A Experimental settings

A.1 IID and non-IID Datasets

Following previous work [4, 5], in different datasets for classification tasks, if there are C classes, we construct C groups. The probability of a training sample with label k assigned to group k is $q > 0$ and to other group is $\frac{1-q}{C-1}$. Each client randomly chooses one group and samples a certain number of samples uniformly from this group for its local dataset. The probability q controls the degree of non-IID. If $q = \frac{1}{C}$, the dataset for each client is IID. For the root dataset, we randomly sample a certain number of samples from the training dataset with a bias probability p whose meaning is similar to q and then split the rest for each client in the way mentioned above.

We use Fashion-MNIST, CIFAR-10, and CIFAR-100 dataset [7] as the evaluation datasets, which have been widely used in prior FL studies. The Fashion-MNIST dataset consists of a 60K image samples training set and a 10K samples test set, where each image is associated with a label from 10 classes. The CIFAR-10 dataset composes 50K training images and 10K test images in 10 classes. They are colorful with three channels, and have varying object scales. Similar to CIFAR-10, the CIFAR-100 dataset comprises 100 classes with 600 images in each class, totaling 500 training images and 100 testing images per class. For all three datasets, we use $q = 0.5$ to conduct the non-IID experiments as [5]. By setting $q = 0.1$, we also conduct the IID experiments in §B.1 as [13].

A.2 Details for baseline attacks

The experiments cover untargeted and targeted attacks, including the following baselines:

- **Fang attack** [5]: Fang attack maximizes the deviation of the global model towards the opposite direction of the unattacked global model update. By using Trim and Krum as the aggregation algorithm in the attack, we obtain two different attacks: **Fang-Trim** and **Fang-Krum**.
- **Label-flipped (LF)** [14]: Following the same setting as Fang [5], we flip label k to $C-k-1$ for each sample on malicious clients, where $k \in \{0, \ldots, C-1\}$ and C denotes the total number of dataset classes.
- **AGR-agnostic** [13], *i.e.*, Aggregation algorithm agnostic attack, aims to limit the range of malicious updates and maximize the perturbation. It has two variants: **Min-Max attack** restricts the range based on the maximum value of the L-2 distance between any pair of benign model updates. **Min-Sum attack** is based on the minimum value of the sum of squares of the distances between any benign model update and all other benign model updates.
- **Scaling attack** [1]: Scaling attack is a targeted backdoor attack. It poisons the local dataset with trigger-embedded samples and scales up the model updates before sending them to the server. We use the same pattern trigger in LIE [2] and follow the same data augmentation scheme as FLTrust [4].

– **DBA** [16]: DBA disassembles a global trigger pattern into discrete local patterns, distributing them across diverse attacker training sets, thereby diminishing the noticeable distinction between benign and malicious gradients.
– **3DFed** [8]: 3DFed is a framework for covert FL backdoor attacks in a black-box setting, featuring three evasion modules: constrained loss backdoor training, noise mask, and decoy model. It implants indicators into backdoor models to obtain feedback from the global model and dynamically adjusts hyper-parameters for evasion modules based on this feedback.

A.3 Details for baseline defense

– **FLTrust** [4] trains a root model with a small root dataset, and takes the cosine similarity between a local model and the root model as the weight in aggregation.
– **Trimmed-Mean (Trim)** [17] are coordinate-wise aggregation rules assuming the number of malicious clients n_m is known, remove the parameters of the smallest and largest n_m values, and average the remaining.
– **Krum** [3] assumes the number of malicious clients is known as n_m. The server computes a score for each model update according to the distance between it and the $n - n_m - 2$ nearest ones. The one with the smallest score is selected as the global model update.
– **Tolpegin defense** [14] identifies malicious clients by standardizing model updates, employing the principal component analysis (PCA) for dimension reduction, and detecting them through clustering.
– **FLDetector** [18] employs consistency analysis on model updates to identify malicious clients. It predicts a client's model update based on historical data and compares it with the received update. Inconsistencies indicate potential malice.
– **DeepSight** [12] identifies malicious clients through the analysis of their cosine distances, Normalized Update Energies and Division Differences. Subsequently, it clips the aggregated gradient to enhance defense.
– **FLAME** [11] detects malicious clients through cosine distance analysis and subsequently utilizes a customized weak DP approach, integrating noise boundary proofing and dynamic clipping bound.

A.4 Hyperparameters

We set both global model learning rate α and local model learning rate β to 0.5 in all experiments. For mask learning rate γ, we set 0.5 in all three datasets. The default binarization threshold τ is 0.5. In CIFAR-10 and CIFAR-100, we allow the system to execute server-client communication for 500 and 1000 rounds, respectively. In each communication round, each client executes local training for $l = 5$ iterations. While in Fashion-MNIST, the system runs $2,500$ communication rounds, and each client only does a single local training iteration per round.

