

The Impact of Adversarial Node Placement in Decentralized Federated Learning Networks

Adam Piaseczny*, Eric Ruzomberka[†], Rohit Parasnis[‡], Christopher G. Brinton*

*Purdue University, Email: {apiasecz, cgb}@purdue.edu

[†]Princeton University, Email: er6214@princeton.edu

[‡] Massachusetts Institute of Technology, Email: rohit100@mit.edu

Abstract—As Federated Learning (FL) grows in popularity, new decentralized frameworks are becoming widespread. These frameworks leverage the benefits of decentralized environments to enable fast and energy-efficient inter-device communication. However, this growing popularity also intensifies the need for robust security measures. While existing research has explored various aspects of FL security, the role of adversarial node placement in decentralized networks remains largely unexplored. This paper addresses this gap by analyzing the performance of decentralized FL for various adversarial placement strategies when adversaries can jointly coordinate their placement within a network. We establish two baseline strategies for placing adversarial node: random placement and network centrality-based placement. Building on this foundation, we propose a novel attack algorithm that prioritizes adversarial spread over adversarial centrality by maximizing the average network distance between adversaries. We show that the new attack algorithm significantly impacts key performance metrics such as testing accuracy, outperforming the baseline frameworks by between 9% and 66.5% for the considered setups. Our findings provide valuable insights into the vulnerabilities of decentralized FL systems, setting the stage for future research aimed at developing more secure and robust decentralized FL frameworks.

I. INTRODUCTION

Federated Learning (FL) [1] has emerged as a popular method for distributed machine learning applications across various domains [2]. By sharing the workload of the training process across different nodes, FL enables parallel computation and eliminates the need for shared access to training data, enhancing both efficiency and privacy of the training process. While traditional FL frameworks follow the client-server architecture, recent work [3], [4] has focused on the fully decentralized setting in which nodes participating in the learning process exchange their models and conduct distributed consensus via D2D communication. This eliminates the need for a central server and offers advantages in terms of communication efficiency, scalability, and robustness.

However, the decentralized setup also introduces new challenges, particularly in terms of security in the training process. Centralized FL frameworks are known to be vulnerable to attacks on the training process by adversarial clients [5]. These adversaries can employ an array of common and well-established attacks [6]–[8] to compromise the global model's

integrity and prevent it from achieving its target. In the absence of a centralized coordinator, we expect that the attack potency is affected not only by the strength of the adversaries, but also by the placement of the adversaries in the network.

In this paper, we investigate this relationship. We aim to answer the following question: *How does the placement of adversarial nodes in the network affect the attack potency?* By addressing this question, we aim to provide valuable insights into the vulnerabilities of decentralized FL systems and lay the groundwork for future research focused on enhancing the security of such frameworks.

Related Work: While the influence of adversarial placement has not been extensively explored in decentralized FL, it has garnered significant attention in other networked systems [9], [10]. In many of these systems, the centrality measures of nodes play a pivotal role in understanding the dynamics of worst-case adversarial placement [11], [12]. This insight has been instrumental in formulating strategies for applications like virus defense and detection, underscoring the importance of centrality in network security.

Thus, recent developments in decentralized FL would reflect a growing interest in centrality-based approaches for placing adversarial nodes [13]. Given this research trend and similarities between decentralized FL and other networked systems, it would be reasonable to assume that centrality is a key factor in determining adversarial impact in FL environments. However, our research presents a nuanced perspective. We demonstrate that centrality-based approaches do not necessarily align with the most effective strategy for adversarial placement, and in fact significantly underperform other placement strategies in certain FL scenarios.

Correspondingly, growing concern over security vulnerabilities in deep learning has led to an increase in research efforts aimed at enhancing the security of FL architectures. A variety of strategies have been proposed [14], [15] that modify the aggregation process or fine-tune model parameters for higher training resilience. More recently, some attention has been directed towards the robustness of FL frameworks with different network topologies [4]. These studies analyzed various configurations, from semi-centralized to fully decentralized setups. However, many questions remain unanswered, particularly regarding how network topologies and the placement of adversarial nodes interact. This paper aims to address

This project was supported in part by ONR under grants N000142112472 and N000142212305, and by NSF under grants CNS-2146171 and ITE-2326898

some of these questions and fill this research gap.

Summary of Contributions In this paper, we make the following key contributions:

- **Model Poisoning Attack Analysis in Decentralized FL:** We analyze the effects of adversarial placement in poisoning attacks for different network topologies within decentralized FL settings. Our analysis introduces two baselines for evaluating the potency of placement attack strategies: random choice and centrality-based attacks.
- **MaxSpAN-FL Attack Framework:** We introduce a novel algorithm for placing attack nodes in specific network types. The resulting algorithm, MaxSpAN-FL (Maximally Spread-out Adversaries in a Network), relies on the principle of (approximately) maximizing the average inter-adversarial distance. MaxSpAN-FL demonstrates superior performance over baseline methods, achieving from 9% up to 66.5% performance improvement over next best baseline for the scenarios considered.

II. METHODOLOGY

In this section, we first introduce our decentralized FL model in Section II-A. Given this model, we then define our baseline algorithms we used to evaluate adversarial node placement in decentralized FL in Section II-B, and finally we define our proposed new attack algorithm in Section II-C.

A. System Model

As depicted in Fig. 1, a network of D2D nodes is a strongly connected, time-invariant, directed graph $G = (V, E)$ where V is the set of nodes and E is the set of edges (i.e., communication links) between the nodes. The set V can be partitioned into two sets: a set of adversarial nodes A and a set of honest nodes $H := V \setminus A$.

