Poisoning with Cerberus: Stealthy and Colluded Backdoor Attack against Federated Learning

Xiaoting Lyu[1], Yufei Han[2], Wei Wang[1]*, Jingkai Liu[1], Bin Wang[3], Jiqiang Liu[1], Xiangliang Zhang[4]

[1]Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, China
[2]INRIA, France
[3]Zhejiang Key Laboratory of Multi-dimensional Perception Technology, Application and Cybersecurity, China
[4]University of Notre Dame, USA
{xiaoting.lyu, wangwei1, jingkai.liu, jqliu}@bjtu.edu.cn, yfhan.hust@gmail.com, bin_wang@zju.edu.cn, xzhang33@nd.edu

Abstract

Are Federated Learning (FL) systems free from backdoor poisoning with the arsenal of various defense strategies deployed? This is an intriguing problem with significant practical implications regarding the utility of FL services. Despite the recent flourish of poisoning-resilient FL methods, our study shows that carefully tuning the collusion between malicious participants can minimize the trigger-induced bias of the poisoned local model from the poison-free one, which plays the key role in delivering stealthy backdoor attacks and circumventing a wide spectrum of state-of-the-art defense methods in FL. In our work, we instantiate the attack strategy by proposing a distributed backdoor attack method, namely *Cerberus Poisoning* (CerP). It jointly tunes the backdoor trigger and controls the poisoned model changes on each malicious participant to achieve a stealthy yet successful backdoor attack against a wide spectrum of defensive mechanisms of federated learning techniques. Our extensive study on 3 large-scale benchmark datasets and 13 mainstream defensive mechanisms confirms that *Cerberus Poisoning* raises a significantly severe threat to the integrity and security of federated learning practices, regardless of the flourish of robust Federated Learning methods.

Introduction

The distributed nature of federated learning makes it vulnerable to backdoor attacks carried out by malicious participants, as unveiled in recent studies (Fung, Yoon, and Beschastnikh 2018; Baruch, Baruch, and Goldberg 2019; Xie et al. 2020; Bagdasaryan et al. 2020; Wang et al. 2020; Fang et al. 2020; Shejwalkar and Houmansadr 2021). Several malicious participants can collude to embed a well-designed backdoor trigger into their local training data to poison the global aggregated model. The perturbed global model then misclassifies the input instances embedded with the backdoor trigger as the target label specified by the adversary. However, the perturbed global model performs normally on clean input instances. In security-critical applications, e.g., distributed video surveillance and credit risk assessment, the adversary can compromise several local computing devices of the target federated learning system to launch such backdoor attacks. The adversary can mislead the jointly trained global model to misidentify malicious activities with the trigger signal on one hand. On the other hand, the poisoned model works normally on other inputs without the trigger. It is thus difficult for the FL service owners to flag the malfunction caused by colluding backdoor attacks.

There has been a long line of efforts exploring various defensive mechanisms to prevent distributed poisoning attacks, including backdoor attacks. These defense methods detect anomalies in the submitted local models to mitigate potential poisoning effects (Blanchard et al. 2017; Yin et al. 2018; Mhamdi, Guerraoui, and Rouault 2018; Pillutla, Kakade, and Harchaoui 2019; Shejwalkar and Houmansadr 2021; Cao et al. 2021), introduce additional random perturbations to model parameters to gain certified robustness against the backdoor noise (Geyer, Klein, and Nabi 2017; Sun et al. 2019; Wei et al. 2021; Xie et al. 2021; Sun et al. 2021), down-weigh the poisoned local model updates sharing similar model parameters to mitigate colluding poisoning (Fu et al. 2019; Fung, Yoon, and Beschastnikh 2020), and build a voting-based defense mechanism to filter poisoned local models (Cao, Jia, and Gong 2021; Andreina et al. 2021). As confirmed in our empirical study, these defensive mechanisms can indeed effectively mitigate the existing backdoor attacks. They provide a rich arsenal of robust learning solutions for service providers of FL systems.

However, our study shows that the distributed backdoor poisoning threat against federated learning is far from being well addressed. We demonstrate that adversaries can organize the collusion of malicious participants to easily dodge various state-of-the-art defensive mechanisms while successfully launching distributed backdoor attacks. The adversary achieves this goal by exploiting the fundamental assumptions of different defensive mechanisms and adjusting the learning objectives of backdoor attacks accordingly.

Our contributions can be summarized in the following perspectives.

1) Theoretically, we establish lower and upper bounds on the trigger-induced local model changes for malicious participants in general federated learning tasks. Instead of requiring the local model parameters of different participants to be IID, we assume the gradients of the local models are bounded and Lipschitz-continuous. This assumption

*Corresponding author.

holds for most practically deployed machine learning models and is generic enough for both IID and non-IID federated learning scenarios. Our analysis identifies the key factors controlling the local model bias induced by backdoor triggers, thus deciding the feasibility of federated backdoor attacks exposed to defense methods.

2) We propose a stealthy distributed backdoor attack, namely Cerberus Poisoning (*CerP*) by exploiting the algorithmic principles of current defense methods in federated learning. Despite originating from different perspectives and threat model settings, these defensive mechanisms share the same core assumption: *regardless of the data distribution of participants, each poisoned local model trained with poisoned data is biased largely from those trained by the poison-free data.* Exploiting the limit of this assumption, *CerP* casts the distributed backdoor attack as a joint optimization process of three learning objectives.

Automatic fine-tuning of backdoor triggers. Based on the theoretical analysis, we believe that the injected trigger is an important factor in determining the magnitude of variation in the parameter bias of the poisoned local models. In *CerP*, we propose to *fine-tune the backdoor trigger* to facilitate the learning of the poisoned data and reduce the parameter bias of the poisoned local models.

