FedCom: Byzantine-Robust Federated Learning Using Data Commitment

Bo Zhao
Nanjing University of Aeronautics and Astronautics
bozhao@nuaa.edu.cn

Tao Wang
Nanjing University of Aeronautics and Astronautics
wangtao1@nuaa.edu.cn

Liming Fang*
Nanjing University of Aeronautics and Astronautics Shenzhen Research Institute
fangliming@nuaa.edu.cn

ICC 2023 - IEEE International Conference on Communications | 978-1-5386-7462-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC45041.2023.10279255

***Abstract*—Federated learning is a promising distributed edge learning methodology that allows multiple clients to collaboratively train statistical models without disclosing private training data. However, there may exist Byzantine clients launching data/model poisoning attacks to compromise global model's performance or convergence. Most of the existing Byzantine-robust FL schemes are either ineffective against several advanced poisoning attacks, and their robustness suffers from further degradation when local datasets are highly non-independently and identically distributed (non-IID). To address these issues, we propose FedCom, which could achieve data/model poisoning tolerant FL under practical non-IID data partitions, even the attackers do not honestly follow FedCom's protocol. The cardinal design of FedCom is privacy-respectfully asking clients to generate the commitments, which commit and verify the honesty of their local data distributions or local model updates. An extensive performance evaluation demonstrates FedCom's superior performance compared to the state-of-the-art Byzantine-robust schemes under various rigorous settings, even the attackers do not honestly follow the protocol of FedCom and fabricate the commitments.**

***Index Terms*—federated learning, poisoning attack, Byzantine-robust**

I. Introduction

The recently emerging federated learning (FL) [1] [2] is a promising edge-network-empowered machine learning paradigm, which allows multiple clients to collaboratively train a global statistical model under the orchestration of a central server while keeping data localized. Such schemes could significantly alleviate the privacy concerns in smart large-scale computation infrastructures like AI-empowered healthcare systems [3], smart UAV networks [4], etc.

However, the absence of legal accesses to clients' private datasets and/or the local training processes allows Byzantine clients to launch data poisoning [5] [6] [7] or model poisoning [8] [9] attacks to compromise or manipulate the global model. Without effective defences against these attacks to ensure Byzantine-robustness in FL, the global model's performance or convergence would drastically deteriorate.

To address this issue, several Byzantine-robust FL schemes are proposed, where the server tries to identify the poisoning local model updates and excludes them from global model aggregation. Some are categorized [10] [11] [12] as Distance Statistical Aggregation (DSA), as such schemes remove geometrical outliers among received local models before global model aggregation. Others could be categorized [13] [14] as Contribution Statistical Aggregation (CSA), as such schemes give priority to local models which contribute the most to elevating global model's performance.

However, existing Byzantine-robust FL schemes have various limitations. First, the recently proposed attack [8] could generate poisoned local models geometrically close to benign models and could circumvent the DSA schemes. Second, the CSA methods require collecting a public validation dataset before the training starts, which violates the privacy principle in FL. Finally, most of the existing schemes are ineffective when data is non-independent and identically distributed (non-IID) among clients, which is usually of practical as different users bear heterogenous data distributions.

To address the issues mentioned above, in this paper, we propose FedCom, a novel Byzantine-robust federated learning framework, which achieves both data/model poisoning tolerant FL under practical non-IID data distribution among clients, even when attackers do not honestly follow the protocol of FedCom. Specifically, before initiating FL training process, each client is firstly required to privacy-securely generate a commitment to its local training data distribution, which is considered as a crafted dataset by server that is subject to the differential-privately (DP) perturbed data distribution of the original local training dataset, without disclosing any original data. Then, the central server collects these commitments, compares the Wasserstein distances among them, and lowers the scores of ones with outlying distances. Finally, the server evaluates each local model on their corresponding commitment dataset, and lowers the score of ones with obvious asynchronous loss degradations. The final score is determined by the multiplication of the two scores, and local models with lower final scores are excluded from model aggregation.

In summary, the major contributions of this paper are summarized as follows:

*Liming Fang is the corresponding author.

This work is supported in part by the National Key R&D Program of China under Grant 2021YFB3100700, in part by the National Natural Science Foundation of China under Grant 62272228 and U22B2029, in part by the Key R&D Program of Jiangsu Province under Grant ZD201913, and in part by Shenzhen Science and Technology Program under Grant JCYJ20210324134408023.

