An investigation on the Defense system of Data Poisoning Attacks Against Federated Learning Systems

Adnan Rahman Eshan
School of Data and Sciences
Brac University
Dhaka, Bangladesh
adnan.rahman.eshan@g.bracu.ac.bd

***Abstract*—Federated learning has emerged as a promising approach for training machine learning models while preserving data privacy. However, the vulnerability of federated learning systems to data poisoning attacks poses a significant threat to their integrity and reliability. In this report, we critically analyze the defense system proposed by Tolpegin et al. for mitigating data poisoning attacks in federated learning. Through a systematic evaluation, we uncover vulnerabilities and limitations within the defense system, highlighting the challenges of effectively detecting and mitigating sophisticated adversarial strategies. Our findings emphasize the need for further research and development in the domain of defense mechanisms for federated learning systems, particularly in addressing adaptive attacks and ensuring robust detection of malicious data. We also stress the importance of collaborative efforts among researchers, practitioners, and industry stakeholders to collectively advance the security and trustworthiness of federated learning systems. By enhancing the resilience of defenses and fostering open discussions, we can strive towards the widespread adoption of federated learning while safeguarding data integrity, privacy, and model reliability.**

***Index Terms*—Federated Learning, Data Poisoning**

I. Introduction

In recent years, federated learning has been a good approach to preserve the data privacy of different models.FL systems allow global model training without the sharing of raw private data. Instead, individual participants only share model parameter updates. Consider a deep neural network (DNN) model. DNNs consist of multiple layers of nodes where each node is a basic functional unit with a corresponding set of parameters. Nodes receive input from the immediately preceding layer and send output to the following layer; with the first layer nodes receiving input from the training data and the final layer nodes generating the predictive result. However, as the popularity of federated learning grows, so does the concern for its vulnerability to adversarial attacks. Among these attacks, data poisoning has proven to be a significant threat to the integrity and reliability of federated learning systems.

Data poisoning is an attack on machine learning models wherein the attacker adds examples to the training set to manipulate the behavior of the model at test time. This paper explores poisoning attacks on neural nets. The proposed attacks use “clean-labels”; they don’t require the attacker to have any control over the labeling of training data. They are also targeted; they control the behavior of the classifier on a specific test instance without degrading overall classifier performance.

The paper titled ”Data Poisoning Attacks Against Federated Learning Systems” by Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu sheds light on the potential dangers posed by data poisoning attacks in the context of federated learning. It investigates the effectiveness of defense mechanisms employed to mitigate these attacks and highlights their limitations, emphasizing the urgent need for improved defenses.

The report is structured as follows: first, we will provide a brief overview of federated learning and data poisoning attacks to establish a foundation for understanding the problem. Next, we will summarize the defense system proposed by Tolpegin et al. and discuss its strengths and weaknesses. Subsequently, we will present a comprehensive analysis of the identified vulnerabilities and propose potential strategies to strengthen the defense against data poisoning attacks in federated learning systems.

Through this report, we hope to raise awareness about the pressing need for robust defenses against data poisoning attacks and contribute to the ongoing efforts in enhancing the security of federated learning. By identifying and addressing the weaknesses in existing defense systems, we aim to pave the way for the development of more resilient solutions that can withstand adversarial threats.

II. Related Works

The growing adoption of federated learning has prompted extensive research into the security and privacy challenges associated with this distributed learning paradigm. In particular, the threat of data poisoning attacks has gained significant attention, as adversaries can manipulate the integrity of the shared model by injecting malicious data during the training process. Several research studies have focused on understanding the vulnerabilities of federated learning systems to data poisoning attacks and proposing defense mechanisms to mitigate these threats. In this section, we discuss notable papers that have contributed to the understanding and defense against data poisoning attacks in federated learning.

One prominent study in this domain is the work by Bagdasaryan et al. [1], which introduced the concept of model inversion attacks in federated learning. The authors demonstrated that adversaries with access to the model updates can infer sensitive information from the aggregated model, thus compromising user privacy. While this study primarily focuses on privacy concerns, it highlights the need for robust defenses to ensure the integrity and reliability of the federated learning process.

