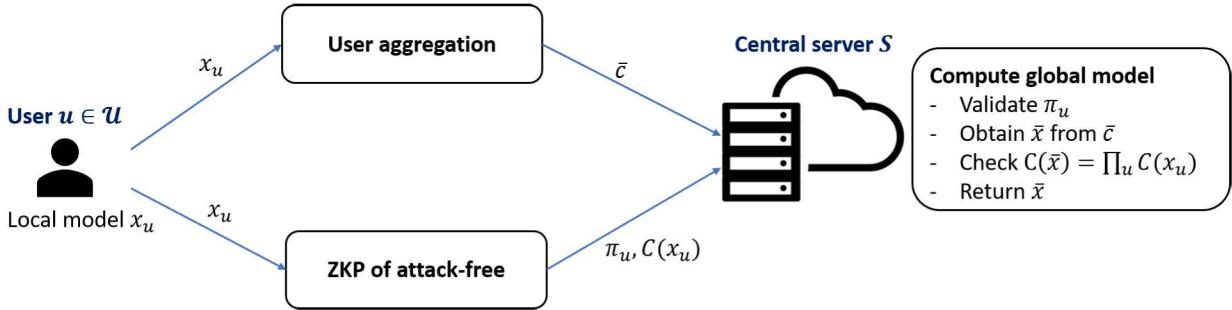


Fig. 1. Secure framework for FL. Each user $u \in \mathcal{U}$ trains a local model x_u that is used as an input to (1) **User aggregation** and (2) **ZKP of attack-free model**. The **User aggregation** component returns $\bar{c} = E_{pk}(\sum_{u \in \mathcal{U}'} x_u)$ which is the encryption of the sum over the local models of honest users. The **ZKP of attack-free model** component returns the proof π_u and the commitment $C(x_u)$ for each x_u . The outputs of these two component are then used by the central server as inputs to the **Computing global model** component. This component validates $\{\pi_u\}_{u \in \mathcal{U}}$, obtains the global model \bar{x} from \bar{c} , checks if $C(\bar{x}) = \prod_u C(x_u)$ is consistent with $\{C(x_u)\}_{u \in \mathcal{U}}$, and then returns \bar{x} .

on securing the key generation and distribution processes but instead assumes that the keys are created securely and honest users' keys are not leaked to others.

2) *Design Goals*: Besides the security aspects, our design also aims to address unique challenges in federated learning, including operating on high-dimensional vectors and users dropping out. With respect to the threat model, we summarize the design goals as follows:

- 1) The server learns nothing except what can be inferred from \bar{x} .
- 2) Each user does not learn anything about the others' inputs.
- 3) The server can verify that \bar{x} is the sum of clean models.
- 4) The protocol operates efficiently on high-dimensional vectors
- 5) Robust to users dropping out.
- 6) Maintain undisrupted service (i.e., liveness) in case a small subset of users deviate from the protocol.

III. SECURE FRAMEWORK FOR FEDERATED LEARNING

To address the new security model proposed in Section II, we propose a framework to integrate secure aggregation with defenses against poisoning attacks in a way that they can maintain their respective security and privacy properties. Our framework is designed to ensure the privacy of users' local model while preventing certain attacks from users. As shown in Fig. 1, our framework consists of three components: (1) users aggregation, (2) zero-knowledge proof of attack-free model, and (3) computing global model. We discuss the properties of each component as follows.

A. Users Aggregation

This component securely computes the sum of local models from the users. First, each user $u \in \mathcal{U}$ trains a local model update x_u and uses it as input to this component. Then, the component outputs $\bar{c} = E_{pk}(\sum_{u \in \mathcal{U}'} x_u)$ which is an encryption over the sum of the model updates of all honest users $\mathcal{U}' \subseteq \mathcal{U}$. $E_{pk}(\cdot)$ denotes a viable additive homomorphic encryption scheme, and we require that it must have the property of indistinguishability under chosen plaintext attack (IND-CPA). The encryption function will be discussed in detail in Section IV.

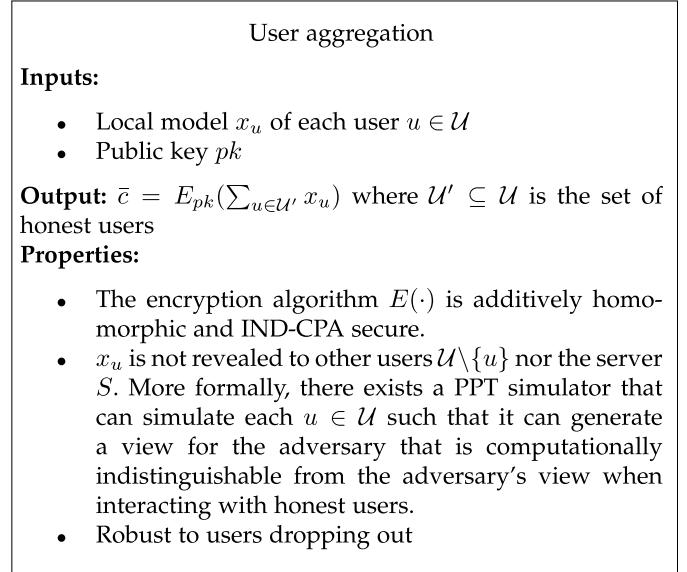


Fig. 2. User aggregation component.

Furthermore, this aggregation must maintain the privacy with respect to each local model x_u , particularly, we should be able to simulate the output of each user by a Probabilistic Polynomial Time (PPT) simulator. We specify the security properties of the Users aggregation component in Fig. 2.

B. Zero-Knowledge Proof of Attack-Free Model

This component generates a zero-knowledge proof proving the local model is free from a certain poisoning attack. In specific, we first assume the existence of a function that verifies an attack-free model as follows:

$$R(x) = \begin{cases} C(x) & \text{if } x \text{ is attack-free} \\ \perp & \text{otherwise} \end{cases} \quad (1)$$

where x is a local model, $C(x)$ is the commitment of x . Next, the component implements a zero-knowledge proof protocol for the users to attest the correct execution of $R(x)$ without revealing x . Note that the ZKP itself is not a defense mechanism, but it is a means to verify that the users have executed the function $R(\cdot)$ on their side. Hence, the effectiveness of