Control over local model bias. For each malicious participant, we suppress the model parameter bias between the poisoned local model and its poison-free counterpart that would have been derived if no trigger noise was injected. We also explain the theoretical rationality of explicitly suppressing the bias in the poisoned local models.

Diversity of poisoned local models. To bypass the sybil-attack mitigation methods (Fung, Yoon, and Beschastnikh 2018), we require the poisoned local models submitted by malicious participants to be as dissimilar as possible. We enlarge the divergence of the poisoned local models to avoid being flagged by the similarity-based defenses.

3) Our comprehensive empirical evaluation shows that the proposed *CerP* method circumvents all the 13 defense methods on 3 large-scale benchmark datasets (*CIFAR-100*, *Fashion-MNIST*, and *LOAN*). Compared to other state-of-the-art distributed backdoor attack methods (*Sybil attack* (Fung, Yoon, and Beschastnikh 2018), *LIE* (Baruch, Baruch, and Goldberg 2019), and *DBA* (Xie et al. 2020)), the experimental results show that *CerP* achieves consistently higher attack success rates and maintains the accuracy of the main learning task, no matter which defense strategy is employed. In contrast, none of the 3 distributed backdoor attack methods can neutralize all defenses.

Related Work

Backdoor Attacks against Federated Learning. Pioneering studies on backdoor attacks against federated learning systems (Fung, Yoon, and Beschastnikh 2018; Bagdasaryan et al. 2020) assume that each malicious participant trains their local models individually, without any collusion between them. Since they use the same backdoor trigger, the poisoned local models tend to share similar parameter values and are largely deviated from benign local models. These attacks are thus easily mitigated by the Byzantine-robust aggregation methods and sybil-attack mitigation methods like *Foolsgold* (Fung, Yoon, and Beschastnikh 2018).

More advanced distributed backdoor threats (Baruch, Baruch, and Goldberg 2019; Sun et al. 2019) consider how to evade Byzantine-robust aggregation rules. They clip the parameters of the poisoned local model according to the parameter range of the benign local models. However, they either assume the parameters of the benign local models are IID Gaussian variables so that the bounds on benign parameter values can be estimated (Baruch, Baruch, and Goldberg 2019), or assume that the benign parameter bounds are known as prior knowledge (Sun et al. 2019). Neither of these assumptions holds in practice, especially in non-IID federated learning tasks. Therefore, the manually configured parameter clipping bounds may be overestimated (downgrading the learning capability), or underestimated (failing the attack task). Besides, some methods require knowing the model parameters committed by the benign participants, which violates the protocol of federated learning.

Alternatively, *DBA* (Xie et al. 2020) manages stealthy backdoor attacks by manually decomposing a global backdoor trigger into different local triggers and assigning separately the local triggers to each malicious participant. The malicious participants learn to fit different local triggers, and thus have dissimilar poisoned local models to bypass the sybil-attack mitigation methods. However, how to properly decompose global triggers to guarantee successful backdoor attacks remains an open issue. Manual decomposition of global triggers unavoidably introduces artifacts into local triggers. *DBA* may thus lead to large deviations of the poisoned local models from the benign ones. Therefore, this method fails to attack the popular Byzantine-robust aggregation methods such as *Krum* and *Bulyan*.

Byzantine-robust aggregation methods. These methods (Blanchard et al. 2017; Yin et al. 2018; Mhamdi, Guerraoui, and Rouault 2018; Pillutla, Kakade, and Harchaoui 2019; Shejwalkar and Houmansadr 2021; Cao et al. 2021) follow the spirit of anomaly detection. The core hypothesis assumes that the parameters of all benign local models stay within a bounded l_p-ball centered on the global model. Therefore, the poisoned local models are considered to be outliers that largely deviate from benign local models. Nevertheless, they may fail to detect distributed backdoor attacks that are dedicated to minimizing the distance between the malicious and benign local models.

Differential privacy-based methods. These defense methods (Geyer, Klein, and Nabi 2017; Wei et al. 2021; Sun et al. 2021; Xie et al. 2021) adopt the core ideas of *differential privacy theory* and *random smoothing* (Cohen, Rosenfeld, and Kolter 2019). These methods add Gaussian or Laplace noises to the parameters of the global model or local models. The injected perturbation makes the disturbed models insensitive to backdoor triggers. However, how to properly set the noise magnitude is still open in practice. Too strong or too weak noise may either harm the utility of the target model or weaken the capability to defend against the attack.

Other defense methods like *Foolsgold* (Fung, Yoon, and Beschastnikh 2018) identify poisoning sybils based on the

similarity of local model updates. Nevertheless, the adversary can encourage the diversity of poisoned local models to bypass these defense methods. *Ensemble FL* method (Cao, Jia, and Gong 2021) trains multiple global models to make a majority vote on prediction decisions. Similarly, *BaFFLe* (Andreina et al. 2021) is also a voting-based defense, where participants validate the global model on their local data and vote to accept or reject the global model. Since benign participants are agnostic to attacker-designed triggers, it is difficult to recognize and flag backdoor poisoning efforts only by inspecting the overall classification performance of the poison-free testing data.

Algorithm Description of Cerberus Poisoning

Preliminaries

Federated Learning. We focus on the setting of Federated Learning with *partial participation*, i.e. N_p out of N participants are selected in each training iteration of federated learning. Compared to the full participation setting, the partial participation setting of FL is better adapted to real-world machine learning applications, such as mobile edge computing, where local devices may join or leave the FL service at will. Each participant i hosts a local dataset $D_i = \left\{\{x_{i,j} \in R^m, y_{i,j}\}_{j=1}^{d_i}\right\}$, where $d_i = |D_i|$ and $\{x_{i,j}, y_{i,j}\}$ represents the features and label of each data instance. At each training iteration t, we use g^{t-1} and h_i^t to represent the global model shared with the selected participants and the local model of each participant i respectively. The server updates the global model by aggregating the local model updates with a learning rate η: $g^t = g^{t-1} + \frac{\eta}{N_p}\sum_{i=1}^{N_p}(h_i^t - g^{t-1})$.