A.5 Details for real-world experiment setting in Section 1

In a swarm of 100 FL clients, we compromise 20 as malicious clients to apply Min-Sum attack. The system executes server-client communication for $2,500$ rounds and each client does a single local training per communication round. We train a four-layer CNN model on Fashion-MNIST dataset with the non-IID setting described in §A.1. We compare our SkyMask with a model-level defense, Tolpegin defense [14], which analyzes the PCA dimension reduction and clustering results of the local model updates for detecting malicious client.

B Additional experiment results

B.1 The Defense Effectiveness of SkyMask

We perform IID experiments as [13], to demonstrate that SkyMask is applicable to various data distributions, providing supplementary evidence for §4.2.

Table 1: FL testing accuracy under different attacks and attack success rates of targeted attack. The experimental results of targeted attack are in the form of "testing accuracy/attack success rate".

Dataset (Model)	Attack	FedAvg	FLTrust	Trim	Krum	DeepSight	FLAME	SkyMask-NR	SkyMask
Fashion -MNIST (CNN)	None	**0.90**	0.89	0.89	0.86	0.89	**0.90**	0.89	**0.90**
	LF	0.81	0.88	0.84	0.86	0.85	0.85	0.89	**0.90**
	Min-Max	0.63	0.89	0.76	0.86	0.73	0.75	**0.90**	**0.90**
	Min-Sum	0.85	0.89	0.78	0.55	0.82	0.86	**0.90**	**0.90**
	Fang-Trim	0.53	0.89	0.74	0.86	0.76	0.82	**0.90**	**0.90**
	Fang-Krum	0.88	0.89	0.86	0.26	0.86	0.87	**0.90**	**0.90**
	Scaling	0.83/0.21	0.89/0.10	0.88/0.14	0.89/0.11	0.90/0.17	0.90/0.11	0.89/0.10	**0.90/0.10**
	DBA	0.87/0.73	0.90/0.20	0.89/0.36	0.85/0.12	0.90/0.17	0.90/0.18	**0.90/0.10**	**0.90/0.10**
CIFAR-10 (ResNet20)	None	**0.79**	0.78	0.78	0.63	0.78	**0.79**	**0.79**	**0.79**
	LF	0.76	0.76	0.74	0.64	0.77	0.75	0.77	**0.78**
	Min-Max	0.68	0.61	0.58	0.56	0.64	0.70	**0.79**	**0.79**
	Min-Sum	0.73	0.75	0.62	0.30	0.67	0.71	**0.79**	**0.79**
	Fang-Trim	0.10	0.70	0.38	0.62	0.45	0.48	**0.78**	**0.78**
	Fang-Krum	0.68	0.77	0.60	0.42	0.55	0.56	**0.79**	**0.79**
	Scaling	0.10/1.00	0.76/0.12	0.75/0.67	0.62/0.10	0.76/0.13	0.76/0.12	0.78/0.11	**0.79/0.10**
	DBA	0.72/1.00	0.78/0.89	0.77/0.98	0.24/0.10	0.78/0.14	0.77/0.14	0.78/0.11	**0.78/0.10**
CIFAR-100 (ResNet20)	None	**0.48**	**0.48**	0.47	0.33	**0.48**	**0.48**	**0.48**	**0.48**
	LF	0.44	**0.48**	0.44	0.33	0.47	0.47	**0.48**	**0.48**
	Min-Max	0.33	0.44	0.25	0.22	0.31	0.33	**0.48**	**0.48**
	Min-Sum	0.41	0.47	0.34	0.11	0.40	0.39	**0.48**	**0.48**
	Fang-Trim	0.10	0.46	0.13	0.28	0.15	0.14	0.47	**0.48**
	Fang-Krum	0.07	0.45	0.10	0.03	0.06	0.10	**0.48**	**0.48**
	Scaling	0.42/1.00	0.47/0.43	0.46/1.00	0.33/0.01	0.48/0.97	0.48/0.94	0.48/0.02	**0.48/0.01**
	DBA	0.43/0.98	0.48/0.86	0.48/0.96	0.28/0.02	0.49/0.94	0.48/0.94	**0.48/0.01**	**0.48/0.01**

Under a low fraction of attacks. Table 1 illustrates that in IID case, across all attacks, SkyMask not only achieves the highest testing accuracy but also attains accuracy levels equivalent to the result of FedAvg under no attacks. While FLTrust demonstrates similar results on Fashion-MNIST dataset, its defense on CIFAR-10 and CIFAR-100 datasets is notably weaker, particularly against Min-Max attack. Additionally, FLTrust exhibits a high attack success rate under DBA, indicating vulnerability to targeted attacks. All other defense methods have poor performance under the fine-grained attacks.

Under no attack. In IID case, as shown in Table 1, the impact of the existing robust aggregation algorithms on the convergence prediction accuracy under no attacks are not as noticeable as in the non-IID case. The lesser data heterogeneity within the IID local datasets is attributed to the decreased overreaction induced by data heterogeneity across all defenses within this context. With the exception of Krum, all the other defenses lose less than 1% in accuracy. Our SkyMask still performs well with the highest testing accuracy.