Several poisoning attacks were developed for popular ML models including SVM [6,12,44,45,50,52], regression [19], dimensionality reduction [51], linear classifiers [12,23,57], unsupervised learning [7], and more recently, neural networks [12,30,42,45,53,58]. However, most of the existing work is concerned with poisoning ML models in the traditional setting where training data is first collected by a centralized party. In contrast, our work studies poisoning attacks in the context of FL. As a result, many of the poisoning attacks and defenses that were designed for traditional ML are not suitable to FL. For example, attacks that rely on crafting optimal poison instances by observing the training data distribution are inapplicable since the malicious FL participant may only access and modify the training data s/he holds. Similarly, server-side defenses that rely on filtering and eliminating poison instances through anomaly detection or k-NN [36,37] are inapplicable to FL since the server only observes parameter updates from FL participants, not their individual instances

Another relevant contribution is the research conducted by Bhagoji et al. [2], which investigated the vulnerability of federated learning to backdoor attacks. The authors demonstrated that malicious participants can inject poisoned models into the federated learning process, leading to compromised model performance and potential leakage of sensitive information. Their work emphasizes the importance of detecting and mitigating such attacks to maintain the trustworthiness of federated learning systems.

Furthermore, Shokri et al. [3] explored the possibility of membership inference attacks in federated learning, where an adversary aims to determine if a particular data sample was part of the training dataset. By analyzing the output of the shared model, adversaries can infer the presence or absence of specific data instances, raising concerns about data privacy. The study highlights the need for defenses that prevent such membership inference attacks and preserve the confidentiality of user data.

In addition to these studies, recent research by Yang et al. [4] delves into the domain of adaptive poisoning attacks in federated learning. The authors proposed a novel attack strategy where adversaries can adaptively inject poisoned data based on the knowledge gained during the training process. Their work sheds light on the dynamic nature of data poisoning attacks and emphasizes the importance of developing defense mechanisms that can adapt to evolving adversarial strategies.

Moreover,The rising popularity of FL has led to the investigation of different attacks in the context of FL, such as backdoor attacks [2,46], gradient leakage attacks [18,27,59] and membership inference attacks [31,47,48]. Most closely related to our work are poisoning attacks in FL. There are two types of poisoning attacks in FL: data poisoning and model poisoning. Our work falls under the data poisoning category. In data poisoning, a malicious FL participant manipulates their training data, e.g., by adding poison instances or adversarially changing existing.instances [16,43]. The local learning process is otherwise not modified. In model poisoning, the malicious FL participant modifies its learning process in order to create adversarial gradients and parameter updates. [4] and [14] demonstrated the possibility of causing high model error rates through targeted and untargeted model poisoning attacks. While model poisoning is also effective, data poisoning may be preferable or more convenient in certain scenarios, since it does not require adversarial tampering of model learning software on participant devices, it is efficient, and it allows for non-expert poisoning participants. Finally, FL poisoning attacks have connections to the concept of Byzantine threats, in which one or more participants in a distributed system fail or misbehave. In FL, Byzantine behavior was shown to lead to sub-optimal models or non-convergence [8,20]. This has spurred a line of work on Byzantine-resilient aggregation for distributed learning, such as Krum [8], Bulyan [28], trimmed mean, and coordinate-wise median [55]. While model poisoning may remain successful despite Byzantine-resilient aggregation [4,14,20], it is unclear whether optimal data poisoning attacks can be found to circumvent an individual Byzantineresilient scheme, or whether one data poisoning attack may circumvent multiple Byzantine-resilient schemes. We plan to investigate these issues in future work.

Building upon this existing body of work, the paper by Tolpegin et al. [5] specifically focuses on data poisoning attacks against federated learning systems and examines the effectiveness of defense mechanisms in mitigating these threats. By critically analyzing their proposed defense system, we aim to identify its limitations and contribute to the ongoing efforts in strengthening the security of federated learning against data poisoning attacks.

III. Research Methodology

The objective of the analysis of the paper is to assess the effectiveness of the defense system proposed by Tolpegin et al. in mitigating data poisoning attacks in federated learning.In this paper, the authors study the vulnerability of FL systems to malicious participants seeking to poison the globally trained model. They also made minimal assumptions on the capability of a malicious FL participant – each can only manipulate the raw training data on their device. This allows for non-expert malicious participants to achieve poisoning with no knowledge of model type, parameters, and FL process. Under this set of assumptions, label flipping attacks become a feasible strategy to implement data poisoning, attacks which have been shown to be effective against traditional, centralized ML models The authors also investigated their application to FL systems using complex deep neural network models.