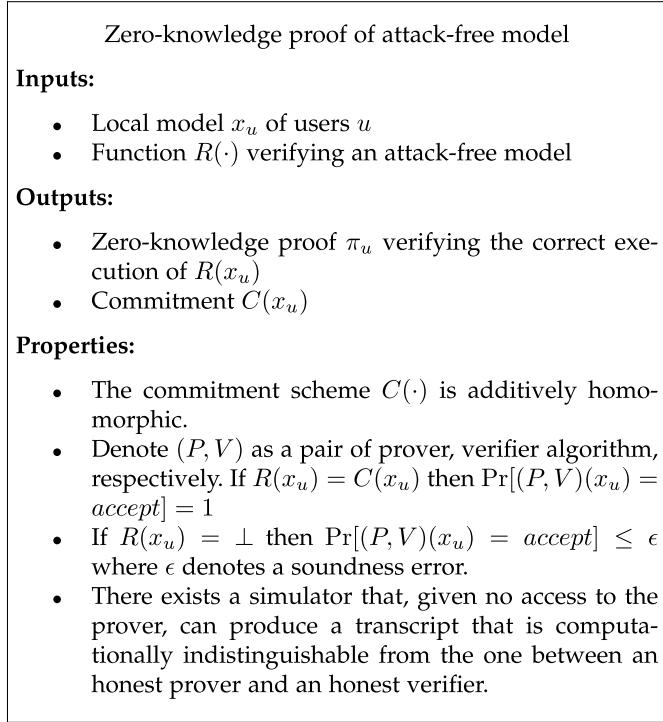


Fig. 3. Zero-knowledge proof component.

this component in preventing poisoning attacks depends on the performance of the defense $R(\cdot)$. A specific implementation of this component for defending against backdoor attacks is shown in Section V. The properties of this component are shown in Fig. 3.

The prover algorithm P is run by the users, while the verifier algorithm V is run by the server S .

C. Computing Global Model

This component takes as input the output of the other two components and produces the corresponding global model. Hence, it serves to combine two previous components and it is run solely by the server. Specifically, it needs to be able to decrypt \bar{c} from the first component and to validate the zero-knowledge proofs $\{\pi_u\}_{u \in \mathcal{U}'}$ from the second component. Then, it outputs the sum of the model updates from the honest users. We define this component in Fig. 4.

From the third component, we can prove that this framework preserves the privacy of users' model updates given that all the components satisfy their properties. Let Adv be a random variable representing the joint views of all parties that are corrupted by the adversary according to the threat model. We show that it is possible to simulate the combined view of any subset of honest-but-curious parties given only the inputs of those parties and the global model \bar{x} . This means that those parties do not learn anything other than their own inputs and \bar{x} .

Theorem 1: There exists a PPT simulator \mathcal{S} such that for $Adv \subseteq \mathcal{U} \cup S$, the output of the simulator \mathcal{S} is computationally indistinguishable from the output of Adv

Proof: Refer to Appendix A. \square

IV. SECURE AGGREGATION IN FEDERATED LEARNING

In this section, we describe a secure aggregation scheme that realizes the User aggregation component of our framework.

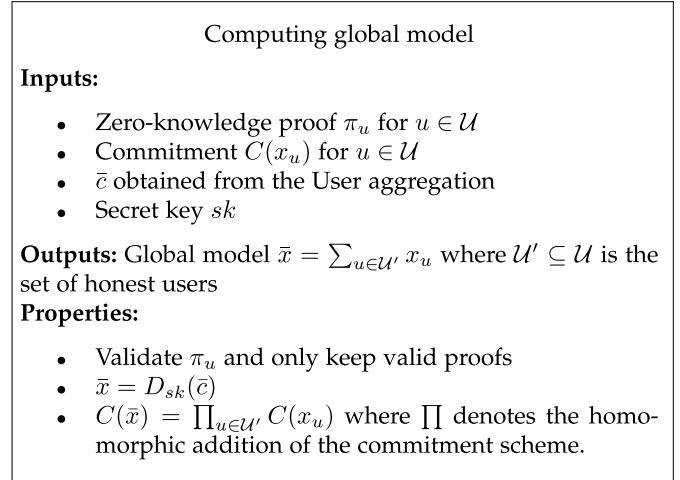


Fig. 4. Computing global model component.

Although some secure aggregation schemes have been proposed for federated learning, they are not secure against our threat model in Section II. Particularly, the aggregation schemes in [2] and [39] assume no collusion between the server and a subset of users. The work of Bonawitz et al. [8] fails to reconstruct the global model if one malicious user arbitrarily deviates from the secret sharing protocol. A detailed discussion of these related work is provided in Section VII.

The main challenge here is that we want to provide robustness in spite of malicious users, i.e., the protocol should be able to tolerate users that do not behave according to the protocol and produce correct results based on honest users. Moreover, we need to minimize the users' communication overhead when applying this scheme.

A. Cryptographic Primitives

1) *Damgård-Jurik and Paillier Cryptosystems* [10], [31]: Given $pk = (n, g)$ where n is the public modulus and $g \in \mathbb{Z}_{n^{s+1}}^*$ is the base, the Damgård-Jurik encryption of a message $m \in \mathbb{Z}_{n^s}$ is $\mathcal{E}_{pk}(m) = g^m r^{n^s} \bmod n^{s+1}$ for some random $r \in \mathbb{Z}_n^*$ and a positive natural number s . The Damgård-Jurik cryptosystem is additively homomorphic such that $\mathcal{E}_{pk}(m_1) \cdot \mathcal{E}_{pk}(m_2) = \mathcal{E}_{pk}(m_1 + m_2)$. This cryptosystem is semantically secure (i.e., IND-CPA) based on the assumption that computing n -th residue classes is believed to be computationally difficult. The Paillier's scheme is a special case of Damgård-Jurik with $s = 1$.

In the threshold variant of these cryptosystems [10], several shares of the secret key are distributed to some decryption parties, such that any subset of at least t of them can perform decryption efficiently, while less than t cannot decrypt the ciphertext. Given a ciphertext, each party can obtain a partial decryption. When t partial decryptions are available, the original plaintext can be obtained via a *Share combining* algorithm.

2) *Pedersen Commitment* [32]: A general commitment scheme includes an algorithm C that takes as inputs a secret message m and a random r , and outputs the commitment $commit$ to the message m , i.e., $commit = C(m, r)$.

A commitment scheme has two important properties: (1) *commit* does not reveal anything about m (i.e., hiding), and (2) it is infeasible to find $(m', r') \neq (m, r)$ such that $C(m', r') = C(m, r)$ (i.e., binding). Pedersen commitment [32] is a realization of such scheme and it is also additively homomorphic such that $C(m, r) \cdot C(m', r') = C(m + m', r + r')$

B. Secure Aggregation Protocol

Our secure aggregation scheme relies on additive homomorphic encryption schemes, that is $\mathcal{E}_{pk}(m_1 + m_2) = \mathcal{E}_{pk}(m_1) \cdot \mathcal{E}_{pk}(m_2)$. In this work, we use the Damgård-Jurik public-key cryptosystem [10] which is both IND-CPA secure and additively homomorphic. The cryptosystem supports threshold encryption and operates on the set \mathbb{Z}_n^* .