Table 2: Testing accuracy, FPR and FNR of different malicious client detection methods under different attacks. The experimental results of targeted attack are in the form of "testing accuracy/attack success rate".

Dataset (Model)	Attack	Testing accuracy			FPR			FNR		
		Tolpegin	FLDetector	**SkyMask**	Tolpegin	FLDetector	**SkyMask**	Tolpegin	FLDetector	**SkyMask**
Fashion-MNIST (CNN) non-IID	None	0.89	0.86	0.89	/	/	/	/	/	/
	LF	0.89	0.86	0.89	0.11%	0.05%	0.00%	0.38%	0.00%	0.02%
	Min-Max	0.88	0.08	0.89	0.38%	96.94%	0.27%	0.80%	100.00%	0.40%
	Min-Sum	0.64	0.38	0.89	41.10%	100%	0.26%	84.80%	100.00%	0.00%
	Fang-Trim	0.88	0.86	0.89	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.85	0.10	0.89	28.20%	99.46%	0.20%	60.40%	100.00%	0.00%
	Scaling	0.89/0.10	0.89/0.10	0.89/0.10	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%
	DBA	0.89/0.10	0.89/0.10	0.89/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fashion-MNIST (CNN) IID	None	0.9	0.88	0.9	/	/	/	/	/	/
	LF	0.89	0.87	0.9	0.07%	0.02%	0.00%	0.00%	0.00%	0.00%
	Min-Max	0.89	0.87	0.9	0.06%	0%	0.00%	0.00%	0.81%	0.00%
	Min-Sum	0.88	0.49	0.9	15.34%	100%	0.01%	7.60%	100.00%	0.00%
	Fang-Trim	0.9	0.88	0.9	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.89	0.13	0.9	8.04%	99.60%	0.19%	6.40%	100.00%	0.00%
	Scaling	0.89/0.11	0.9/0.11	0.9/0.10	0.00%	0.00%	0.00%	0.68%	0.00%	0.00%
	DBA	0.90/0.8	0.90/0.9	0.90/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CIFAR-10 (ResNet20) non-IID	None	0.76	0.75	0.76	/	/	/	/	/	/
	LF	0.7	0.7	0.76	13.10%	0.00%	4.72%	19.90%	100.00%	2.60%
	Min-Max	0.61	0.11	0.77	36.50%	100%	0.00%	88.00%	100.00%	0.00%
	Min-Sum	0.59	0.31	0.77	38.80%	100%	0.00%	78.00%	100.00%	0.00%
	Fang-Trim	0.76	0.74	0.76	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.17	0.31	0.77	37.40%	87.18%	0.00%	84.00%	100.00%	0.00%
	Scaling	0.76/0.10	0.74/0.10	0.77/0.09	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%
	DBA	0.65/0.47	0.77/0.10	0.77/0.10	0.00%	0.00%	0.00%	46.10%	0.00%	0.00%
CIFAR-10 (ResNet20) IID	None	0.75	0.79	0.79	/	/	/	/	/	/
	LF	0.73	0.75	0.78	3.45%	0.00%	3.24%	5.80%	4.26%	0.00%
	Min-Max	0.74	0.25	0.79	0.05%	100%	0.00%	0.00%	100.00%	0.00%
	Min-Sum	0.74	0.31	0.79	0.14%	100%	0.00%	1.00%	100.00%	0.00%
	Fang-Trim	0.74	0.77	0.78	0.03%	0.00%	0.00%	0.00%	0.05%	0.00%
	Fang-Krum	0.74	0.18	0.79	4.04%	94.60%	0.00%	10.00%	100.00%	0.00%
	Scaling	0.79/0.09	0.78/0.10	0.79/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	DBA	0.73/0.10	0.79/0.10	0.79/0.10	0.00%	0.00%	0.00%	2.55%	0.00%	0.00%
CIFAR-100 (ResNet20) non-IID	None	0.28	0.44	0.44	/	/	/	/	/	/
	LF	0.27	0.36	0.44	27.66%	0.14%	1.28%	25.00%	98.97%	0.00%
	Min-Max	0.27	0.1	0.44	3.05%	100.00%	0.00%	8.00%	100.00%	0.00%
	Min-Sum	0.2	0.02	0.44	45.37%	100%	0.00%	66.00%	100.00%	0.00%
	Fang-Trim	0.44	0.44	0.44	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.02	0.05	0.44	27.39%	16.49%	0.00%	66.00%	100.00%	0.00%
	Scaling	0.44/0.01	0.44/0.01	0.44/0.01	0.02%	0.00%	0.00%	0.75%	0.00%	0.00%
	DBA	0.44/0.25	0.44/0.01	0.44/0.01	0.00%	0.00%	0.00%	17.40%	0.00%	0.00%
CIFAR-100 (ResNet20) IID	None	0.36	0.48	0.48	/	/	/	/	/	/
	LF	0.35	0.47	0.47	1.04%	0.00%	0.03%	2.63%	0.00%	2.17%
	Min-Max	0.36	0.02	0.48	0.36%	100%	0.00%	1.00%	100%	0.00%
	Min-Sum	0.36	0.04	0.48	1.80%	100%	0.00%	1.50%	100%	0.00%
	Fang-Trim	0.35	0.47	0.48	1.04%	0.00%	0.00%	2.00%	0.00%	0.00%
	Fang-Krum	0.34	0.03	0.48	4.11%	78.35%	0.00%	7.00%	100%	0.00%
	Scaling	0.36/0.02	0.48/0.01	0.48/0.01	0.00%	0.00%	0.00%	0.20%	0.00%	0.00%
	DBA	0.35/0.03	0.48/0.01	0.48/0.01	0.24%	0.00%	0.00%	9.32%	0.00%	0.00%