Adversary's goal. The goal of backdoor attacks on FL is twofold: 1) The classifier derived by the poisoned federated training process should produce the expected decision output set by the backdoor trigger. 2) The accuracy of the main learning task on poison-free data should not be perturbed.

Adversary's capability. The adversary compromises C of the total N participants ($C \ll N$). In a training iteration of partial-participation federated learning, *each selected malicious participant* controlled by the adversary injects backdoor poisoned instances into the local training set. Note that the adversary cannot know or tamper with *the global aggregation rules* or *the local training process of benign participants*. The malicious participants can collude by sharing their poisoned local models with the adversary. The adversary can then share back to each malicious participant the poisoned local models submitted by other malicious participants as additional knowledge to guide how to control the local model changes of the malicious participants.

The Objective of Cerberus Poisoning

We define *CerP* as a distributed optimization problem with an objective function given in Eq.1. The aggregated global model is trained to fit both the clean training data and poisoned training data with the trigger. Given a federated training iteration t, the adversary launches the backdoor attack to derive the poisoned local models $h_i^{*,t}$ of the compromised participants via jointly optimizing the learning objective with compromised participants, which gives in Eq.1.

$$\begin{aligned}\Delta x^{*,t}, \{h_i^{*,t}\}_{i\in S} = \underset{\Delta x, h_i^t(i\in S)}{\arg\min} \{&\sum_{i\in S}\Big(\sum_{j\in D_i^{nor}} \ell_{h_i^t}(x_{i,j}, y_{i,j}) \\ &+ \sum_{j\in D_i^{mal}} \ell_{h_i^t}(x_{i,j}+\Delta x, \hat{y}_{i,j})\Big) + \alpha\sum_{i\in S}\left\|h_i^t - h_i^{nor,t}\right\|_{Fro} \\ &+ \beta\sum_{i,i'\in S} cs(h_i^t, h_{i'}^t)\} \\ s.t.\ &\left\|\Delta x - \Delta x^0\right\|_2 \le \varphi\end{aligned} \tag{1}$$

where $\ell_{h_i^t}$ denotes the classification loss function given the labelled data instance (x, y) and the local classifier model h_i^t of each participant i. $\|\|_{Fro}$ is the Frobenius norm of a matrix. We use S to represent the set of compromised participants in the t-th federated training iteration. The attack process of *CerP* can be summarized from two perspectives:

For a compromised participant $i \in S$, D_i^{mal} and D_i^{nor} represent the backdoor-poisoned and the poison-free training data hosted by the participant i, respectively. $h_i^{nor,t}$ is the poison-free model trained by the compromised participant i at the iteration t. $cs(h_i^t, h_{i'}^t)$ is the pairwise cosine similarity between the local models submitted by **a pair of compromised participants** i and i' ($i, i' \in S$). According to the threat model setting, each compromised participant can access the poisoned local models derived by other compromised participants via the adversary. Except that, all of the compromised participants follow the standard federated learning setting. **Besides,** we consider the backdoor trigger Δx as an optimization variable in *CerP*. Δx^0 is the initial backdoor trigger designed by the adversary before launching the optimization process of Eq.1. The parameter φ in Eq.1 limits the distance between the finely tuned trigger Δx and the initial trigger Δx^0.

At the iteration t, solving the constrained optimization problem in Eq.1 produces the tuned backdoor trigger $\Delta x^{*,t}$ and the poisoned local model $h_i^{*,t}$. The finely tuned backdoor trigger $\Delta x^{*,t}$ can be used in the attack. The poisoned local models $h_i^{*,t}$ are committed to the central server to generate the poisoned global model for the next iteration. The learning objective of *CerP* (Eq.1) is four-fold:

Objective 1. Learning both clean and backdoor poisoned training data. The classification accuracy of the main learning task and the backdoor attack task is optimized. The main learning task is to train h_i^t on the clean training data hosted by the compromised participants (**the first term** in Eq.1). The backdoor attack task is to make h_i^t fit the backdoor poisoned training data hosted by the compromised participants (**the second term** of Eq.1).

Objective 2. Trigger fine-tuning for stealthy backdoor attacks. The backdoor trigger Δx is considered as an optimization variable of the attack objective. Intuitively, the backdoor trigger Δx and the poisoned local models are jointly tuned to minimize the learning loss of the backdoor poisoned training data. This is used to facilitate the poisoned local model h_i^t to accurately fit the backdoor poisoned data without inducing large biases in the local model parameters.

We establish the following theoretical analysis to explain the rationality of backdoor trigger tuning.

Without loss of generality, we use a multi-layer neural network $\mathcal{H}$ for C-class classification with K fully connected layers as the target model for the federated learning task. The classification function can be written as $\mathcal{H}(x) = \sigma_{K-1}h_{K-1}(\sigma_{K-2}h_{K-2}(\sigma_{K-3}h_{K-3}(\cdots\sigma_0 h_0(x))))$, where σ_k and $h_k \in R^{d_k \times d_{k+1}}$ ($k = 0, 1, 2, ..., K-1$ and $d_K = C$) are the activation function and the parameter matrix of each layer, respectively. Following (Wang et al. 2020), we define an ϵ-adversarial equivalent sample x' to a backdoor poisoned sample $x + \Delta x$.