We also assume that the communication channel between each pair of users is secured with end-to-end encryption, such that an external observer cannot eavesdrop the data being transmitted between the users. The secure communication channel between any pair of users can be mediated by the server.

In the one-time setup phase, the secret shares $\{s_u\}_{u \in \mathcal{U}}$ of the private key sk are distributed among n users in \mathcal{U} via the threshold encryption scheme. The scheme requires a subset $\mathcal{U}_t \subset \mathcal{U}$ of at least t users to perform decryption with sk . This process of key distribution can be done via a trusted dealer or via an MPC protocol. After that, each secure aggregation process is divided into 2 phases: (1) Encryption and (2) Decryption.

1) *Encryption Phase*: To securely compute the sum $\bar{x} = \sum_{u \in \mathcal{U}} x_u$, each user $u \in \mathcal{U}$ first computes the encryption $c_u = \mathcal{E}_{pk}(x_u)$ and then send c_u to the server. The server then computes the product of all $\{c_u\}_{u \in \mathcal{U}}$, denoting as \bar{c} :

$$\begin{aligned} \bar{c} &= \prod_{u \in \mathcal{U}} c_u \pmod{n^{s+1}} = \prod_{u \in \mathcal{U}} \mathcal{E}_{pk}(x_u) \pmod{n^{s+1}} \\ &= \mathcal{E}_{pk} \left(\sum_{u \in \mathcal{U}} x_u \pmod{n^{s+1}} \right) \end{aligned} \quad (2)$$

Additionally, per IND-CPA, knowing the ciphertext c_u implies nothing about the plaintext x_u for $u \in \mathcal{U}$. As each user only sends out c_u , neither the server S nor other entities $\mathcal{U} \setminus \{u\}$ with access to c_u can infer any information about x_u . This satisfies the security properties that are specified in Fig. 2.

2) *Decryption Phase*: Suppose that up to this phase, some amount of less than $N - t$ users dropout, that means we still have the subset $\mathcal{U}_t \subset \mathcal{U}$ of at least t users remaining to perform decryption.

First, the server sends \bar{c} to all users in \mathcal{U}_t . Each user $u \in \mathcal{U}_t$ computes $c'_u = \bar{c}^{2s_u \Delta}$ where $\Delta = t!$, and then sends c'_u back to the server. The server then applies the *Share combining* algorithm in [10] to compute $\bar{x} = \sum_{u \in \mathcal{U}} x_u$ from $\{c'_u\}_{u \in \mathcal{U}_t}$. As a result, the server can securely obtain \bar{x} without knowing each individual x_u .

3) *User Drop-Outs*: Furthermore, by design, it can be seen from this protocol that we can tolerate up to $N - t$ users dropping out and the server is still able to construct the global

model of all users. In other words, as long as there are at least t users remaining, it will not affect the outcome of the protocol. Therefore, our protocol is robust to user drop-outs up to $N - t$.

4) *Security Against Active Adversary*: We analyze the security in the active adversary model. First, the protocol is secure in spite of collusion between the server and a subset of users in a way that allows the server to decrypt some c_u to obtain the corresponding x_u . By using the threshold encryption, as long as there is at least one honest user in \mathcal{U}_t , the server will not be able to learn x_u of any user. Hence, the protocol can tolerate up to $t - 1$ malicious users, with $t < |\mathcal{U}|$.

Second, in the encryption phase, if a user deviates from the protocol, e.g., by computing c_u as a random string or using a different key other than pk , the server will not be able to decrypt \bar{c} . To tolerate this behavior, the server can iteratively drop each c_u from \bar{c} until it obtains a decryptable value.

In the decryption phase, if a user u^* does not cooperate to provide c'_{u^*} , then the server will not have all the necessary information to perform decryption. However, if we still have at least t honest users (not counting u^*), we can eliminate u^* from \mathcal{U}_t and add another user to \mathcal{U}_t , hence, the server can still perform decryption.

In summary, our secure aggregation protocol can tolerate $t - 1$ malicious users and be robust against $|\mathcal{U}| - t$ users dropping out (where $t < |\mathcal{U}|$). Therefore, t should be determined by calibrating the trade-off between the usability (small t) and security (large t).

C. Efficient Encryption of x_u

We discuss two methods for reducing the ciphertext size when encrypting x_u . Since x_u is a vector of dimension m , naively encrypting every element of x_u using either the Paillier or Damgård-Jurik cryptosystems will result in a huge ciphertext with the expansion factor of $2 \log n$ (i.e., the ratio between the size of c_u and of x_u). Below, we present (1) a polynomial packing method and (2) a hybrid encryption method that includes a key encapsulation mechanism (KEM) and a data encapsulation mechanism (DEM).

1) *Polynomial Packing*: We can reduce the expansion factor by packing x_u into polynomials as follows. Suppose $x_u \equiv (x_u^{(1)}, x_u^{(2)}, \dots, x_u^{(m)})$, we transform this into:

$$x'_u = x_u^{(1)} + x_u^{(2)} \cdot 2^b + x_u^{(3)} \cdot 2^{2b} \dots + x_u^{(m)} \cdot 2^{(m-1)b} \quad (3)$$

where $b = \lceil \log N \rceil$. Hence, x'_u becomes a mb -bit number.

Then, to encrypt x_u , we simply run the Damgård-Jurik encryption on x'_u . Specifically, we obtain the ciphertext $c_u = \mathcal{E}(x'_u)$. This c_u will have $(\lceil \frac{mb}{\log n} \rceil + 1) \log n$ bits. The expansion factor can be calculated as:

$$\lim_{m \rightarrow \infty} \frac{(\lceil \frac{mb}{\log n} \rceil + 1) \log n}{mb} = 1 \quad (4)$$

Therefore, as m increases, i.e., larger models, the expansion factor approaches 1, thereby implying minimal overhead. Note that since we are packing into polynomials, the additive homomorphic property of the ciphertexts is retained.

Additionally, if we use this polynomial packing scheme with the Paillier encryption, it would result in an expansion factor

of 2 as Paillier requires the plaintext size to be smaller than the key size. To keep the expansion factor of 1 when using Paillier, we introduce a KEM-DEM scheme as follows.

2) *Hybrid Encryption (KEM-DEM Technique)*: In [9], the authors propose a KEM-DEM approach to compress the ciphertext of Fully Homomorphic Encryption (FHE). With FHE, they can use blockciphers to generate the keystreams. However, since our work only deals with Partially HE (PHE) like Paillier, we cannot use blockciphers, hence, we need to propose a different approach to generate the keystream with only PHE.