Table 3: The impact of the total number of clients. (The first line in the table is the testing accuracy result of FedAvg under no attack.)

Dataset (Model)	Attack	Testing accuracy		FPR		FNR	
		500	1000	500	1000	500	1000
Fashion -MNIST (CNN) non-IID	**FedAvg**	0.89	0.89	/	/	/	/
	None	0.89	0.89	/	/	/	/
	LF	0.89	0.89	0.40%	0.40%	0.94%	0.62%
	Min-Max	0.89	0.89	0.40%	0.30%	0.00%	0.00%
	Min-Sum	0.89	0.89	0.40%	1.60%	0.00%	0.00%
	Fang-Trim	0.89	0.89	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.89	0.89	0.60%	1.20%	0.00%	0.00%
	Scaling	0.89 / 0.11	0.89 / 0.10	0.00%	0.00%	0.00%	0.00%
	DBA	0.89 / 0.11	0.89 / 0.10	0.00%	0.00%	0.00%	0.00%

B.2 The Significance of Learnable Masks

Due to the page limitation, we only show the tabular for the experimental results on non-IID CIFAR-10 dataset in §4.3. Here we present more experiment results on different datasets in both IID and non-IID cases.

In the Fashion-MNIST part of Table 2, we can see that it is hard for Tolpegin defense to detect malicious clients stably, especially under Fang-Krum attack and Min-Sum attack. Its FPR and FNR are pretty high, so it fails to defend against attacks effectively and loses a lot of benign information. Quite the opposite, SkyMask perfectly completes the detection task and achieves an FNR of less than 4‰, which means that the server hardly misses any malicious client, and it can eliminate the harm of attacks. SkyMask also achieves an FPR of less than 3‰, so it can preserve the most benign clients and better characterize the global data distribution.

On CIFAR-10 and CIFAR-100 datasets, the fine-grained attacks consistently evade model-level malicious client detection algorithms, *i.e.*, Tolpegin defense and FLDetector. Particularly under Min-Max attacks in the non-IID case, the FNR of Tolpegin defense and FLDetector reach 88% and 100% respectively, indicating that almost all malicious updates elude their defenses. SkyMask achieves a FNR of less than 3% under LF attack, while maintaining a 0% FNR under all other attacks.

B.3 SkyMask's Scalability

To demonstrate the scalability of SkyMask, in addition to the experiments on CIFAR-10 dataset in §4.4, we conduct experiments utilizing a CNN global model trained on Fashion-MNIST. For Fashion-MNIST, we assessed the prediction accuracy of the main task, along with the FPR and FNR of SkyMask under various attack scenarios, encompassing 500 and 1000 clients.

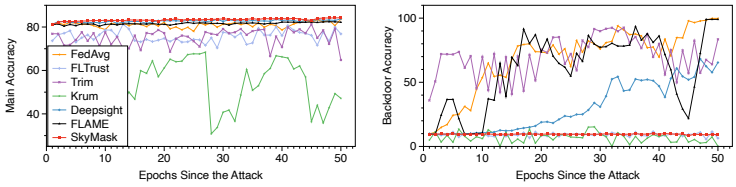
Table 3 shows that SkyMask can work well even if the number of clients increases to 1000. As the number of clients increases, the FNR of SkyMask is still lower than 1.6% and the FPR is lower than 1%. In addition, the missed malicious local models do not affect the performance of the global model. We

can see that the global model of SkyMask achieves the same testing accuracy as the unattacked global model's obtained by FedAvg.

B.4 The effectiveness against SOTA Backdoor Attack

Recently, 3DFed [8], an adaptive and extensible framework for backdoor attack gets a lot of attention. It selects specific parameters for poisoning and constrains backdoor training loss to improve the stealthiness of malicious model updates. Simultaneously, it generates decoy model updates to provide further cover for malicious model updates. Therefore, to demonstrate the effectiveness of SkyMask against the fine-grained backdoor attack, we conduct the experiments on CIFAR-10 dataset with the ResNet20 global model based on the official repository of 3DFed.[1] According to the experiment settings in [8], we set the starting epoch of the attack as $E_{start} = 150$, and the attack lasts for 50 rounds. The attacker compromises 20 of 100 clients. We record the main task testing accuracy and the backdoor accuracy at each epoch for each defense.