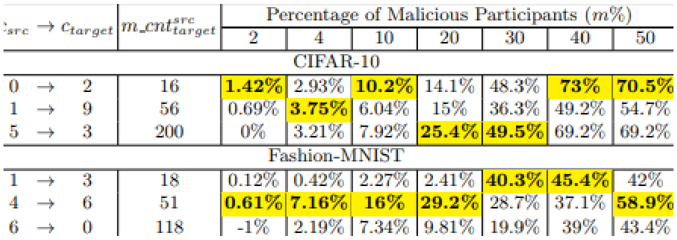
The authors implemented FL in Python using the PyTorch library. By default, they had N = 50 participants, one central aggregator, and k = 5.They used an independent and identically distributed data distribution or in other words they assumed the total training dataset is uniformly randomly distributed among all participants with each participant receiving a unique subset of the training data. The testing data is used for model evaluation only and is therefore not included in any participant's trained dataset .Both DNN models converge after fewer than 200 training rounds.Then the authors set their FL experiments to run for R = 200 rounds total.

In order to simulate the label flipping attack in a FL system with N participants of which meach experiment the authors randomly designated N × mas malicious. The rest are honest according to the paper. To address the impact of random selection of malicious participants, by default the authors repeat each experiment 10 times and report the average results. Unless otherwise stated, we use m = 10

Now in our investigation we,analyse the defense process.According to the authors,the parameter updates sent from malicious participants have unique characteristics because in the attack process they used a replacement method which replaces the parameters from N × m

IV. Results

Updated Attack: After the change of the replacement method,the vulnerability on the dataset remains the same.Like the authors,we start by investigating the feasibility of poisoning FL systems using label flipping attacks.. Results demonstrate that as the malicious participant percentage increases the global model utility (test accuracy) decreases. Even with small m, we observe a decrease in model accuracy compared to a non-poisoned model (denoted by MNP in the graphs), and there is an even larger decrease in source class recall. In experiments with CIFAR-10, once m reaches 40class decreases to 0



$_{src} \rightarrow c_{target}$	$m\_cnt_{target}^{src}$	Percentage of Malicious Participants (m%)						
		2	4	10	20	30	40	50
CIFAR-10								
0 → 2	16	**1.42%**	2.93%	**10.2%**	14.1%	48.3%	**73%**	**70.5%**
1 → 9	56	0.69%	**3.75%**	6.04%	15%	36.3%	49.2%	54.7%
5 → 3	200	0%	3.21%	7.92%	**25.4%**	**49.5%**	69.2%	69.2%
Fashion-MNIST								
1 → 3	18	0.12%	0.42%	2.27%	2.41%	**40.3%**	**45.4%**	42%
4 → 6	51	**0.61%**	**7.16%**	**16%**	**29.2%**	28.7%	37.1%	**58.9%**
6 → 0	118	-1%	2.19%	7.34%	9.81%	19.9%	39%	43.4%

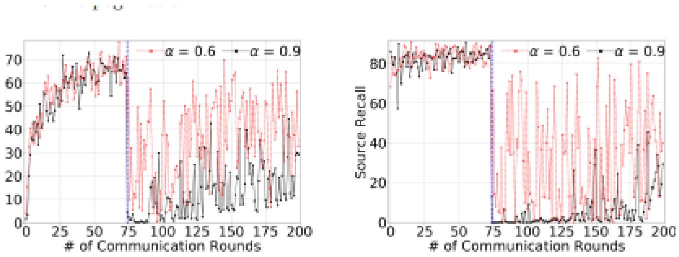
Fig. 1. Caption

While both datasets are vulnerable to the updated attack, the degree of vulnerability varies between datasets with CIFAR-10 demonstrating more vulnerability than Fashion-MNIST.On the other hand, vulnerability variation based on source and target class settings is less clear. In Table 2, we report the results of three different combinations of source → target attacks for each dataset. Consider the two extreme settings for the CIFAR-10 dataset: on the low end the 0 → 2 setting has a baseline misclassification count of 16 while the high end count is 200 for the 5 → 3 setting. Because of the DNN's relative challenge in differentiating class 5 from class 3 in the non-poisoned setting, it could be anticipated that conducting a label flipping attack within the 5 → 3 setting would result in the greatest impact on source class recall. However, this was not the case. Table 2 shows that in only two out of the six experimental scenarios did 5 → 3 record the largest drop in source class recall. In fact, four scenarios' results show the 0 → 2 setting, the setting with the lowest baseline misclassification count, as the most effective option for the adversary. Experiments with Fashion-MNIST show a similar trend, with label flipping attacks conducted in the 4 → 6 setting being the most successful rather than the 6 → 0 setting which has more than 2× the number of baseline misclassifications. These results indicate that identifying the most vulnerable source and target class combination may be a non-trivial task for the adversary, and that there is not necessarily a correlation between non-poisoned misclassification performance and attack effectiveness.