978-1-5386-7462-8/23/$31.00 ©2023 IEEE

- To the best of our knowledge, this is a novel lightweight Byzantine-robust FL framework that provides a fundamental methodology of defending against both data/model poisoning attacks with non-IID local datasets.
- We identify poisoning local datasets as ones generating outlying commitments referring to Wasserstein distance between each other.
- We identify local models compromising commitments as ones with significantly asynchronous convergence processes on corresponding commitments.
- We make extensive comparisons between FedCom and the state-of-the-art robust FL schemes defending against various data/model poisoning attacks on two non-IID datasets. The results demonstrate the superior performances of FedCom, even with attackers who undermine FedCom's protocol by fabricating benign commitments.

II. Preliminaries and Related Works

1) Federated Learning: Federated learning (FL) allows multiple clients to collaboratively train a machine learning model without disclosing any private data. FL requires a central server to coordinate clients' training processes. The vanilla FL system (i.e., `FedAverage`) follows the steps below:

Step 1: the central server distributes the latest global model W to each client.

Step 2: the ith client updates W on local datasets to obtain local model w_i, and submit w_i to central server.

Step 3: central server obtain the latest global model via weighted averaging. Formally, $W = \sum_{i=0,1,\ldots n} \frac{|d_i|}{\sum_{j=0,1,\ldots n} |d_j|} w_i$, where $|d_i|$ is the size of local dataset ith client holds.

2) Poisoning Attack: In FL, poisoning attacks aim to manipulate global models via poisoning local training datasets (LDP) or local models (LMP) [15]. LDP has been much explored, which usually launch attack by injecting malicious samples into training dataset to train a compromised or manipulated global model [5], [16], [17]. LMP is not wildly explored until recently proposed work [8] gives novel idea of jamming the training process of FL, even if the aggregation rule is claimed to be Byzantine-robust.

3) Distance Statistical Aggregation: Some Byzantine-robust aggregation rules concretize such assumption: benign local model holds majority in all local models, and poisoning local models are the ones with outlying geometrical characteristics. `Krum` [10] firstly calculates the sum of Euclidean distances between w_i and the $n-f-1$ closest local models, n is the total number of clients while f is the upper bound of Byzantine clients number. `Krum` works if $f < (n-2)/2$. `Krum` will select the local model with lowest score as global model in one iteration. `Krum` always select local model which is the closest to the centroid of majority model cluster. Similar schemes including `Bulyan` [18], `Trimmed Mean`, and `Median` [11], which further remove outliers referring to each dimension of local models.

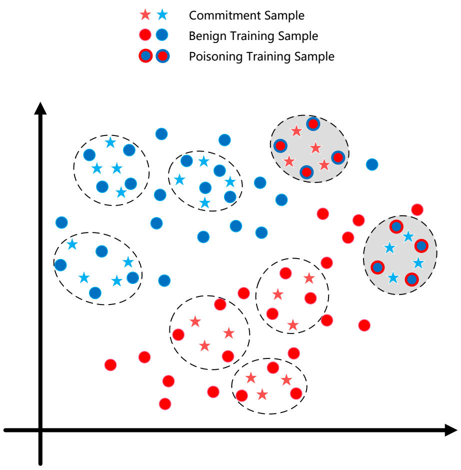


Fig. 1. Demonstration of data commitment.

However, local datasets in federated learning are usually non-ID, which breaks the assumption of schemes mentioned above, and introduce vulnerabilities. Fang et al [8] exploited such vulnerabilities, and successfully impact the robustness of all the schemes mentioned above.

4) Contribution Statistical Aggregation: There is also an assumption that benign local models could positively impact global model accuracy, or negatively impact global model loss on certain validation set. Such an assumption derives two types of aggregation rules: loss function based rejection (LFR) and error rate based rejection (ERR) [8]. Representative schemes adopting ideas above are `Zeno` [13] and `Zeno++` [14].

However, such schemes could be infeasible in federated learning, as forming such a public dataset by IID sampling violates the privacy principle of federated learning. Besides, the local models trained on highly non-IID local datasets might also obtain bad performance on public central datasets. Since their infeasibility in FL, DSA is out of the scope of our evaluation.

III. Design of FedCom

A. Overview of FedCom

The workflow of FedCom is described as follows.