As shown in Fig. 1(b), only FLTrust, Krum, and our SkyMask keep the backdoor accuracy lower than 10%. When the attack concludes, the backdoor successfully evades the defenses of other methods and the backdoor accuracy is 77.47% for Trim, 99.6% for FLAME, and 99.8% for DeepSight. While in Fig. 1(a), we can see that Krum and FLTrust achieve defence at the cost of reduced main task testing accuracy. In contrast, our SkyMask maintains a high testing accuracy comparable to unattacked FedAvg's performance.



(a) Testing Accuracy (b) Backdoor Accuracy

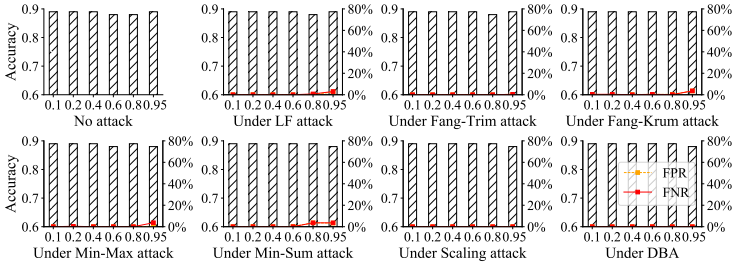
Fig. 1: Results of all defenses under 3DFed attack.

C Sensitivity analysis

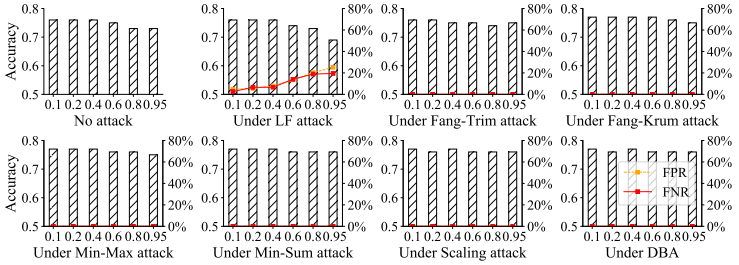
C.1 The Impact of Root Dataset Distribution

The server's root dataset is randomly sampled from the global data distribution by default. It is most likely to collect a root dataset with bias in the real world. However, we can prove that as soon as the bias is not too large ($p \leq 0.8$), our

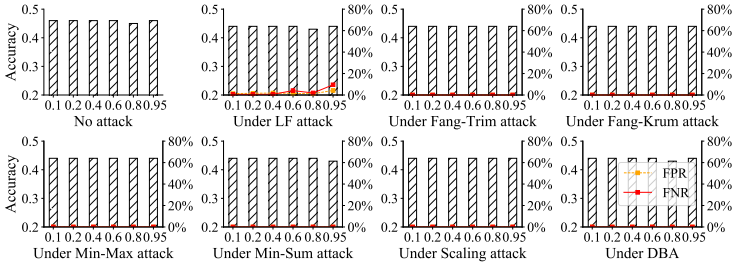
[1] https://github.com/haoyangliASTAPLE/3DFed



(a) On FashionMNIST Dataset (non-IID) with CNN



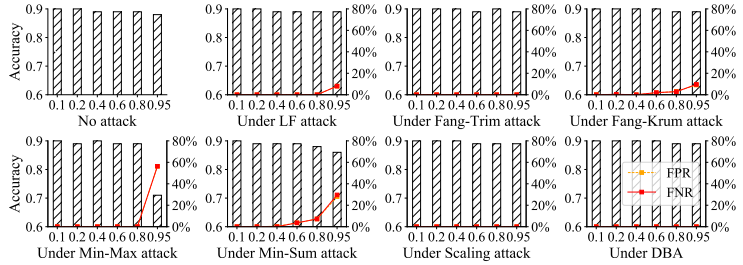
(b) On CIFAR-10 Dataset (non-IID) with ResNet20



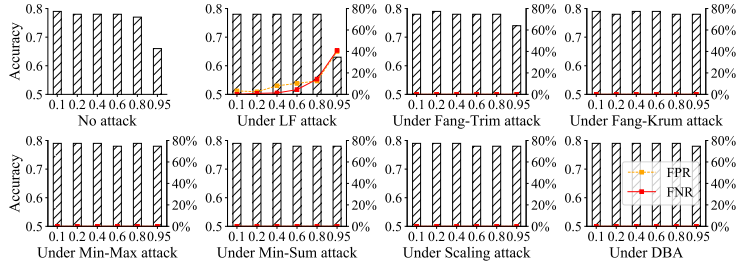
(c) On CIFAR-100 Dataset (non-IID) with ResNet20

Fig. 2: The impact of root data distribution on SkyMask. (Part I)