Attack Timing in the updated attacks: Like the label flipping attack,we consider two scenarios in which the adversary is restricted in the time in which they are able to make malicious participants available: one in which the adversary makes malicious participants available only before the 75th training round, and one in which malicious participants are available only after the 75th training round. As the rate of global model accuracy improvement decreases with both datasets by training round 75, we choose this point to highlight how pre-established model stability may effect an adversary's ability to launch an effective attack. similar to label flipping attack.

Malicious Participant availability: Given the impact of malicious participation in late training rounds on attack effectiveness, the authors introduced a malicious participant availability parameter . By varying the authors simulated the adversary's ability to control compromised participants' availability (i.e. ensuring connectivity or power access) at various points in training. Specifically, represents malicious participants' availability and therefore likelihood to be selected relative to honest participants. For example, if = 0.6, when selecting each participant Pi Pr for round r, there is a 0.6 probability that Pi will be one of the malicious participants. Larger implies higher likeliness of malicious participation. In cases where k ¿ N × mrecall by round when = 0.6 and = 0.9 for both the CIFAR-10 and FashionMNIST datasets. In both datasets, when malicious participants are available more frequently, the source class recall is effectively shifted lower in the graph, i.e., source class recall values with = 0.9 are often much smaller than those with = 0.6. We note that the high round-by-round variance in both graphs is due to the probabilistic variability in number of malicious participants in individual training rounds. When fewer malicious participants are selected in one training round relative to the previous round, source recall increases. When more malicious participants are selected in an individual round relative to the previous round, source recall falls. We further explore and illustrate our last remark with respect to the impact of malicious parties' participation in consecutive rounds

in consecutive rounds, i.e., (of malicious Pr) – (of malicious Pr1). The reported results are then averaged across multiple runs of FL and all cases in which each participation difference was observed. The results confirm our intuition that, when Pr contains.



(a) CIFAR-10

(b) Fashion-MNIST

Fig. 2. Caption

Defense System: The defense system which has been proposed in the paper has this insight: the parameter updates sent from malicious participants have unique characteristics compared to honest participants' updates for a subset of the parameter space.The description of the defense strategy can be described through the Algorithm . Let R denote the set of vulnerable FL training rounds and csrc be the class that is suspected to be the source class of a poisoning attack. We note that if csrc is unknown, the aggregator can defend against potential attacks such that csrc = c c C. We also note that for a given csrc, Algorithm 1 considers label flipping for all possible ctarget. An aggregator therefore will conduct —C— independent iterations of Algorithm 1, which can be conducted in parallel. For each round r R and participant Pi Pr, the aggregator computes the delta in participant's model update compared to the global model, i.e., ,i ← r,i r. Recall from Section 2.1 that a predicted probability for any given class c is computed by a specific node nc in the final layer DNN architecture. Given the aggregator's goal of defending against the label flipping attack from csrc, only the subset of the parameters in ,i corresponding to ncsrc is extracted. The outcome of the extraction is denoted by src ,i and added to a global list U built by the aggregator. After U is constructed across multiple rounds and participant deltas, it is standardized by removing the mean and scaling to unit variance. The standardized list U 0 is fed into Principal Component Analysis (PCA), which is a popular ML technique used for dimensionality reduction and pattern visualization. For ease of visualization, we use and plot results with two dimensions (two components). Data Poisoning Attacks Against Federated Learning SIn our updated attack we have tried to break that unique characteristic by mixing the characteristic.We have replaced 30 percent of 1 parameter with another which will give a different graphical representation comparing to the representation described in the paper.Unfortunately the code given for the defense in the paper has some technical glitch for which the graphical representation of the updated attack is not possible to show.