- **Step 1: Generating Data Commitments:** Before initiating FL training process, each client performs the nearest samples averaging on original local datasets under differential privacy mechanism to synthesize commitments.
- **Step 2: Scoring Data Commitments:** After collecting commitments from each client, the server compares Wasserstein distances between commitments, and score each of them referring to such distances.
- **Step 3: Scoring Local Models:** In each round of FL, the server evaluates each local model on their corresponding commitment datasets, and score them referring to the asynchronism of loss degradation.

- **Step 4: Local Model Aggregation:** Referring to above two scores, the local models are aggregated to update the global model.

B. Generating Data Commitment

The essences of various poisoning attacks are that Byzantine clients dishonestly train local models on its benign local dataset $D^{(i)}_{training}$. To detect such dishonesty, we ask $client_i$ to submit the description of $\mathbb{E}(D^{(i)}_{training})$, which we call *commitment*. Specifically, we define commitment $D^{(i)}_{commitment}$ is a dataset that subjects to a perturbed distribution of $D^{(i)}_{training}$ with differential privacy (DP) mechanism, and have no intersection with $D^{(i)}_{training}$.

Inspired by data mixup of FL [19], For each sample p_k in $D^{(i)}_{training}$, we firstly find the m nearest samples $\Gamma = \{p_{k+1}, p_{k+2}, ..., p_{k+m}\}$ to p_k. Then, take $c_j = \frac{1}{m}\sum_{p\in\Gamma} p$ as crafted sample c_k. Traversing all samples in $D^{(i)}_{training}$, and we could finally construct a set of crafted samples $\mathbb{C} = \{c_1, c_2, ..., c_{|D^{(i)}_{training}|}\}$ The final commitment dataset of $D^{(i)}_{training}$ is $\mathbb{C}$ under DP perturbation. Specifically,

$$D^{(i)}_{commitment} = \mathbb{C} + N(0, (\sigma^{(i)})^2) \tag{1}$$

In which N is the random noise subjects to normal distribution, $(\sigma^{(i)})^2$ is the intensity of the DP mechanism. We could guarantee that $D^{(i)}_{commitment}$ will not disclose either the original sample or the distribution of $D^{(i)}_{training}$, but is strongly related to it. The demonstration of data commitment is presented in Fig. 1.

C. Scoring Data Commitments

After receiving data commitments from all clients, the server is able to evaluate them to figure out potential poisoning local data distributions. Although local datasets are highly non-IID, which indicates significant divergences between local data distributions $\mathbb{E}(D^{(i)}_{training})$, but there is a reasonable assumption that the divergences between $\mathbb{E}(D^{(i)}_{training})$, which is approximately equivalent to $\mathbb{E}(D^{(i)}_{commitment})$, are bounded by a certain distribution.

Hence, to identify the outlying distributions, we first define $x^{(i)}_j$ as the set of jth dimension elements of $D^{(i)}_{commitment}$, $\overline{x^{(i)}_j}$ as the set of jth dimension elements of all other commitments server received. Then, we adopt the Wasserstein distance [20] between $x^{(i)}_j$ and $\overline{x^{(i)}_j}$ of $D^{(i)}_{commitment}$ to measure the difference between $D^{(i)}_{commitment}$ and other commitments on *j*th dimension. After determining the divergences on each dimension between all commitments, each local model w_i's *Data Credit* (DC) could be determined as

$$\begin{aligned} &DC_i = 1 - \tfrac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{d^{(i)}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \\ &d^{(i)} = \tfrac{1}{n}\sum_{j\in 1,2,...,n} Wass\left(x^{(i)}_j, \overline{x^{(i)}_j}\right) \end{aligned} \tag{2}$$

where $d^{(i)}$ is the total distribution divergence of $D^{(i)}_{commitment}$, and μ and σ are the mean and variance of d of each commitment.