The general idea of our protocol is as follows. Suppose a user wants to send $E_{pk}(x)$ to the server (for simplicity, we use the notation x instead of x_u). Instead of sending $c = E_{pk}(x)$ directly, the user chooses a secret random symmetric key k that has the same size as an element in x and computes $c' = (E_{pk}(k), SE_k(x))$ where $SE_k(x) = x - k$, which means subtracting k from every element in x . c' is then sent to the server. When the server receives c' , it can obtain $c = E_{pk}(x)$ by computing $E_{pk}(k) \times E_{pk}(SE_k(x)) = E_{pk}(x) = c$.

The correctness of this scheme holds because the server can obtain $E_{pk}(x)$ from the user. The security also holds because the server does not know the key k chosen by the user, hence, it cannot know x . However, this scheme is not semantically secure, i.e., IND-CPA secure, because by defining $SE_k(x) = x - k$, identical elements in x will result in the same ciphertext.

To resolve this, we use an IV (initialization vector) for the encryption function. IV is a nonce which could be a random number or a counter that is increased each time it is used. Denote $x = (x_1, \dots, x_m)$, \oplus as a bitwise XOR operation, on the user's side:

- 1) Choose the secret l -bit random key k and a nonce IV.
- 2) compute the keystream ks : For $i \in [1, m]$, compute $ks_i = k \times (IV \oplus i) \bmod p$ where p is an l -bit number.
- 3) Compute $SE_k(x) = x - ks = (x_1 - ks_1, \dots, x_m - ks_m)$
- 4) Send $(IV, E_{pk}(k), SE_k(x))$ to the server

On the server's side, receiving $(IV, E_{pk}(k), SE_k(x))$:

- 1) Compute the encrypted keystream $E_{pk}(ks)$: For $i \in [1, m]$, compute $E_{pk}(k)^{IV \oplus i} \bmod n^2 = E_{pk}(k \times (IV \oplus i)) = E_{pk}(ks_i)$
- 2) Compute $E_{pk}(x) = E_{pk}(SE_k(x)) \cdot E_{pk}(ks) = E_{pk}(x_1 - ks_1 + ks_1, \dots, x_m - ks_m + ks_m) = E_{pk}(x_1, \dots, x_m)$

Lemma 1: By picking k uniformly at random, and p as an l -bit prime number, the DEM component (i.e., $SE_k(x)$) is IND-CPA secure.

Proof: To show that the DEM component is IND-CPA secure, it is sufficient prove that $k \times (IV \oplus i) \bmod p$ is uniformly random with respect to k (Theorem 3.32 in [22]). In fact, in \mathbb{Z}_p^* , by picking k uniformly at random, $k \times (IV \oplus i) \bmod p$ is uniformly distributed if and only if neither k nor $(IV \oplus i)$ shares prime factors with p . Hence, p should be an l -bit prime number and k can be chosen randomly from \mathbb{Z}_p^* . \square

Since the KEM component (i.e., $E_{pk}(k)$) is IND-CPA secure by default using Paillier, and the DEM component is

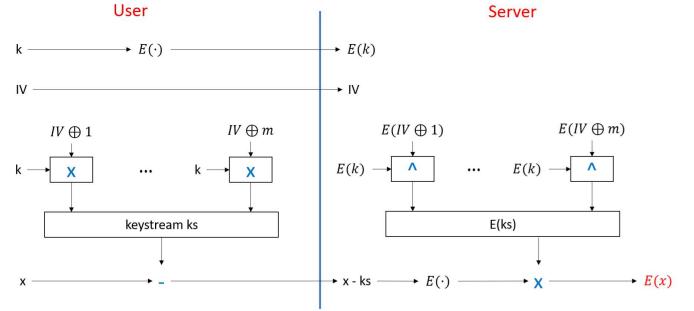


Fig. 5. KEM-DEM for PHE. \oplus denotes a bitwise XOR operation.

also IND-CPA secure as previously shown in Lemma 1, we can derive the following theorem:

Theorem 2: The KEM-DEM scheme is IND-CPA secure.

Figure 5 illustrates this scheme. The bandwidth expansion factor of this scheme can be computed as:

$$\lim_{m \rightarrow \infty} \frac{mb + 2 \log n + c}{mb} = 1 \quad (5)$$

where c is the size of the IV. Therefore, this scheme also achieves the expansion factor close to 1 as m grows, however, it incurs some computational overhead as compared to the polynomial packing scheme.

Although both the KEM-DEM scheme and polynomial packing have the same asymptotic expansion factor (e.g., 1), each has its own advantages and disadvantages. The polynomial packing is simple and works out-of-the-box if we use the Damgård-Jurik cryptosystem. This is because Damgård-Jurik allows the plaintext size to be greater than the key size. However, it will not work with conventional cryptosystems like Paillier as they require the plaintext size to be smaller than the key size. The KEM-DEM scheme would work for all cryptosystems but it is more complex and incurs more computational overhead than the polynomial packing scheme. Since our protocol is using the Damgård-Jurik cryptosystem, Section VI only evaluates the polynomial packing scheme.

V. USE CASE: ZKP FOR BACKDOOR DEFENSES

This section describes an implementation that realizes the Zero-knowledge proof component of our framework. Using the backdoor attack as a use case, we provide a ZKP protocol that can detect backdoor attacks from users. Before that, we establish some background regarding multi-party protocol (MPC) and ZKP.

A. Preliminaries

1) *Multi-Party Computation (MPC) Protocol:* In a generic MPC protocol, we have n players P_1, \dots, P_n , each player P_i holds a secret value x_i and wants to compute $y = f(x)$ with $x = (x_1, \dots, x_n)$ while keeping his input private. The players jointly run an n -party MPC protocol Π_f . The view of the player P_i , denoted by $View_{P_i}(x)$, is defined as the concatenation of the private input x_i and all the messages received by P_i during the execution of Π_f . Two views are *consistent* if the messages reported in $View_{P_i}(x)$ as incoming

from P_i are consistent with the outgoing message implied by $View_{P_j}(x)$ ($i \neq j$). Furthermore, the protocol Π_f is defined as perfectly correct if there are n functions $\Pi_{f,1}, \dots, \Pi_{f,n}$ such that $y = \Pi_{f,i}(View_{P_i}(x))$ for all $i \in [n]$. An MPC protocol is designed to be t -private where $1 \leq t < n$: the views of any t -subset of the parties do not reveal anything more about the private inputs of any other party.