Definition 1. Given a targeted classifier $\mathcal{H}^{nor}$ and a backdoor poisoned classifier $\mathcal{H}^{mal}$, an ϵ-adversarial equivalent x' is defined as:

$$\begin{aligned} &\mathcal{H}^{nor}(x') = \mathcal{H}^{mal}(x + \Delta x), \\ &x' = x + \Delta x + \epsilon_x, \\ &s.t. \quad \|\epsilon_x\|_2 \leq \epsilon \end{aligned} \tag{2}$$

Assume that the backdoor attack is successfully delivered at the training iteration t. Let $h_{i,k}^{nor,t}$ and $h_{i,k}^{mal,t}$ denote the parameter matrix of the layer k of the poison-free and poisoned local model of the compromised participant i. Let $g_k^{nor,t-1}$ be the parameter matrix of the layer k of the aggregated global model at the training iteration $t-1$. The distance between $h_{i,k}^{mal,t}$ and $g_k^{nor,t-1}$ (the trigger-induced local model bias) can be bounded in Theorem 1.

Theorem 1 *Let $\ell_{\mathcal{H}}(x, y)$ be the classification risk function of a federated learning task. Its gradient with respect to each coordinate j of $h_{i,k}^t$, $\nabla_{h_{i,k}^t,j}\ell_{\mathcal{H}}(x, y)$ is bounded by the Lipschitz constant L, i.e. $|\nabla_{h_{i,k}^t,j}\ell_{\mathcal{H}}(x, y)| \leq L$. Thus the coordinate-wise gradient follows a γ-subgaussian distribution with mean μ. $X(k)$ represents the input to the layer k of $\mathcal{H}$ by feeding a set of input instances X to $\mathcal{H}$. For a given layer k ($1 \leq k \leq K$), the distance between $h_{i,k}^{mal,t}$ and $g_k^{nor,t-1}$ can be bounded from above with a probability of $p \geq 1 - 2d_k d_{k+1} m N_\delta e^{\{-n \min\{\sqrt{d_k d_{k+1}}L/\gamma, 2d_k d_{k+1}L^2/\gamma^2\}\}}$. The constant N_δ is defined such that $N_\delta \leq (1 + D/\delta)^{d_k d_{k+1}}$, where δ is the covering number of the layer k parameter matrix $h_{i,k}^{nor,t}$ (Vershynin 2011).*

$$\begin{aligned} \|h_{i,k}^{mal,t} - g_k^{nor,t-1}\|_{Fro} \leq \frac{\|\epsilon\|_2\sqrt{d_i^{mal}}\prod_{s=0}^{k}\|h_{i,s}^{nor,t}\|_*}{\rho_k} \\ +2\eta_t\sqrt{d_k d_{k+1}}L + \|g_k^{nor,t} - g_k^{nor,t-1}\|_{Fro} \end{aligned} \tag{3}$$

where ρ_k is the minimum eigenvalue of $X(k)$. ϵ is the perturbation bound of the adversarial equivalent to the backdoor poisoned sample $x + \Delta x$. $\|\|_*$ denotes the spectral norm of the parameter matrix. η_t is the learning rate of the federated training iteration t.

Corollary 1.1 *Inheriting the setting of Theorem 1, if the classification loss $\ell_{\mathcal{H}}(x, y)$ is L_c-Lipschitz continuous, the upper bound Eq.3 can be further formulated as:*

$$\begin{aligned} \|h_{i,k}^{mal,t} - g_k^{nor,t-1}\|_{Fro} \leq \frac{\|\epsilon\|_2\sqrt{d_i^{mal}}\prod_{s=0}^{k}\|h_{i,s}^{nor,t}\|_*}{\rho_k} \\ + 2\eta_k\sqrt{d_k d_{k+1}}L + \eta_t L_c\|g_k^{nor,t} - g_k^{nor,*}\|_{Fro} \end{aligned} \tag{4}$$

where $g_k^{nor,*}$ as the optimal parameters of the layer k derived once converged.

Theorem 2 *Following the same setting in Theorem 1, if the classification loss $\ell_{\mathcal{H}}(x, y)$ is L_c-Lipschitz continuous, for a given layer k ($1 \leq k \leq K$), the distance between $h_{i,k}^{mal,t}$ and $g_k^{nor,t-1}$ can be bounded from below as:*

$$\begin{aligned} \|h_{i,k}^{mal,t} - g_k^{nor,t-1}\|_{Fro} \geq \frac{\nu_k\|\epsilon\|_2}{\max\limits_{x_{i,k}^{nor}, x_{i,k}^{mal}}\|x_{i,k}^{nor} - x_{i,k}^{mal}\|_2} \\ - \eta_t L_c\|g_k^{nor,t} - g_k^{nor,*}\|_{Fro} \end{aligned} \tag{5}$$

where ν_k is the minimum non-zero singular value of the product of the parametric matrices $h_{i,k}^{nor}, h_{i,k-1}^{nor} \cdots h_{i,0}^{nor}$. $x_{i,k}^{nor}$ and $x_{i,k}^{mal}$ are the poison-free and poisoned samples in $X(k)$ hosted by the participant i respectively.

Our unveilings in the analysis can be summarized in two aspects. **First**, under the adversary-free scenario, $\|g_k^{nor,t} - g_k^{nor,t-1}\|_{Fro}$ in Eq.3 and $\|g_k^{nor,t} - g_k^{nor,*}\|_{Fro}$ in Eq.5 vanish when the federated training is close to convergence. Bearing this in mind, it is generally impossible to ensure the success of backdoor attacks without causing local model changes on compromised participants, unless $\epsilon = 0$ according to Eq.3 and Eq.5. However, the exception with $\epsilon = 0$ holds only when $\mathcal{H}^{nor}(x + \Delta x) = \hat{y}$. In this case, the trigger Δx should be chosen in a way that a classifier can produce accurately the target label of an input instance carrying the trigger, even without poisoning the classifier with the backdoor data. This situation is difficult to meet in practice.