D. Scoring Local Models

We evaluate each local model w_i by calculating the difference of its validation loss on its corresponding commitment $D^{(i)}_{commitment}$ between neighboring iterations of FL. Specifically, let $l_i = l^{before}_i - l^{after}_i$, in which l^{before}_i is w_i's loss on $D^{(i)}_{commitment}$ before a single round local training, while l^{after}_i is loss after a single round local training. l_i must be positive if hyperparameters are well-tuned, or we consider w_i is not honestly trained on $D^{(i)}_{training}$ in current round local training. We also make a reasonable assumption that under the same machine learning task with L-Lipschitz continuity, each local model's convergence should be approximately synchronous, and the differences between l_i should not be too large. Hence, in k-client FL system, each local model w_i's *Training Credit* (TC) could be determined as

$$TC_i = \left\{ \begin{array}{c} 0, l_i \leq 0 \\ \frac{1}{\sum_{l_j>0}|l_i - l_j|}, l_i > 0 \end{array} \right. \tag{3}$$

Obviously, w_i with a negative l will receive a TC score of zero, and such w_i will be abandoned by central server. l with larger distances to other ls brings a lower TC score to corresponding local model, as such a local model's local training progress may not synchronized to others.

E. Local Model Aggregation

Specifically, in k-client FL system, the final weight I_i of w_i in each round of FL is determined

$$I_i = \frac{Flag_i\left|D^{(i)}_{\text{training}}\right|}{\sum_{j=1,2,...,k} Flag_j\left|D^{(j)}_{\text{training}}\right|} \tag{4}$$

where

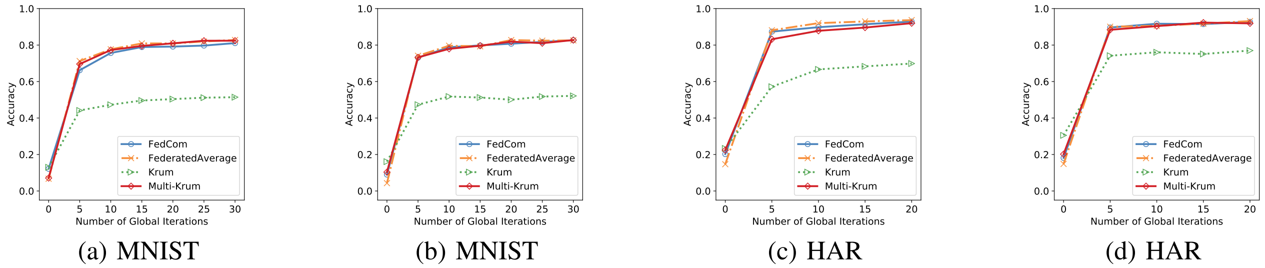
$$Flag_i = \left\{ \begin{array}{l} 0, DC_i * TC_i < \text{Median}(DC * TC) \\ 1, DC_i * TC_i \geq \text{Median}\,(DC * TC) \end{array} \right. \tag{5}$$

Under such settings, the attackers either committing poisoning local data distributions, or dishonestly behaving by fabricating benign distributions and submitting poisoning local models, will obtain low scores, which confine the poisoning local models' effectiveness. After determining each local model's final weight, the local models could be aggregated via `FedAverage` to update the global model.

IV. Evaluation

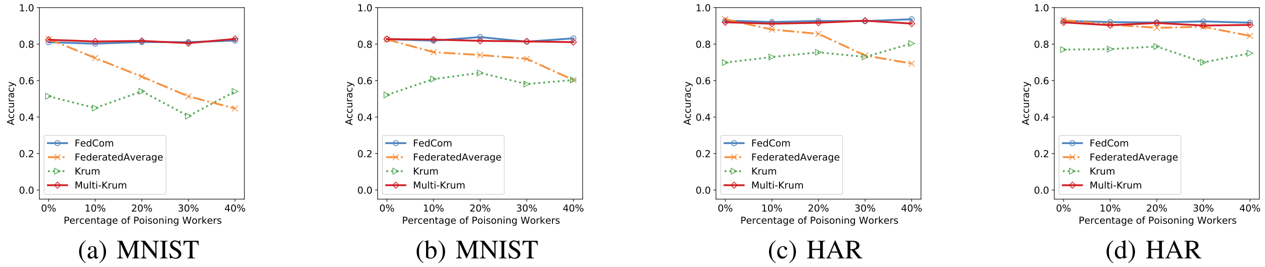
A. Experimental Setup

1) Datasets: We employ two non-IID datasets for performance evaluation, i.e., Human Activity Recognition (HAR) [21] and MNIST [22]. HAR is a dataset of motion data sampled from motion sensors of smart phones, partitioned by different users. MNIST is a dataset for handwritten digit recognition, each sample is a 28*28 gray-level image.