2) *Zero-Knowledge Proof (ZKP)*: Suppose that we have a public function f , a secret input x and a public output y . By using a ZKP protocol, a prover wants to prove that it knows an x s.t. $f(x) = y$, without revealing what x is. A ZKP protocol must have three properties as follows: (1) if the statement $f(x) = y$ is correct, the probability that an honest verifier accepting the proof from an honest prover is 1 (i.e., completeness), (2) if the statement is incorrect, with a probability less than some small soundness error, an honest verify can accept the proof from a dishonest prover showing that the statement is correct (i.e., soundness), and (3) during the execution of the ZKP protocol, the verifier cannot learn anything other than the fact that the statement is correct (i.e., zero-knowledge).

3) *Backdoor Attacks and Defenses*: Backdoor attacks aim to make the target model to produce a specific attacker-chosen output label whenever it encounters a known trigger in the input. At the same time, the backdoored model behaves similarly to a clean model on normal inputs. In the context of federated learning, any user can replace the global model with another so that (1) the new model maintains a similar accuracy on the federated-learning task, and (2) the attacker has control over how the model behaves on a backdoor task specified by the attacker [3].

To defend against backdoor attacks, Liu et al. [25] propose a *pruning defense* that can disable backdoors by removing *backdoor neurons* that are dormant for normal inputs. The defense mechanism works in the following manner: for each iteration, the defender executes the backdoored DNN on samples from the testing dataset, and stores the activation of each neuron per test sample. The defender then computes the neurons' average activations over all samples. Next, the defender prunes the neuron having the lowest average activation and obtains the accuracy of the resulted DNN on the testing dataset. The defense process terminates when the accuracy drops below a predetermined threshold.

When implementing this pruning defense, to prune a neuron, we simply set the weights and biases associated to that neuron to zero. In this way, we can retain the structure of the local models x_u .

B. ZKP Protocol to Backdoor Detection

To inspect a local model for backdoors while keeping the model private, we propose a non-interactive ZKP such that each user $u \in \mathcal{U}$ can prove that x_u is clean without leaking any information about x_u .

First, we devise an algorithm *backdoor* that checks for backdoor in a given DNN model x as in Algorithm 1.

Theorem 3: $backdoor(x)$ returns 1 $\Rightarrow x$ is a non-backdoored model

Proof: Refer to Appendix B. \square

Algorithm 1 *backdoor*(x) – Check for Backdoors in a DNN Model

Input: DNN model x , validation dataset
Output: 1 if x is backdoored, 0 otherwise

- 1 Test x against the validation dataset, and record the average activation of each neuron;
- 2 $k \leftarrow$ the neuron that has the minimum average activation;
- 3 Prune k from x ;
- 4 Test x against the validation dataset;
- 5 **if** accuracy drops by a threshold τ **then**
- 6 **return** 1
- 7 **else**
- 8 **return** 0

Given a model x , we define $R(r, x)$ for some random r as follows:

$$R(r, x) = \begin{cases} C(r, x) & \text{if } backdoor(x) \text{ returns 1} \\ \perp & \text{otherwise} \end{cases} \quad (6)$$

where $C(r, x)$ is the commitment of x

Based on [21], we construct the ZKP protocol as follows. Each user $u \in \mathcal{U}$ generates a zero-knowledge proof π_u proving x_u is clean, i.e., $R(r_u, x_u) = C(r_u, x_u)$. To produce the proof, each user u runs Algorithm 2 to obtain π_u . Note that Π_f is a 3-party 2-private MPC protocol, which means a subset of any 2 views do not reveal anything about the input x_u . Then, π_u and $commit_u = C(r_u, x_u)$ are sent to the server where they will be validated as in Algorithm 3.

Finally, for the server to validate that the \bar{x} obtained from secure aggregation is the sum of the models that were validated by the ZKP protocol, it collects $\{r_u\}_{u \in \mathcal{U}}$ and verify that $\prod_{u \in \mathcal{U}} commit_u = C(\sum_{u \in \mathcal{U}} r_u, \bar{x})$. This is derived from the additive homomorphic property of the Pedersen commitment scheme, where

$$\begin{aligned} \prod_{u \in \mathcal{U}} commit_u &= \prod_{u \in \mathcal{U}} C(r_u, x_u) = C\left(\sum_{u \in \mathcal{U}} r_u, \sum_{u \in \mathcal{U}} x_u\right) \\ &= C\left(\sum_{u \in \mathcal{U}} r_u, \bar{x}\right) \end{aligned} \quad (7)$$

Theorem 4: The proposed zero-knowledge proof protocol satisfies the completeness, zero-knowledge, and soundness properties with soundness error $\frac{2}{3}^\lambda$ where λ is a security parameter.

Proof: Refer to Appendix C. \square

VI. EVALUATION

In this section, we analyze the computation and communication cost of the proposed protocol on both the users and the server. Then we implement a proof-of-concept and benchmark the performance.

A. Performance Analysis

1) *Users:* For the communication cost, w.r.t the secure aggregation, a user has to send out the encrypted model,

Algorithm 2 Generate Zero-Knowledge Proof

Input: $x_u, \lambda \in \mathbb{N}$, a perfectly correct and 2-private MPC Π_f protocol among 3 players

Output: Proof π_u

- 1 $\pi_u \leftarrow \emptyset;$
- 2 **for** $k = 1, \dots, \lambda$ **do**
- 3 Samples 3 random vectors x_1, x_2, x_3 such that $x_u = x_1 + x_2 + x_3$;
- 4 Consider an 3-input function $f(x_1, x_2, x_3) = R(r, x_1 + x_2 + x_3)$ and emulate the protocol Π_f on inputs x_1, x_2, x_3 . Obtain the views $v_i = \text{View}_{P_i}(x)$ for all $i \in [3]$;
- 5 Computes the commitments $\text{commit}_1, \dots, \text{commit}_3$ to each of the n produced views v_1, v_2, v_3 ;
- 6 For $j \in \{1, 2\}$, computes $e_j = H(j, \{\text{commit}_i\}_{i \in [3]})$;
- 7 $\pi_u \leftarrow \pi_u \cup (\{e_j\}_{j \in \{1,2\}}, \{v_{e_j}\}_{j \in \{1,2\}}, \{\text{commit}_i\}_{i \in [3]})$
- 8 **return** π_u

Algorithm 3 Validate Zero-Knowledge Proof

Input: π_u, commit_u

Output: 1 if accept, 0 if reject

- 1 **for** $p_k \in \pi_u$ **do**
- 2 $(\{e_j\}_{j \in \{1,2\}}, \{v_{e_j}\}_{j \in \{1,2\}}, \{\text{commit}_i\}_{i \in [3]}) \leftarrow p_k$;
- 3 **if** $\{\text{commit}_i\}_{i \in [3]}$ is invalid **then**
- 4 **return** 0
- 5 **if** $e_j \neq H(j, \{\text{commit}_i\}_{i \in [3]})$ for $j \in \{1, 2\}$ **then**
- 6 **return** 0
- 7 **if** $\exists e \in \{e_1, e_2\} : \Pi_{f,e}(\text{View}_{P_e}(x)) \neq \text{commit}_u$ **then**
- 8 **return** 0
- 9 **if** v_{e_1} and v_{e_2} are not consistent with each other **then**
- 10 **return** 0
- 11 **return** 1

which cost $O(m)$. Regarding the zero-knowledge proof, the user needs to submit a proof of size $O(mk\lambda)$ where k is the size of the commitments.