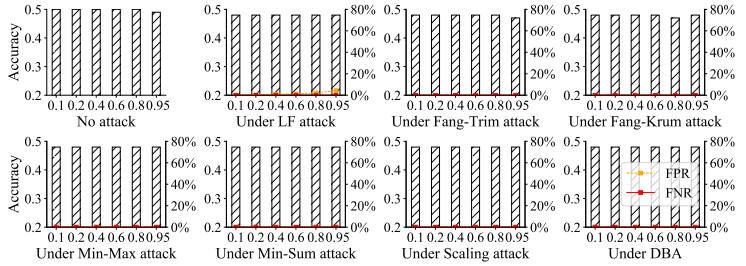
method can still take advantage of the root dataset and complete the malicious client detection task. We experiment on Fashion-MNIST using the CNN global model and on CIFAR-10 and CIFAR-100 using the Resnet20 model. We consider both IID and non-IID cases for local datasets as described in §A.1. The results are as shown in Fig. 2 and Fig. 3: when the bias probability of the root dataset ≤ 0.8, all the testing results of the global model obtained by our method are very close or equal to the model obtained by FedAvg under no attack. Only in some cases when the bias probability is 0.95, such as Fig. 3(a) under Min-Max attack, and Fig. 3(b) under no attack and LF attack, there is a drop in testing accuracy. Considering that the root dataset has a total of 100 samples, the lack of some classes of samples in the root dataset is possible when the bias probability of the root dataset is 0.95, so this result is reasonable.



(a) On FashionMNIST Dataset (IID) with CNN



(b) On CIFAR-10 Dataset (IID) with ResNet20



(c) On CIFAR-100 Dataset (IID) with ResNet20

Fig. 3: The impact of root data distribution on SkyMask. (Part II)

The FPR and FNR results in the figures show that the malicious client detection accuracy of SkyMask fluctuates as the degree of non-IID of the root data changes but remains very low. Our method does not fail until the bias probability of the root dataset up to 95%. As long as the generated root dataset has a bias not so significant, SkyMask can complete its malicious client detection task and keep the global model a good performance.

C.2 The Impact of Thresholds in Binarization

As the description of the processing of binary masks in §3.2, the server should choose a threshold τ. Our optimization strategy makes the parameters in masks converge to 0 or 1, so SkyMask should not be sensitive to threshold τ. We

Table 4: The impact of threshold in binarization on SkyMask.

Dataset (Model)	Attack	Acc.					FPR					FNR				
		0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7	0.3	0.4	0.5	0.6	0.7
Fashion -MNIST (CNN) non-IID	None	0.89	0.89	0.89	0.89	0.89	/	/	/	/	/	/	/	/	/	/
	LF	0.88	0.88	0.89	0.89	0.88	0.01%	0.02%	0.00%	0.03%	0.32%	0.08%	0.02%	0.02%	0.00%	0.04%
	Min-Max	0.89	0.89	0.89	0.89	0.89	0.00%	0.00%	0.27%	0.00%	0.00%	0.00%	0.00%	0.40%	0.00%	0.00%
	Min-Sum	0.88	0.89	0.89	0.89	0.88	0.04%	0.20%	0.26%	0.34%	0.89%	0.00%	0.40%	0.00%	0.00%	1.60%
	Fang-Trim	0.88	0.89	0.89	0.88	0.88	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.89	0.89	0.89	0.89	0.89	0.18%	0.28%	0.20%	0.40%	0.28%	0.40%	0.00%	0.00%	0.40%	0.00%
	Scaling	0.89/0.10	0.89/0.12	0.89/0.10	0.89/0.10	0.89/0.11	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	DBA	0.89/0.10	0.89/0.10	0.89/0.10	0.89/0.10	0.89/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Fashion -MNIST (CNN) IID	None	0.88	0.89	0.90	0.89	0.89	/	/	/	/	/	/	/	/	/	/
	LF	0.89	0.89	0.90	0.90	0.89	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Min-Max	0.89	0.90	0.90	0.90	0.89	0.08%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
	Min-Sum	0.89	0.89	0.90	0.90	0.89	0.04%	0.05%	0.01%	0.01%	0.34%	0.13%	0.00%	0.00%	0.00%	1.00%
	Fang-Trim	0.89	0.90	0.90	0.90	0.90	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.89	0.89	0.90	0.90	0.89	0.77%	0.56%	0.19%	0.05%	0.07%	1.00%	0.20%	0.00%	0.00%	0.20%
	Scaling	0.89/0.12	0.90/0.11	0.90/0.10	0.90/0.11	0.90/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	DBA	0.90/0.10	0.90/0.11	0.90/0.10	0.90/0.11	0.90/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CIFAR-10 (ResNet20) non-IID	None	0.76	0.76	0.76	0.76	0.76	/	/	/	/	/	/	/	/	/	/
	LF	0.76	0.76	0.76	0.76	0.76	4.72%	4.33%	2.03%	2.40%	6.65%	5.70%	2.80%	2.60%	4.00%	6.30%
	Min-Max	0.77	0.77	0.77	0.77	0.77	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Min-Sum	0.77	0.77	0.77	0.77	0.77	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Trim	0.76	0.77	0.76	0.76	0.76	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.77	0.77	0.77	0.77	0.77	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Scaling	0.76/0.09	0.76/0.12	0.77/0.12	0.77/0.10	0.77/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	DBA	0.77/0.10	0.76/0.10	0.77/0.10	0.77/0.11	0.77/0.10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CIFAR-10 (ResNet20) IID	None	0.79	0.78	0.79	0.78	0.78	/	/	/	/	/	/	/	/	/	/
	LF	0.76	0.78	0.78	0.79	0.78	5.31%	3.40%	3.24%	1.25%	8.04%	5.45%	1.75%	0.00%	4.45%	8.45%
	Min-Max	0.78	0.79	0.79	0.79	0.78	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Min-Sum	0.78	0.79	0.79	0.79	0.79	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Trim	0.78	0.79	0.78	0.78	0.78	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Fang-Krum	0.78	0.79	0.79	0.78	0.78	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Scaling	0.79/0.11	0.79/0.10	0.79/0.10	0.79/0.09	0.79/0.11	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	DBA	0.79/0.10	0.79/0.11	0.79/0.10	0.79/0.10	0.79/0.11	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