(a) MNIST (b) MNIST (c) HAR (d) HAR

Fig. 2. Results of model accuracy on different datasets under all aggregation rules when there is no attack. (a) and (c) are results of LR, while (b) and (d) are result of DNN.



(a) MNIST (b) MNIST (c) HAR (d) HAR

Fig. 3. Results of model accuracy on benign datasets under Gaussian attack. (a) and (c) are results of LR, while (b) and (d) are result of DNN.

2) Non-IID Settings: For HAR, it is naturally user-wise moderately non-IID, and we randomly select different users' datasets and distribute them to different clients. For MNIST, we follow the method in LotteryFL [22] to generate a more significantly non-IID data partition.

3) Models and Training Settings: We consider 20 clients in FL with both convex and non-convex models: multi-class logistic regression classifiers (LR) and deep neural networks (DNN) with 150 hidden neurons.

4) Baseline Aggregation Rules:

- **FedAverage**: the classic aggregation rule of FL, global model is aggregated by averaging local models with weights, in which weights are the ratio of local dataset's size and global dataset's size.
- **Krum [10]**: a representative a series of DSA schemes that elevates the score of local models with higher degree of outlying, and takes the model with lowest score as global model.
- ***Multi*-Krum [10]**: *Multi*-Krum adopt the same scoring rule as Krum, but take $n - f - 1$ local models with the lowest scores, and aggregate these local models via FedAverage.

Noting that we consider schemes of CSA are difficult to be implemented in FL, hence such schemes are not in the scope of our evaluation.

5) Evaluated Poisoning Attacks:

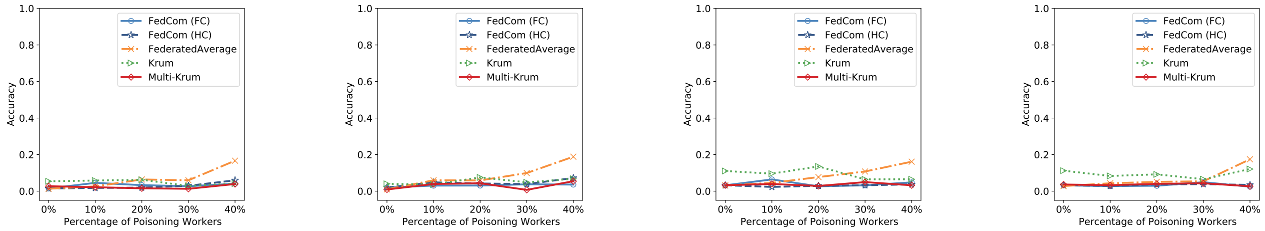
- **Label flipping**: a naive LDP attack by simply mislabelling training samples to undermine the original distribution of dataset.
- **Gaussian attack**: a naive LMP attack by submitting Gaussian noise as a local model, attempting to prevent global model's convergence.
- **Back-gradient** [6]: an advanced LDP attack by solving a reversed optimizing problem on training set, which could generate a poisoning training dataset that elevates trained model's loss on a clean validation dataset.
- **Krum attack** [8]: an advanced LMP attack aimed at Krum or similar aggregation rules. Such an attack attempts to construct a poisoning model that is the closest to other local models, but obtain reverse direction to correct global gradient.

6) Adversary Settings: For all aggregation rules in our experiment, we set the percentage of Byzantine clients from 0% (no attack) to 40%. Besides, for FedCom, we design *honest commitment* (HC) where Byzantine clients generate data commitment honestly on poisoning local datasets, and *fake commitment* (FC) where they generate fake commitments on clean datasets.

7) Performance Metrics: For LMP, we evaluate global model's accuracy on benign validation set. For LDP, we evaluate global model's accuracies on poisoning dataset. Either a higher accuracy on clean dataset, or a lower accuracy on poisoning dataset, indicates a higher performance of the global model.