As regards the computation cost, the computation includes packing the data vector into polynomials and perform encryption which take $O(m + s)$ in total for the secure aggregation. The zero-knowledge proof protocol requires λ runs of $O(L)$ model inferences where L is the size of the validation dataset, thus the total cost is $O(\lambda L)$

2) *Server*: In the secure aggregation scheme, since the server receives data that are sent from all users, the communication complexity is $O(m|\mathcal{U}|)$. As regards the computation cost, for the secure aggregation, the computation includes aggregating the ciphertext from each user and perform decryption on the aggregated model, which takes $O(m|\mathcal{U}|^2)$ in total. Validating the zero-knowledge proofs from all users requires another $O(|\mathcal{U}|k\lambda)$ computations.

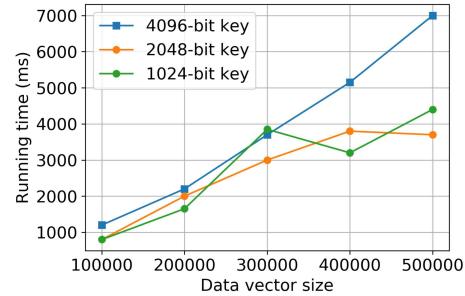


Fig. 6. Wall-clock running time per user, as the size of the data vector increases.

B. Proof-of-Concept and Benchmarking

We evaluate the overhead of our proposed framework based on the running time and communication costs. To that extent, we use the following metrics

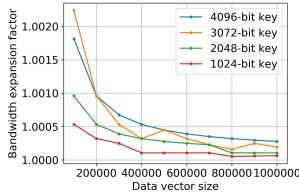
- Wall-clock running time: the total time it takes from the beginning to the end of the framework execution.
- Bandwidth consumption: the total amount of data sent/received by each entity in the system
- Bandwidth expansion factor: the ratio between the total bandwidth consumption when using the framework divided by the total bandwidth consumption when not using the framework.

1) *Secure Aggregation*: To measure the performance, we implement a proof-of-concept in C++ and Python 3. The Damgård-Jurik cryptosystem is used with 1024-, 2048-, 3072-, and 4096-bit keys. To handle big numbers efficiently, we use the GNU Multiple Precision Arithmetic Library [18]. We assume that each element in $\sum_{u \in \mathcal{U}} x_u$ can be stored with 3 bytes without overflow. This assumption conforms with the experiments in [8]. All the experiments are conducted on a single-threaded machine equipped with an Intel Core i7-8550U CPU and 16GB of RAM running Ubuntu 20.04.

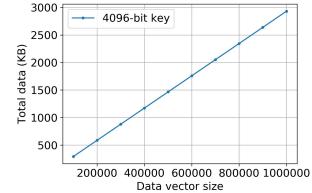
a) *Key generation*: We first discuss the overhead of distributed key generation. Note that this process can be included during the one-time setup phase of the FL system. An implementation of this phase can be adopted from [13]. It is shown that with 128 users, the running time for key generation per user is less than 50 seconds, and each user consumes about 15MB of network bandwidth. Therefore, the key generation phase imposes negligible overhead on the system.

b) *Users*: Wall-clock running times for users is plotted in Fig. 6. We measure the running time with key sizes of 1024, 2048, and 4096, respectively. As can be seen, under all settings, the result conforms with the running time complexity analysis. Also, note that the computation of each user is independent of each other, hence, the number of users does not affect the user's running time. By polynomial packing the data vector and optimizing the encryption process, the user can encrypt the entire data vector within seconds.

Fig. 7a shows the bandwidth expansion factor of the model encryption per user as we increase the data vector size. It can be observed that the factor approaches 1 as the model size increases, thereby conforming to the analysis of reducing the

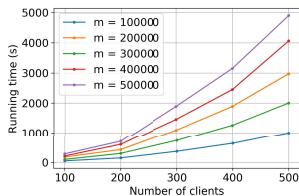


(a) Total bandwidth expansion factor per user, as compared to sending the raw data vector to the server. Different lines represent different values of key sizes k

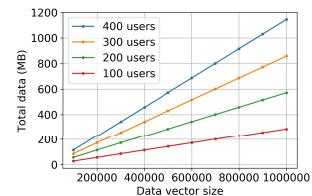


(b) The amount of transmitted data keys are nearly identical to 4096-bit key

Fig. 7. Bandwidth consumption per user.



(a) Wall-clock running time of the server at different data vector sizes, as the number of users increases.



(b) Total bandwidth consumption of the server at various numbers of users, as the data vector size increase.

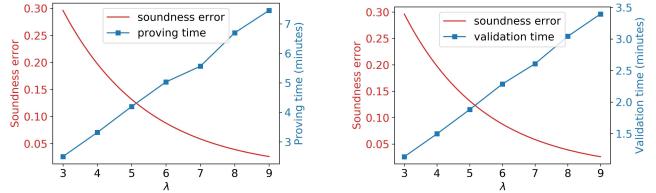
Fig. 8. Running time and bandwidth consumption of the server. The key size is fixed to 4096 bits.

ciphertext size. Moreover, the number of users also does not affect the bandwidth consumption. This result implies that our secure aggregation scheme induces minimal overhead as compared to sending the raw data vector, and can scale to a large number of users.

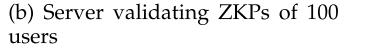
Fig. 7b illustrates the total data transfer of each user. The total data increases linearly with the data vector size, which follows the communication complexity $O(m)$. Furthermore, as shown in Fig. 7a, the amount of transmitted data is very close to the raw data size under all key sizes, therefore, the total data transfer is approximately the vector size multiplied by 3 bytes.

c) *Server*: On the server's side, Fig. 8a shows the running time on the server as we increase the number of users. As can be seen, with 100 users, the server only needs less than 6 minutes to perform decryption, and the result conforms to the running time complexity of $O(m|\mathcal{U}|^2)$. Note that our testing machine (Intel Core i7, 16GB RAM) has much less computation power than state-of-the-art servers. Furthermore, Fig. 8b gives the amount of data received by the server. Since we already optimized the size of data sent by users, the bandwidth consumption of the server is also optimized, which follows the communication complexity $O(m|\mathcal{U}|)$.