Second, according to Eq.3 in Theorem 1, minimizing the value of the product $\|\epsilon\|_2\prod_{s=0}^{k}\|h_{i,s}^{nor,t}\|_*$ is the key to minimize the trigger-induced parameter changes to evade defensive methods. We propose to achieve this goal by adjusting the designated backdoor trigger Δx via Eq.6 (**the second term** in Eq.1). It aims at adapting the trigger to minimize the classification loss of the poisoned local model h_i^t on the backdoor poisoned instance $(x + \Delta x, \hat{y})$. According to the definition of ϵ-adversarial equivalent, Eq.6 directly minimizes the adversarial noise magnitude $\|\epsilon\|_2$ with respect to the local model h_i^t. Consequently, the derived finely tuned trigger $\Delta x^{*,t}$ can reduce both the upper and lower bounds of the trigger-induced model changes, as $\|\epsilon\|_2$ decreases without changing the local model h_i^t (thus $\prod_{s=0}^{k}\|h_{i,s}^{nor,t}\|_*$ keeps unchanged in the product $\|\epsilon\|_2\prod_{s=0}^{k}\|h_{i,s}^{nor,t}\|_*$). Meanwhile, fine-tuning the trigger helps reduce the learning loss on the backdoor poisoned training data, which alleviates the difficulty of memorizing the trigger-induced feature-label correlation.

$$\begin{aligned} &\Delta x^{*,t} = \arg\min_{\Delta x} \sum_{i \in S, j \in D_i^{mal}} \ell_{h_i^t}(x_{i,j} + \Delta x, \hat{y}_{i,j}) \\ &s.t. \|\Delta x - \Delta x^0\|_2 \leq \varphi \end{aligned} \tag{6}$$

Without loss of generality, we adopt L_2 distance to measure the magnitude of changes adapted to the finely tuned trigger.

Objective 3. Deviation regularization on local models. To bypass the defense methods, we control the parameter changes of the poisoned local models at the participant level.

For each compromised participant i, we minimize the distance between the poisoned local model h_i^t and the poison-free local model that could be derived if no trigger noise was injected (noted as $h_i^{nor,t}$). The distance is measured using the Frobenius norm as $\|h_i^t - h_i^{nor,t}\|_{Fro}$. Enforcing the distance regularization (**the third term** of Eq.1) helps reduce the spectral norm of h_i^t. Assuming $h_i^{nor,t}$ is close to convergence on the poison-free training data, $h_i^{nor,t}$ can be considered as a constant as the gradient vanishes. Following the matrix norm inequality, it can be found that minimizing $\|h_i^t - h_i^{nor,t}\|_{Fro}$ suppresses the upper bound of the spectral norm of h_i^t:

$$\|h_i^t - h_i^{nor,t}\|_{Fro} \geq \|h_i^t\|_* - const \tag{7}$$

where all the factors independent of the poisoned training data are absorbed into *const*. More intuitively, dragging h_i^t and $h_i^{nor,t}$ close together helps evade from the Byzantine-robust aggregation methods, e.g., *Trimmed mean* and *Krum*. The closer the poisoned local model h_i^t stays to the benign one $h_i^{nor,t}$, the more difficult it is for these defense methods to identify and exclude the compromised participants without raising false alarms.

The joint application of the trigger fine-tuning and the deviation regularization term over the poisoned local models makes both $\|\epsilon\|_2$ and the classifier's spectral norm in Eq.3 and Eq.5 decrease. They form the core of the proposed *CerP* attack method: *controlling trigger-induced local model changes to deliver stealthy yet successful federated backdoor attacks*.

Objective 4. Pairwise similarity regularization. We suppress the cosine similarity between the poisoned local models of the compromised participants (**the fourth term** of Eq.1). *Foolsgold* (Fung, Yoon, and Beschastnikh 2018) against sybil attacks reduces the aggregated weights of participants that repeatedly contribute similar local model updates. We enhance the diversity of the malicious local models to evade similarity-based defense methods.

Optimization Algorithm

We provide the pseudocodes of *CerP* in Algorithm 1. Multiple malicious participants introduce the backdoor trigger designated by the adversary into local training data. The malicious participants jointly optimize the *CerP*'s attack objective by fine-tuning the backdoor trigger and suppressing the trigger-induced local model biases, as defined in Eq.1. The fine-tuning trigger stage utilizes only the poisoned training data hosted by the malicious participants. Both the poisoned and poison-free local models are aggregated at the central server to produce a poisoned global model.

Experimental Evaluation

We evaluate the attack performance of the distributed backdoor attack methods using 3 benchmark datasets of different application scenarios. We implement all the involved algorithms using PyTorch on an Ubuntu workstation with NVIDIA 3090 GPUs. Our code can be found at the link [1].

[1] https://github.com/xtlyu/CerP

Algorithm 1: Cerberus Poisoning

Input: The global model g and the local model h_i of malicious participant i. The set of malicious participants S.