conduct experiments with different thresholds from 0.3 to 0.7, and the results are shown in Table 4.

On Fashion-MNIST dataset, it can be seen that SkyMask loses a little accuracy under some attacks when the threshold goes far away from 0.5. Still, it also has a similar accuracy as unattacked FedAvg's level, with about 1% accuracy loss. SkyMask achieves FNR lower than 1.6%, so there are almost no undetected malicious clients; FPR is lower than 0.9%, so most benign clients are retained.

On CIFAR-10 dataset, especially under LF attack, FPR and FNR increase as the threshold deviates from 0.5. But they are still less than 9%, and the testing accuracy loss is less than 1%. And for HAR, we can see similar results that SkyMask keeps a very low FPR and FNR, achieving the same testing accuracy as unattacked FedAvg. Therefore, the change in threshold does not affect our system's performance, and SkyMask is not sensitive to the threshold in binarization.

C.3 The Impact of Dimension Reduction

The clustering results of SkyMask are influenced by the dimensionality of the features after dimensionality reduction. In Fig. 4, we can see that when the number of features after dimensionality reduction rises from 2 to 15, although the testing accuracy does not decrease severely in most cases, there is an increase in the FPR and FNR, especially under LF attack. The reason is that the framework uses the Gaussian mixture model to cluster the dimension-reduced masks. Since the total number of masks is not large enough, the clustering results may not be satisfying in high-dimension space due to the curse of dimensionality. Therefore, we set the number of features after dimensionality reduction as two empirically.

D Discussion and Limitations

The current Byzantine attacks are launched with either a single client or multiple clients executing the same attack strategy, which may only account for an increased proportion of the model aggregation stage on the server, but the malicious clients do not collaborate. Increasingly, some researchers find that some specific layers are more vulnerable to Byzantine attacks [15]. We also observe that some Byzantine attacks modify parameters near the output layer significantly more than those far from the output layer.

Therefore, each client can be responsible for a different part of model poisoning, using the philosophy of "not putting eggs in the same basket." The attacker hides poisoning factors in all the uploaded models of the malicious clients it compromises. When most or all malicious model updates bypass the malicious client detection mechanism and get aggregated on the server, it can form a powerful attack on the global model, which is truly a collaborative attack by multiple malicious clients.

The damage to the characteristics of local model updates by this collaborative attack is unknown, and it remains to be verified whether SkyMask can efficiently detect these malicious attacks. For other malicious client detection algorithms, detecting this abnormal pattern in a single model update may be challenging. Theoretically, using our mask mechanism, the collaborative impact of all malicious clients on the global model could be shown after masked aggregation so that the optimization process can change their corresponding masks' pattern. Actually, Fang attack and AGR-agnostic attack have attempted to attack collaboratively, which weakens the anomaly level of model updates uploaded by each malicious client but influences the performance of the global model when all attacks are combined. Hence, there is a chance for our SkyMask to detect this kind of malicious client.

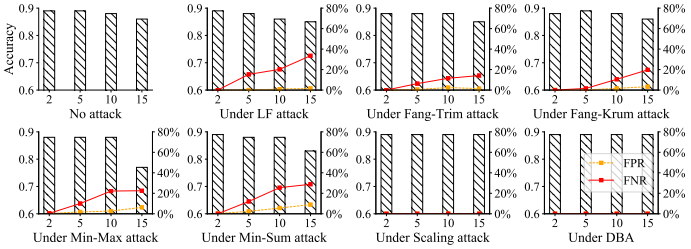
In the future, we will continue to study this topic in-depth, first verify the effectiveness of this collaborative attack, and deepen the ability of our learnable binary mask to extract the features of model updates and realize the defense capability for this unknown attack.

Furthermore, SkyMask can be seen as an additional module to the aggregation algorithm. Not only the basic FedAvg algorithm, the SOTA FL algorithms, such as FedProx [9], FedBN [10], or FedAMP [6], can also be equipped with our malicious client detection module. This feature lets the FL system retain the heterogeneity-handling, high convergence rate, or personalization features the original aggregation algorithm may have. It is also possible to enjoy the defensive capability of our malicious client detection strategy, and we will explore the merits of various combinations in future work.