2) *Zero-Knowledge Proof*: For the implementation of the zero-knowledge backdoor detection, we train a neural network to use as an input. We use the MNIST dataset and train a 3-layer DNN model in the Tensorflow framework using ReLU activation at the hidden layers and softmax at the output layer. The training batch size is 128. Then, we implement an MPC protocol that is run between 3 emulated parties based on [35]. The model's accuracy is 93.4% on the test set.



(a) User generating ZKP



(b) Server validating ZKPs of 100 users

Fig. 9. Wall-clock running time of the zero-knowledge proof protocol on both users and server, as λ increases.

a) *Users*: To benchmark the performance of the ZKP protocol, we measure the proving and validation time against the soundness error. As shown before, the soundness error is $\frac{2}{3}^\lambda$ where λ is the security parameter. Fig. 9a illustrates the proving time to generate a ZKP by each user. We can see that as λ increases, the soundness error decreases, nonetheless, the proving time increases linearly. This is a trade-off since the protocol runs slower if we need to obtain a better soundness error. It only takes more than 5 minutes to achieve a soundness error of 0.06, and roughly 7 minutes to achieve 0.03.

b) *Server*: Fig. 9b shows the validation time of the ZKP protocol when the server validates the proofs of 100 users. Similar to the proving time, there is a trade-off between the running time and the soundness error. For a soundness error of 0.03, we need about 3.5 minutes to validate 100 users, which is suitable for most FL use cases [7].

To ensure that the ZKP has minimal impact on the model performance, we have also evaluated our framework using CIFAR-10 [23] and Imagenette [20]. After training the models on these datasets under our framework with FL, we attain an accuracy of 86.6% and 92.5% on CIFAR-10 and Imagenette, respectively. This shows that our framework imposes negligible trade-offs in terms of model accuracy.

VII. RELATED WORK

A. Secure Aggregation in FL

Leveraging secret sharing and random masking, Bonawitz et al. [8] propose a secure aggregation protocol and utilize it to aggregate local models from users. However, the protocol relies on users honestly following the secret sharing scheme. Consequently, although the privacy guarantee for honest users still retains, a single malicious user can arbitrarily deviate from the secret sharing protocol and make the server fail to reconstruct the global model. Furthermore, there is no mechanism to identify the attacker if such an attack occurs.

In [2] and [39], the decryption key is distributed to the users, and the server uses homomorphic encryption to blindly aggregate the model updates. However, the authors assume that there is no collusion between the server and users so that the server cannot learn the decryption key. Hence, the system does not work under our security model where users can be malicious and collude with the server.

Additionally, there have been studies on how to use generic secure MPC based on secret sharing to securely compute any

function among multiple parties [4], [12], [24]. However, with these protocols, each party has to send a secret share of its whole data vector to a subset of the other parties. In order to make the protocols robust, the size of this subset of users should be considerably large. Since each secret share has the same size as the entire data vector's, these approaches are not practical in federated learning in which we need to deal with high-dimensional vectors.

On the other hand, differential privacy (DP) is considered as a prominent privacy tool with a formal guarantee on the privacy leakage [14], [28]. However, DP only aims to protect membership privacy and does not conceal the local models from the server. Previous work has shown that the users in FL are still susceptible to other privacy attacks, such as data reconstruction, even when DP is used [6], [15]. Furthermore, DP comes with a trade-off that reduces the model performance. It has been suggested that using DP makes it impossible to train a good model for datasets like CIFAR-10 or ImageNet with reasonable accuracy [6], [34]. Our proposed protocol fully hides the local models from the server, disabling any attacks that rely on inspecting the local models, while maintaining the performance of the models.

B. Zero-Knowledge Proof for Detecting Poisoned Models

Although there have been many studies on developing a ZKP protocol based on MPC, they have not been widely used in the machine learning context. The IKOS protocol proposed by Ishai et al. [21] is the first work that leverages secure multi-party computation protocols to devise a zero-knowledge argument. Giacomelli et al. [17] later refine this approach and construct ZKBoo, a zero-knowledge argument system for Boolean circuits using collision-resistant hashes that does not require a trusted setup. Our proposed ZKP protocol is inspired by the IKOS protocol with repetitions to reduce the soundness error to $\frac{2}{3}^\lambda$. We also demonstrate how such protocol can be used in the context of machine learning, especially for attesting non-poisoned models while maintaining privacy guarantees.

Another line of research in ZKP focuses on zkSNARK implementations [5] that have been used in practical applications such as ZCash [33]. However, these systems depend on cryptographic assumptions that are not standard, and have large overhead in terms of memory consumption and computation cost, thereby limiting the statement sizes that they can manage. Therefore, it remains a critical challenge whether zkSNARK could be used in machine learning where the circuit size would be sufficiently large.

C. Defense Mechanisms Against Poisoning Attacks

There have been multiple research studies on defending against poisoning attacks. Liu et al. propose to remove backdoors by pruning redundant neurons [25]. On the other hand, Bagdasaryan et al. [3] devise several anomaly detection methods to filter out attacks. More recently proposed defense mechanisms [26], [36] detect backdoors by finding differences between normal and infected label(s).

Our work leverages such a defense mechanism to construct a ZKP protocol for FL in which the users can run

the defense locally on their model updates, and attest its output to the server, without revealing any information about the local models. Different defense mechanisms may have different constructions of the ZKP protocols, nevertheless, they must all abide by the properties specified in our framework (Section III). As shown in Section V, we construct a ZKP protocol based on the defense proposed by Liu et al. [25].