Output: The finely tuned backdoor trigger $\Delta x^{*,t}$ and the backdoor poisoned local model $h_i^{*,t}$.

for $t = E \rightarrow T$ **do**
  The central server sends the global model g^{t-1} to all the selected participants.
  for $all\ participants\ i \in S\ in\ parallel$ **do**
    $h_i^t \leftarrow g^{t-1}$;
    The adversary uses the data instances hosted by the malicious participants based on the model h_i^t to optimize the trigger $\Delta x^{*,t}$ by Eq. 6;
    Patch $x, x \in D_i^{mal}$ with finely tuned backdoor trigger $\Delta x^{*,t}$;
    Calculate $h_i^{*,t}$ by optimizing Eq.1 with D_i^{mal} and D_i^{nor};
    Send the poisoned local model $h_i^{*,t}$ to the server.
  end for
end for

Datasets, triggers, models, and hyperparameters. We evaluate *CerP* on 3 large-scale benchmark datasets: the applications of image classification (*CIFAR-100* (Krizhevsky, Hinton et al. 2009) and *Fashion-MNIST* (Xiao, Rasul, and Vollgraf 2017)), and the loan/credit risk assessment (*LOAN* (George 2020)). On each dataset, we adopt the setting of non-IID data distribution. The datasets, hyperparameters, and model structures are summarized in Table 1.

Baseline backdoor attacks. To organize a comparative study, the 3 distributed backdoor attacks (*LIE* (Baruch, Baruch, and Goldberg 2019), *Sybil attack* (Fung, Yoon, and Beschastnikh 2018), and *DBA* (Xie et al. 2020)) are involved.

Defense methods. We study the attack performance under the 13 defense methods: *Trimmed mean* (Yin et al. 2018), *Median* (Yin et al. 2018), *Krum* (Blanchard et al. 2017), *MKrum* (Blanchard et al. 2017), *Bulyan* (Mhamdi, Guerraoui, and Rouault 2018), *RFA* (Pillutla, Kakade, and Harchaoui 2019), *DnC* (Shejwalkar and Houmansadr 2021), *FLTrust* (Cao et al. 2021), *Foolsgold* (Fung, Yoon, and Beschastnikh 2018), *CRFL* (Xie et al. 2021), *FedCDP* (Geyer, Klein, and Nabi 2017), *FedLDP* (Wei et al. 2021), and *FL-WBC* (Sun et al. 2021).

Attack settings. To define a *partial-participation FL setting*, the global server selects 20 out of 100 participants on CIFAR-100 and Fashion-MNIST and 80 participants on LOAN respectively for parameter aggregation in each federated training iteration. During the backdoor attack, in each attack iteration, we constraint no more than 4 participants involved in aggregation as malicious participants. We apply this setting to make our studied attack scenario compatible with the feasibility assumption of Byzantine-robust FL algorithms, in order to organize a fair comparison. These defenses are assumed to work in the case where no more than 50% or 25% of the participants are malicious. For the LOAN, CIFAR-100, and Fashion-MNIST datasets, the adversary starts to attack from the **10**-th, **10**-th, and **70**-th training iterations of FL, respectively.

Dataset	Instances	Features	Model	Benign l_r	$\|S\|$	N	Poison l_r	Poison Ratio r	α	β
CIFAR-100	60,000	1024	Resnet-18	0.01	4	100	0.005	5/64	0.0001	0.0001
Fashion-MNIST	60,000	784	2 conv and 2 fc	0.1	4	100	0.05	5/64	0.0001	0.0001
LOAN	887,380	53	3 fc	0.001	3	80	0.0005	5/64	0.0001	0.0001

Table 1: Dataset, model structure, and hyperparameter description.

Attack / Defense	Sybil		DBA		LIE		CerP	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Krum	61.15	89.03	69.41	0.07	70.42	0.98	63.03	**90.76**
MKrum	65.24	88.08	66.01	88.17	63.91	2.57	64.03	**99.34**
Bulyan	68.69	87.48	69.29	5.59	66.80	37.67	68.11	**99.74**
Trimmed mean	66.05	56.72	64.78	3.85	69.77	9.39	66.84	**90.24**
Median	67.23	60.25	65.64	2.49	70.23	10.36	67.68	**92.25**
RFA	70.60	73.39	70.34	11.60	70.86	6.61	70.53	**81.80**
Foolsgold	69.28	72.73	69.98	0.44	69.97	0.92	69.29	**91.48**
FLTrust	71.20	77.32	71.53	63.55	71.34	5.52	71.30	**94.58**
DnC	62.64	92.38	62.89	73.36	70.24	4.90	66.19	**97.21**
FedLDP	70.91	73.01	70.89	73.50	70.86	8.94	71.19	**93.67**
FedCDP	70.35	85.16	70.36	71.42	70.14	66.58	70.68	**95.52**
CRFL	70.71	71.88	71.14	73.74	70.74	11.18	70.59	**93.66**
FL-WBC+Median	67.17	60.10	32.26	0.44	70.32	10.31	67.67	**92.13**
FL-WBC+Trim	66.22	56.58	34.93	0.72	69.81	7.37	66.94	**89.72**

Table 2: ASR and ACC of different distributed backdoor attacks on CIFAR-100(%).

Evaluation metrics. We involve two popularly used benchmarks *ACC* and *ASR* (Xie et al. 2020) to measure the attack performance of backdoor attack methods. *ASR* denotes the attack success rates, measuring the classification accuracy of the derived poisoned global model on the poisoned testing data. In parallel, *ACC* measures the main task's classification accuracy on the poisoned-free testing data.

Attack Performance

We compare *CerP* with 3 state-of-the-art distributed backdoor attacks against 13 defense methods. We show the *ACC* and *ASR* values of all the backdoor attacks involved in the comparison. We compare the *ASR* values of different backdoor attacks at the same *ACC* level. Due to the space limit, we show the results on CIFAR-100 and LOAN in this section. In the following tables, we use the bolded fonts to highlight the highest ASR values obtained among all the attack methods facing various defense methods. As seen in Tables 2–3, *CerP* can achieve higher *ASR* values than other baseline attack methods given the same *ACC* level and against all the deployed defense methods.