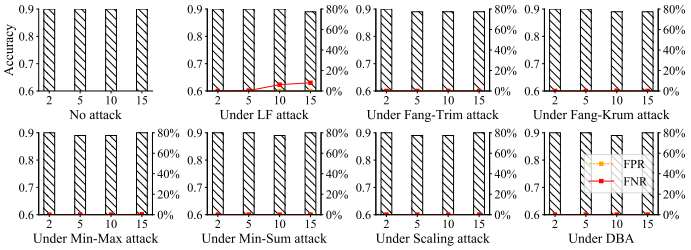
References

1. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to Backdoor Federated Learning. In: Proc. AISTATS (2020) 1

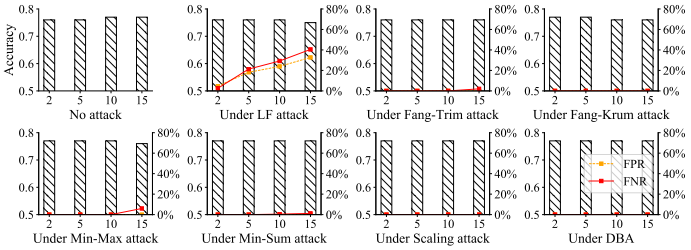
2. Baruch, G., Baruch, M., Goldberg, Y.: A Little is Enough: Circumventing Defenses for Distributed Learning. In: Proc. NeurIPS (2019) 1
3. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In: Proc. NeurIPS (2017) 2
4. Cao, X., Fang, M., Liu, J., Gong, N.Z.: FLTrust: Byzantine-Robust Federated Learning via Trust Bootstrapping. In: Proc. NDSS (2021) 1, 2
5. Fang, M., Cao, X., Jia, J., Gong, N.: Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In: Proc. USENIX Security (2020) 1
6. Huang, Y., Chu, L., Zhou, Z., Wang, L., Liu, J., Pei, J., Zhang, Y.: Personalized Cross-silo Federated Learning on Non-IID Data. In: Proc. AAAI (2021) 10
7. Krizhevsky, A., Hinton, G., et al.: Learning Multiple Layers of Features from Tiny Images. Technical Report (2009) 1
8. Li, H., Ye, Q., Hu, H., Li, J., Wang, L., Fang, C., Shi, J.: 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning. In: Proc. IEEE S&P (2023) 2, 6
9. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proc. MLSys (2020) 10
10. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: FedBN: Federated learning on non-IID features via local batch normalization. In: Proc. ICLR (2021) 10
11. Nguyen, T.D., Rieger, P., De Viti, R., Chen, H., Brandenburg, B.B., Yalame, H., Möllering, H., Fereidooni, H., Marchal, S., Miettinen, M., et al.: {FLAME}: Taming backdoors in federated learning. In: Proc. USENIX Security (2022) 2
12. Rieger, P., Nguyen, T.D., Miettinen, M., Sadeghi, A.R.: Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. In: Proc. NDSS (2022) 2
13. Shejwalkar, V., Houmansadr, A.: Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In: Proc. NDSS (2021) 1, 3
14. Tolpegin, V., Truex, S., Gursoy, M.E., Liu, L.: Data Poisoning Attacks against Federated Learning Systems. In: Proc. ESORICS (2020) 1, 2, 3
15. Varma, K., Zhou, Y., Baracaldo, N., Anwar, A.: LEGATO: A LayerwisE Gradient AggregaTiOn Algorithm for Mitigating Byzantine Attacks in Federated Learning. In: Proc. CLOUD (2021) 10
16. Xie, C., Huang, K., Chen, P.Y., Li, B.: DBA: Distributed Backdoor Attacks against Federated Learning. In: Proc. ICLR (2019) 2
17. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In: Proc. ICML (2018) 2
18. Zhang, Z., Cao, X., Jia, J., Gong, N.Z.: Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In: Proc. KDD (2022) 2



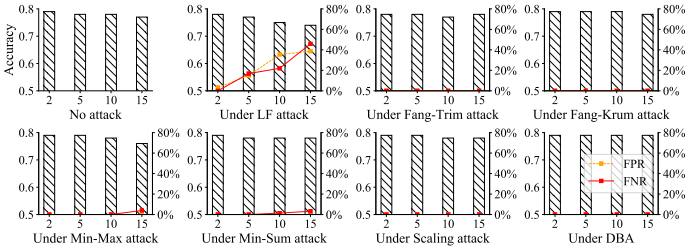
(a) On FashionMNIST Dataset (non-IID) with CNN



(b) On FashionMNIST Dataset (IID) with CNN



(c) On CIFAR-10 Dataset (non-IID) with ResNet20



(d) On CIFAR-10 Dataset (IID) with ResNet20

Fig. 4: The impact of dimensionality reduction on SkyMask.