A. Discussion

Just like the authors approach,we conduct our attacks using two popular image classification datasets: CIFAR-10 [22] and Fashion-MNIST [49]. CIFAR10 consists of 60,000 color images in 10 object classes such as deer, airplane, and dog with 6,000 images included per class. The complete dataset is pre-divided into 50,000 training images and 10,000 test images. Fashion-MNIST consists of a training set of 60,000 images and a test set of 10,000 images. Each image in Fashion-MNIST is gray-scale and associated with one of 10 classes of clothing such as pullover, ankle boot, or bag. In experiments with CIFAR-10, we use a convolutional neural network with six convolutional layers, batch normalization, and two fully connected dense layers. This DNN architecture achieves a test accuracy of 79.90neural network with batch normalization, an architecture which achieves 91.75test accuracy in the centralized scenario without poisoning. Further details of the datasets and DNN model architectures can be found in Appendix A.

V. Conclusion and Future Work

In this report, we have conducted a comprehensive analysis of the defense system proposed by Tolpegin et al. against data poisoning attacks in federated learning systems. By critically examining the system's strengths and weaknesses, we aimed to identify its limitations and contribute to the ongoing efforts in strengthening the security of federated learning.

Through our evaluation, we have uncovered several vulnerabilities within the defense system. These weaknesses highlight the challenges of effectively detecting and mitigating data poisoning attacks in federated learning environments. The system's performance was found to be dependent on various factors, including the attack intensity, defense configuration, and dataset characteristics. While the defense system demonstrated promising results under certain scenarios, it showed limitations in more sophisticated attack settings and dynamic adversarial strategies.

The findings of our analysis emphasize the need for further research and development in the area of defense mechanisms against data poisoning attacks in federated learning systems. Improvements should focus on enhancing the system's resilience to adaptive attacks, minimizing the impact on model convergence, and ensuring robust detection of malicious data.

Additionally, it is crucial to explore the practical feasibility and scalability of the defense system in real-world federated learning deployments. Considerations such as computational overhead, communication efficiency, and privacy preservation should be carefully addressed to enable the adoption of these defenses in practical settings.

For the future work,we want to add to the defense of such sort of attacks in a more comprehensive manner .Moreover,the development of such sort of learning needs to be planned in a more specific so that they are not vulnerable to attacks.

Reference

,;:”’–’”:;,,;:”’–’”:;,. (2022, July 29). ,;:”’–’”:;,,;:”’–’”:;, - YouTube. Retrieved May 12, 2023, from https://spectrum.ieee.org/ai-cybersecurity-data-poisoning

Bhagoji, A. N. (n.d.). Analyzing Federated Learning through an Adversarial Lens. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:634-643, 2019, 1. https://proceedings.mlr.press/v97/bhagoji19a.html

Constantin, L. (2021, April 12). What is data poisoning? Attacks thatcorrupt machine learning models. CSO Online. Retrieved May 12, 2023, from https://www.csoonline.com/article/3613932/how-data-poisoning-attacks-corrupt-machine-learning-models.html

Lyu, L., Yang, Q. (n.d.). Threats to Federated Learning: A Survey. 1. https://arxiv.org/abs/2003.02133

Shen, Z., Shokri, R. (n.d.). Share Your Representation Only: Guaranteed Improvement of the Privacy-Utility Tradeoff in Federated Learning. ICLR 2023 Conference, 1. https://openreview.net/forum?id=oJpVVGXu9i

What is federated learning? (2022, August 24). IBM Research. Retrieved May 12, 2023, from https://research.ibm.com/blog/what-is-federated-learning

Yu, T. (n.d.). Salvaging Federated Learning by Local Adaptation. 1. arXiv:2007.08432v2

Bhagoji, A. N. (n.d.). Analyzing Federated Learning through an Adversarial Lens. Proceedings of the 36th International Conference on Machine Learning, PMLR 97:634-643, 2019, 1. https://proceedings.mlr.press/v97/bhagoji19a.html

Lyu, L., Yang, Q. (n.d.). Threats to Federated Learning: A Survey. 1. https://arxiv.org/abs/2003.02133

Shen, Z., Shokri, R. (n.d.). Share Your Representation Only: Guaranteed Improvement of the Privacy-Utility Tradeoff in Federated Learning. ICLR 2023 Conference, 1. https://openreview.net/forum?id=oJpVVGXu9i

What is federated learning? (2022, August 24). IBM Research. Retrieved May 12, 2023, from https://research.ibm.com/blog/what-is-federated-learning

Yu, T. (n.d.). Salvaging Federated Learning by Local Adaptation. 1. arXiv:2007.08432v2