LIE is proved to conceal the robust aggregation methods under the IID data assumption. However, with the non-IID data setting in our study, we can find that *LIE* cannot bypass most defense methods. The *ASR* of *LIE* varies significantly across different datasets. Moreover, *LIE* requires knowing the model parameters committed by benign participants, which violates the protocol of FL. The empirical results show that the lack of the context knowledge of benign participants brings a noticeable drop to the ASR value of *LIE*. For example, *LIE* can hardly bypass any defense methods except *FedCDP* on CIFAR-100. The results show the limitations of *LIE* for general federated learning.

The decentralized nature of *DBA* facilitates defeating both the robust aggregation methods and *Foolsgold*. However, *DBA* cannot bypass most defense methods on CIFAR-100 and LOAN. The main reason is that *DBA* uses local triggers manually separated from the global trigger. Manually splitting the global trigger into chunks can bring unexpected artifacts to the learning of the backdoor poisoned data, which leads to large deviations of the poisoned local models and make it easy to be flagged by the Byzantine-robust FL methods. For *Sybil attack*, malicious participants submit similar local model updates, which makes it mitigated easily by *Foolsgold*. The ASR of *Sybil attack* against *Foolsgold* is less than 1% on LOAN. In addition, *Sybil attack* does not consider suppressing the deviations between the malicious and benign local models. Therefore, it is difficult for *Sybil attack* to bypass robust aggregation methods.

Attack Effectiveness

The finely tuned trigger. According to the previous theoretical analysis, fine-tuning the trigger is dedicated to reshaping the backdoor trigger to suppress the trigger-induced local model biases and facilitate the learning of the poisoned data to achieve stealthy backdoor attacks. To illustrate the impact of fine-tuning triggers, we compare the attack performance of our proposed method using only the original triggers (denoted as *CerP-NT*) and using the finely tuned triggers (*CerP*). The results in Table 4 show a significant increase in attack performance by simply integrating the trigger tuning module into the *CerP* attack, compared to the *CerP-NT* attack with the original triggers. For example, *CerP* can achieve ASR values higher than 90% against *Krum* and *Bulyan* on LOAN. On the contrary, *CerP-NT* cannot attack successfully, and the ASR values are only 0%.

Ablation study. To better understand the parameter sensitivity of *CerP*, we alternately set one of α (*CerP-ND*) and β

Defense \ Attack	Sybil		DBA		LIE		CerP	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Krum	90.99	0	90.83	0	90.78	0	91.16	**99.98**
MKrum	91.17	0	91.17	0	91.10	0	91.16	**99.98**
Bulyan	91.15	0	91.12	0	91.06	0	91.11	**99.98**
Trimmed mean	91.15	0	91.13	0	91.09	0	91.15	**99.98**
Median	91.18	0	91.16	0	91.10	0	91.19	**99.99**
RFA	92.34	100	92.38	100	92.06	98.97	92.22	**100**
Foolsgold	90.95	0	91.05	0.85	90.95	0	90.80	**99.97**
FLTrust	92.37	100	92.37	100	92.12	99.98	92.31	**100**
DnC	91.38	99.99	91.35	99.98	91.08	0	91.32	**100**
FedLDP	92.30	100	92.53	100	92.05	99.80	92.26	**100**
FedCDP	92.53	100	92.60	100	92.17	99.98	92.31	**100**
CRFL	92.07	100	91.83	100	91.64	99.79	91.66	**100**
FL-WBC+Median	91.19	0	91.16	0	91.10	0	91.14	**99.99**
FL-WBC+Trim	91.15	0	91.13	0	91.10	0	91.11	**99.98**

Table 3: ASR and ACC of different distributed backdoor attacks on LOAN(%).

Attack \ Defense	Krum	MKrum	Bulyan	Trim	Median	Fools	FLtrust	DnC	FedLDP	FL-WBC +Median	FL-WBC +Trim
Fashion-MNIST											
CerP-NT	64.67↓	70.36↓	73.01↓	86.98↓	95.90↓	99.55↓	99.47↓	99.61↓	96.72	82.10↓	87.04↓
CerP-ND	62.46↓	90.75↓	85.16↓	90.67↓	97.63	99.83	99.63↓	99.72↓	96.10↓	91.43↓	91.33↓
CerP-NS	94.75	90.53↓	85.75	91.07↓	97.63	99.80↓	99.65↓	99.71↓	96.16↓	91.53↓	91.37
LOAN											
CerP-NT	0↓	0↓	0↓	0↓	0↓	0↓	100	99.99↓	100	0↓	0↓
CerP-ND	99.98	99.98	99.98	99.98	99.99	99.96↓	100	100	100	99.99	99.98
CerP-NS	99.98	99.99	99.98	99.98	99.99	99.96↓	100	100	100	99.99	99.98

Table 4: ASR of CerP-NT, CerP-NS, and CerP-ND.

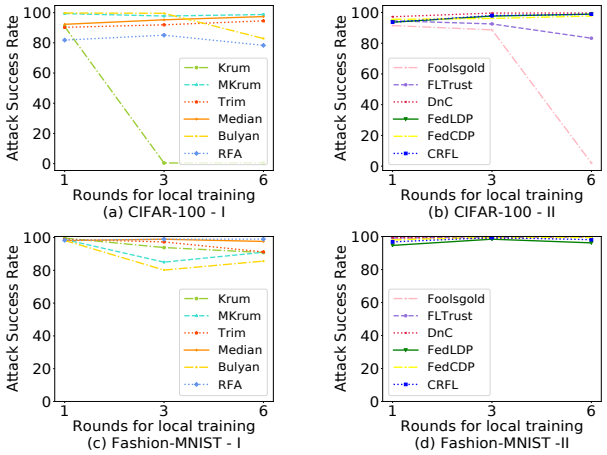

Figure 1: ASR of *CerP* on CIFAR-100 and Fashion-MNIST with different number of the poison local training rounds.