VIII. CONCLUSION

In this paper, we have proposed a secure framework for federated learning. Unlike existing research, we integrate both secure aggregation and defense mechanisms against poisoning attacks under the same threat model that maintains their respective security and privacy guarantees. We have also proposed a secure aggregation protocol that can maintain liveness and privacy for model updates against malicious users. Furthermore, we have designed a ZKP protocol for users to attest non-backdoored models without revealing any information about their models. Our framework combines these two protocols and shows that the server can detect backdoors while preserving the privacy of the model updates. The privacy guarantees for the users' models have been theoretically proven. Moreover, we have presented an analysis of the computation and communication cost and provided some benchmarks regarding its runtime and bandwidth consumption

REFERENCES

- [1] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.
- [2] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1333–1345, May 2018.
- [3] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2938–2948.
- [4] M. L. Ben-Or, S. Goldwasser, and A. Wigderson, "Completeness theorems for non-cryptographic fault-tolerant distributed computation," in *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 351–371, doi: [10.1145/3335741.3335756](https://doi.org/10.1145/3335741.3335756).
- [5] N. Bitansky, A. Chiesa, Y. Ishai, O. Paneth, and R. Ostrovsky, "Succinct non-interactive arguments via linear interactive proofs," in *Proc. Theory Cryptography Conf.* Cham, Switzerland: Springer, 2013, pp. 315–333.
- [6] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "When the curious abandon honesty: Federated learning is not private," in *Proc. IEEE 8th Eur. Symp. Secur. Privacy (EuroS&P)*, Delft, The Netherlands, 2023, pp. 175–199. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/EuroSP57164.2023.00020>
- [7] K. Bonawitz et al., "Towards federated learning at scale: System design," 2019, *arXiv:1902.01046*.
- [8] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1175–1191.
- [9] A. Canteaut et al., "Stream ciphers: A practical solution for efficient homomorphic-ciphertext compression," *J. Cryptol.*, vol. 31, no. 3, pp. 885–916, Jul. 2018.
- [10] I. Damgård, M. Jurik, and J. B. Nielsen, "A generalization of Paillier's public-key system with applications to electronic voting," *Int. J. Inf. Secur.*, vol. 9, no. 6, pp. 371–385, Dec. 2010.
- [11] I. Damgård and M. Koprowski, "Practical threshold RSA signatures without a trusted dealer," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.* Cham, Switzerland: Springer, 2001, pp. 152–165.
- [12] I. Damgård, V. Pastro, N. Smart, and S. Zakarias, "Multiparty computation from somewhat homomorphic encryption," in *Proc. Annu. Cryptol. Conf.* Cham, Switzerland: Springer, 2012, pp. 643–662.

- [13] S. Das, T. Yurek, Z. Xiang, A. Miller, L. Kokoris-Kogias, and L. Ren, "Practical asynchronous distributed key generation," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2022, pp. 2518–2534.
- [14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
- [15] L. H. Fowl, J. Geiping, W. Czaja, M. Goldblum, and T. Goldstein, "Robbing the fed: Directly obtaining private data in federated learning with modified models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–25.
- [16] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.
- [17] I. Giacomelli, J. Madsen, and C. Orlandi, "ZKBOO: Faster zero-knowledge for Boolean circuits," in *Proc. 25th Usenix Secur. Symp. (Usenix Security)*, 2016, pp. 1069–1083.
- [18] T. Granlund. (2010). *Gnu Multiple Precision Arithmetic Library*. [Online]. Available: <http://gmplib.org/>
- [19] A. Hard et al., "Federated learning for mobile keyboard prediction," 2018, *arXiv:1811.03604*.
- [20] J. Howard. (2019). *Imagenette*. Accessed: 2020. [Online]. Available: <https://github.com/fastai/imagenette>
- [21] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, "Zero-knowledge from secure multiparty computation," in *Proc. 39th Annu. ACM Symp. Theory Comput.*, Jun. 2007, pp. 21–30.
- [22] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*. Boca Raton, FL, USA: CRC Press, 2020.
- [23] A. Krizhevsky et al., "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: <https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf>
- [24] Y. Lindell, B. Pinkas, N. P. Smart, and A. Yanai, "Efficient constant round multi-party computation combining BMR and SPDZ," in *Proc. Annu. Cryptol. Conf.* Cham, Switzerland: Springer, 2015, pp. 319–338.
- [25] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Res. Attacks, Intrusions, Defenses*. Cham, Switzerland: Springer, 2018, pp. 273–294.
- [26] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "ABS: Scanning neural networks for back-doors by artificial brain stimulation," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 1265–1282.
- [27] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.
- [28] T. Nguyen, P. Lai, H. Phan, and M. T. Thai, "XRAND: Differentially private defense against explanation-guided attacks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, pp. 11873–11881, 2023.
- [29] T. Nguyen, P. Lai, K. Tran, N. Phan, and M. T. Thai, "Active membership inference attack under local differential privacy in federated learning," in *Proc. 26th Int. Conf. Artif. Intell. Statist.*, vol. 206. PMLR, 2023, pp. 5714–5730. [Online]. Available: <https://proceedings.mlr.press/v206/nguyen23e.html>
- [30] T. Nishide and K. Sakurai, "Distributed Paillier cryptosystem without trusted dealer," in *Proc. Int. Workshop Inf. Secur. Appl.* Cham, Switzerland: Springer, 2010, pp. 44–60.
- [31] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn.* Cham, Switzerland: Springer, 1999, pp. 223–238.
- [32] T. P. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Proc. Annu. Int. Cryptol. Conf.* Cham, Switzerland: Springer, 1991, pp. 129–140.
- [33] E. B. Sasson et al., "Zerocash: Decentralized anonymous payments from Bitcoin," in *Proc. IEEE Symp. Secur. Privacy*, May 2014, pp. 459–474.
- [34] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–14.
- [35] S. Wagh, D. Gupta, and N. Chandran, "SecureNN: 3-party secure computation for neural network training," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 3, pp. 26–49, Jul. 2019.
- [36] B. Wang et al., "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.
- [37] H. Wang et al., "Attack of the tails: Yes, you really can backdoor federated learning," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 16070–16084.
- [38] C. Xie, K. Huang, P.-Y. Chen, and B. Li, "DBA: Distributed backdoor attacks against federated learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.
- [39] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proc. USENIX Annu. Tech. Conf. (USENIX ATC)*, 2020, pp. 493–506.



Truc Nguyen received the B.Eng. degree in computer engineering from the Ho Chi Minh City University of Technology, Vietnam, in 2018, and the Ph.D. degree in computer engineering from the Department of Computer and Information Science and Engineering, University of Florida, in 2023. His current research interests include trustworthy machine learning, cybersecurity, and blockchain.



My T. Thai (Fellow, IEEE) is currently a UF Research Foundation Professor with the Computer and Information Science and Engineering Department, University of Florida. She is also the Associate Director of the Nelms Institute for the Connected World. She has engaged in many professional activities. Her current research interests include trustworthy machine learning, big data analysis, cybersecurity, and optimization in network science and engineering. Her research works have led to six books and more than 250 articles, including several best paper awards. The most recent ones are the 2023 AAAI Distinguished Papers Award and the 2023 Web Science Test-of-Time Award. She is also a recipient of the Department of Defense Young Investigator Award and the National Science Foundation CAREER Award. She is the Founding Editor-in-Chief of the *IET Blockchain* journal and the Editor-in-Chief of the *Journal of Combinatorial Optimization*.