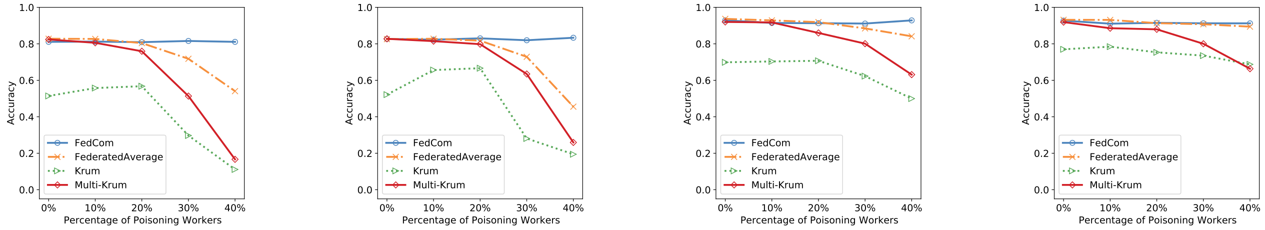
B. Convergence Performance of FedCom

Figure 2 shows the accuracies of global models after several global iterations, from which we could conclude that FedCom obtains similar convergence performance, as well as tolerance of non-IID local datasets, to aggregation rules like FedAverage or *Multi*-Krum.



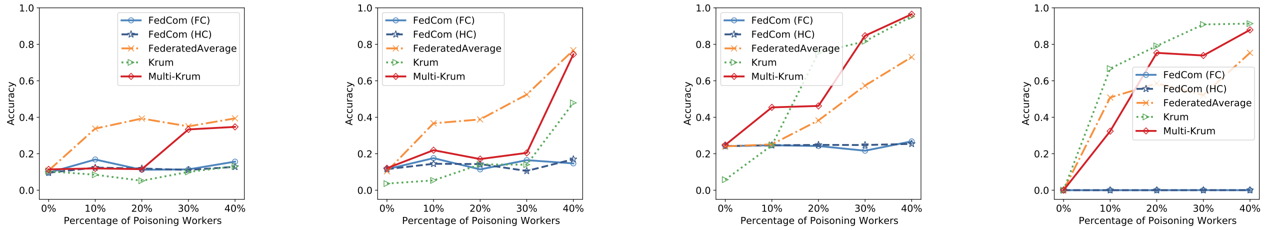
(a) MNIST (b) MNIST (c) HAR (d) HAR

Fig. 4. Results of model accuracy on poisoning datasets under Label-flipping attack. (a) and (c) are results of LR, while (b) and (d) are results of DNN.



(a) MNIST (b) MNIST (c) HAR (d) HAR

Fig. 5. Results of model accuracy on benign datasets under Krum attack. (a) and (c) are results of LR, while (b) and (d) are result of DNN.



(a) MNIST (b) MNIST (c) HAR (d) HAR

Fig. 6. Results of model accuracy on poisoning datasets under Back-gradient attack. (a) and (c) are results of LR , while (b) and (d) are results of DNN.

C. Results for Naive Poisoning Attacks

Figure 3 shows the results of Gaussian attack and Figure 4 shows the results of Label-flipping attack.

1) Robustness: We could summarize from the results that even if facing naive attacks, FedAverage is unable to effectively defend. By contrast, other three aggregation rules all perform various degree of robustness.

2) Impact of Non-IID: We could observe a larger benign accuracy decay or higher poisoning accuracy ascent under FedAverage on MNIST than on HAR. This indicates that a higher degree of non-IID makes attacks more effective. However, the non-IID degree trivially impacts the robustness of FedCom, *Multi*-Krum or Krum.

Through results above, we could summarize that FedCom obtains comparable robustness to *Multi*-Krum or Krum while facing naive attacks like Label-flipping attack or Gaussian attack, even the local datasets are non-IID.

D. Results for Advanced Poisoning Attacks

Figure 5 shows the results of Krum attack and Figure 6 shows the results of Back-gradient attack.

1) Krum Attack: We could easily summarize that Krum attack could significantly deteriorate other baselines' robustness, and a higher level of non-IID will amplify the negative impact. However, Krum attack is almost ineffective to FedCom. Hence, we could consider that FedCom obtains better robustness than schemes like *Multi*-Krum or Krum.

2) Back-Gradient Attack: The results indicates that such an attack could successfully disable other baselines, as global models under these aggregation rules obtains significant rises of accuracy on back-gradient poisoning datasets. However, the performances of global models under FedCom maintains a similar high level as when there is no attack.

Through this part of evaluation, we could conclude that, facing advanced poisoning attacks like Krum attack or Back-gradient attack, *Multi*-Krum or Krum obtain no robustness, but FedCom obtains well robustness to above attacks and tolerance to the effect of non-IID local datasets.