(*CerP-NS*) to 0, while keeping the other unchanged in Eq.1. Parameter α controls the regularization strength over the model deviation. As seen in Table 4, the deviation module is important to bypass the Byzantine robust aggregation methods. Removing it induces a drop in the ASR values of our attack when exposed to multiple defense methods. Removing the regularization of similarity ($\beta = 0$) causes *CerP*'s ASR against *Foolsgold* to drop.

Impact of the poison local training rounds. Figure 1 shows the ASR value changes of *CerP* on Fashion-MNIST and CIFAR-100 when the number of the poison local training rounds increases from 1 to 6. Intuitively, more poison local training rounds can lead to better attack performance. However, when the number of the poison local training rounds is too large, the attack performance of our attack degrades. One possible reason is that too many poison local training rounds make the malicious local models deviate too much from the benign local models, resulting in the malicious local model being identified as an abnormal model by the defense methods.

Concluding Remarks

In this work, we establish theoretical and empirical studies on the feasibility of organizing stealthy yet effective backdoor attacks on FL against defense methods. Our study explicitly unveils the key factors deciding the magnitude of malicious local model biases in general federated learning tasks. Instantiating the theoretical discussion, we propose a unified and highly flexible optimization framework *Cerberus Poisoning (CerP)* to coordinate effective backdoor attacks even with various defense methods deployed, which conducts the fine-tuning of the backdoor triggers and regularizes the trigger-induced local model bias. Substantial experimental results show that the *CerP* attack demonstrates an effective and stealthy backdoor poisoning threat to FL.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China, uder Grant U22B2027, and by the Open Project of the National Engineering Laboratory for Comprehensive Transportation Big Data Application Technology, under Grant No. 2022SK01-A.

References

Andreina, S.; Marson, G. A.; Möllering, H.; and Karame, G. 2021. BaFFLe: Backdoor Detection via Feedback-based Federated Learning. In *ICDCS 2021*, 852–863. IEEE.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How To Backdoor Federated Learning. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, 2938–2948. PMLR.

Baruch, G.; Baruch, M.; and Goldberg, Y. 2019. A Little Is Enough: Circumventing Defenses For Distributed Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8632–8642.

Blanchard, P.; Mhamdi, E. M. E.; Guerraoui, R.; and Stainer, J. 2017. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 119–129.

Cao, X.; Fang, M.; Liu, J.; and Gong, N. Z. 2021. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In *NDSS 2021*.

Cao, X.; Jia, J.; and Gong, N. Z. 2021. Provably Secure Federated Learning against Malicious Clients. In *AAAI 2021*, 6885–6893.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. In *ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, 1310–1320. PMLR.

Fang, M.; Cao, X.; Jia, J.; and Gong, N. Z. 2020. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, 1605–1622. USENIX Association.

Fu, S.; Xie, C.; Li, B.; and Chen, Q. 2019. Attack-Resistant Federated Learning with Residual-based Reweighting. *CoRR*, abs/1912.11464.

Fung, C.; Yoon, C. J. M.; and Beschastnikh, I. 2018. Mitigating Sybils in Federated Learning Poisoning. *CoRR*, abs/1808.04866.

Fung, C.; Yoon, C. J. M.; and Beschastnikh, I. 2020. The Limitations of Federated Learning in Sybil Settings. In *RAID 2020*, 301–316. USENIX Association.

George, N. 2020. Lending club loan data. https://www.kaggle.com/datasets/wordsforthewise/lending-club.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially Private Federated Learning: A Client Level Perspective. *CoRR*, abs/1712.07557.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Mhamdi, E. M. E.; Guerraoui, R.; and Rouault, S. 2018. The Hidden Vulnerability of Distributed Learning in Byzantium. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 3518–3527. PMLR.

Pillutla, V. K.; Kakade, S. M.; and Harchaoui, Z. 2019. Robust Aggregation for Federated Learning. *CoRR*, abs/1912.13445.

Shejwalkar, V.; and Houmansadr, A. 2021. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In *NDSS 2021*.

Sun, J.; Li, A.; DiValentin, L.; Hassanzadeh, A.; Chen, Y.; and Li, H. 2021. FL-WBC: Enhancing Robustness against Model Poisoning Attacks in Federated Learning from a Client Perspective. In *NeurIPS 2021*, 12613–12624.

Sun, Z.; Kairouz, P.; Suresh, A. T.; and McMahan, H. B. 2019. Can You Really Backdoor Federated Learning? arXiv:1911.07963.

Vershynin, R. 2011. Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027.

Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; Sohn, J.; Lee, K.; and Papailiopoulos, D. S. 2020. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Wei, K.; Li, J.; Ding, M.; Ma, C.; Su, H.; Zhang, B.; and Poor, H. V. 2021. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *CoRR*, abs/1708.07747.

Xie, C.; Chen, M.; Chen, P.-Y.; and Li, B. 2021. CRFL: Certifiably Robust Federated Learning against Backdoor Attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 11372–11382. PMLR.

Xie, C.; Huang, K.; Chen, P.; and Li, B. 2020. DBA: Distributed Backdoor Attacks against Federated Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yin, D.; Chen, Y.; Ramchandran, K.; and Bartlett, P. L. 2018. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*,

volume 80 of *Proceedings of Machine Learning Research*, 5636–5645. PMLR.