E. Summary

Through our evaluation, we could conclude that FedCom obtains broader robustness than representative Byzantine ro-

TABLE I
Comparison of Evaluated Aggregation Rules

	FedCom	Krum	*Multi*-Krum	Fed-Average
No attack	√	√	√	√
Gaussian	√	√	√	×
Label-flip	√	√	√	×
Krum	√	×	×	×
Back-gradient	√	×	×	×
Non-IID	√	×	×	×

bust aggregation rules. Moreover, FedCom's robustness will not decay with higher degree of non-IID. We summarize the contribution of FedCom into Table I.

V. Conclusion

In this paper, we proposed FedCom, a data commitment based Byzantine-robust FL framework which provides a fundamental defense against data poisoning and model poisoning attacks, even when local datasets are non-IID. By requiring each client to submit a data commitment, the proposed FedCom could defend against data poisoning attacks by comparing the Wasserstein distance among the data commitments, and thwart model poisoning attacks by evaluating the behavior of different local models using the corresponding data commitment. The exhaustive experimental results demonstrate that FedCom outperforms the state-of-the-art Byzantine-robust FL schemes in defending against typical data poisoning and model poisoning attacks under practical non-IID data distribution among clients.

References

[1] B. McMahan, E. Moore, and D. Ramage, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, ser. Proceedings of Machine Learning Research, A. Singh and X. J. Zhu, Eds., vol. 54. PMLR, 2017, pp. 1273–1282.

[2] P. Kairouz, H. B. McMahan, and B. Avent, "Advances and open problems in federated learning," *CoRR*, vol. abs/1912.04977, 2019.

[3] K. Kapadiya, U. Patel, and R. Gupta, "Blockchain and ai-empowered healthcare insurance fraud detection: an analysis, architecture, and future prospects," *IEEE Access*, vol. 10, pp. 79 606–79 627, 2022.

[4] D. Darsena, G. Gelli, and I. Iudice, "Detection and blind channel estimation for uav-aided wireless sensor networks in smart cities under mobile jamming attack," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11 932–11 950, 2022.

[5] M. Jagielski, A. Oprea, and B. Biggio, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*. IEEE Computer Society, 2018, pp. 19–35.

[6] L. Muñoz-González, B. Biggio, and A. Demontis, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, B. M. Thuraisingham, B. Biggio, and D. M. Freeman, Eds. ACM, 2017, pp. 27–38.

[7] S. Liu, S. Lu, and X. Chen, "Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 6282–6293.

[8] M. Fang, X. Cao, and J. Jia, "Local model poisoning attacks to byzantine-robust federated learning," in *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, S. Capkun and F. Roesner, Eds. USENIX Association, 2020, pp. 1605–1622.

[9] E. Bagdasaryan, A. Veit, and Y. Hua, "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]*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, 2020, pp. 2938–2948.

[10] P. Blanchard, E. M. E. Mhamdi, and R. Guerraoui, "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*, I. Guyon, U. von Luxburg, and S. Bengio, Eds., 2017, pp. 119–129.

[11] D. Yin, Y. Chen, and K. Ramchandran, "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*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5636–5645.

[12] L. Li, W. Xu, and T. Chen, "RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 1544–1551.

[13] C. Xie, S. Koyejo, and I. Gupta, "Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 6893–6901.

[14] ——, "Zeno++: Robust fully asynchronous SGD," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 10 495–10 503.

[15] M. Barreno, B. Nelson, and R. Sears, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS 2006, Taipei, Taiwan, March 21-24, 2006*, F. Lin, D. Lee, B. P. Lin, S. Shieh, and S. Jajodia, Eds. ACM, 2006, pp. 16–25.

[16] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[17] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *CoRR*, vol. abs/1708.06733, 2017.

[18] E. M. E. Mhamdi, R. Guerraoui, and S. Rouault, "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*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 3518–3527.

[19] T. Yoon, S. Shin, and S. J. Hwang, "Fedmix: Approximation of mixup under mean augmented federated learning," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[20] L. Weng, "From GAN to WGAN," *CoRR*, vol. abs/1904.08994, 2019.

[21] D. Anguita, A. Ghio, and L. Oneto, "A public domain dataset for human activity recognition using smartphones," in *21st European Symposium on Artificial Neural Networks, ESANN 2013, Bruges, Belgium, April 24-26, 2013*, 2013.

[22] A. Li, J. Sun, and B. Wang, "Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets," *CoRR*, vol. abs/2008.03371, 2020.