The network of D2D nodes participates in a decentralized FL scheme over a time horizon $t = 1, 2, 3, \dots, N_{\text{epochs}}$. At time t , an honest node $i \in V \setminus A$ trains a local model $x_i^{(t)} \in \mathbb{R}^p$ using a local data-set \mathcal{D}_i and a local loss function $f_i : \mathbb{R}^p \times \mathbb{R}^{|\mathcal{D}_i|} \rightarrow \mathbb{R}$. Training follows the S-AB aggregation procedure [16], in which the local model is computed as

$$x_i^{(t)} = \sum_{j:(j,i) \in E} \frac{x_j^{(t-1)}}{|j : (j,i) \in E|} - \alpha y_i^{(t-1)}$$

where $\alpha > 0$ is the learning rate and $y_i^{(t-1)}$ is the global gradient estimation of node i at time $t-1$ as given by the S-AB aggregation procedure. In contrast to the honest nodes, adversarial nodes do not follow the S-AB aggregation procedure. Instead, an adversarial node $i \in A$ mounts an FGSM attack [6] at time t using a local poisoned data-set $\mathcal{D}_{i,adv}^{(t-1)}$ calculated as:

$$\mathcal{D}_{i,adv}^{(t-1)} = \mathcal{D}_i + \epsilon_i \cdot \text{sign}(\nabla_{\mathcal{D}_i} f_i(x_i^{(t-1)}, \mathcal{D}_i))$$

where $\epsilon_i > 0$ is the attack power of node i . Using the poisoned data, the adversary computes the local model

$$x_i^{(t)} = x_i^{(t-1)} - \alpha \nabla_x f_i(x_i^{(t-1)}, \mathcal{D}_{i,adv}^{(t-1)})$$

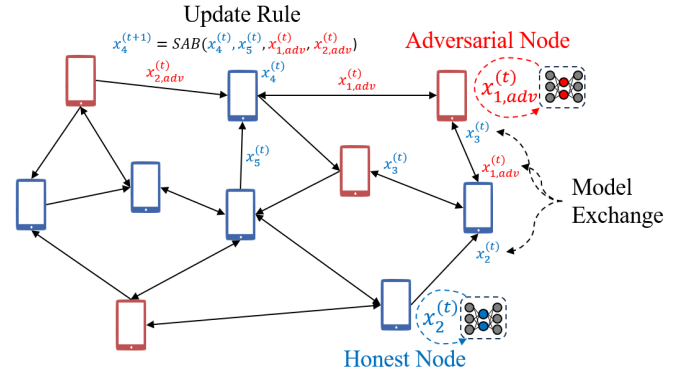


Fig. 1: Decentralized federated learning with adversarial nodes.

We assume that all adversarial nodes use the same attack power $\epsilon > 0$. After training but before time $t+1$, all nodes exchange their local models and global gradient estimations.

The goal of the honest nodes is to find a global model x^* that minimizes the global loss function

$$f(x) = \sum_{i \in V \setminus A} f_i(x_i, \mathcal{D}_i). \quad (1)$$

The S-AB aggregation procedure guarantees all nodes reach consensus and find an optimal model in the setting where all nodes are honest (i.e., A is empty). The adversaries have a goal opposite of the trusted nodes, namely, to maximize the global loss function (1).

B. Baseline Algorithms

To evaluate potency of a placement attack on a network we established the following 2 baseline algorithms for placing adversarial nodes: random placement and network centrality-based placement.

The random placement attack serves as our initial and most fundamental baseline for evaluating the effectiveness of more sophisticated attack strategies. In this approach, the subset $A \subset V$ is randomly selected, with the size $|A|$ specified in advance. Given its random nature, the random placement attack does not incorporate any strategic planning or targeted placement, making it the simplest form of attack in our study and providing a basic point of comparison for other attacks.

The network (eigenvector) centrality-based placement attack serves as our second baseline, offering a more nuanced approach than the random placement attack. This strategy is designed to exploit information that is inherent to the network topology when placing adversaries. It is rooted in the intuitive hypothesis, informed by other fields, that nodes with higher eigenvector centrality could potentially have a greater impact on the network. In this strategy, each node in the network graph G has its eigenvector centrality calculated. For a given size $|A|$, the subset A is formed by selecting the $|A|$ nodes in V with the largest eigenvector centralities. The underlying intuition suggests that these nodes would be most effective in spreading malicious updates across the network.

Eigenvector centrality was chosen for node placement because it is a comprehensive measure that considers the rela-

tionship of the node to its neighbors as well as to the entire network. This provides a mathematical basis for analysis, making it an attractive choice for our study. Furthermore, the computational efficiency of calculating eigenvector centrality via power iteration methods adds to its feasibility for larger networks.

Using this algorithm for a baseline will allow us to evaluate the effectiveness of an attack strategy using network topology information in an intuitive way and quantify the potency and success of other more sophisticated attacks.

C. New Attack Algorithm

In this section we propose an adversarial placement attack termed the Maximally Spread-out Adversaries in a Network for Federated Learning (MaxSpAN-FL) placement attack. The MaxSpAN-FL placement attack is designed to strategically identify influence regions for adversarial nodes in a way that efficiently and approximately maximizes the *average distance between adversarial nodes*

$$d_{\text{avg}} = \frac{1}{\binom{|A|}{2}} \sum_{i \in A} \sum_{\substack{j \in A \\ j > i}} d(i, j)$$

where $d(i, j)$ is the shortest graph hop distance between nodes i and j on G . It does so by employing a breadth-first search (BFS) clustering technique to consider the structure of the neighborhood surrounding a node. This allows us to spread the adversaries throughout the network in an efficient way, which we hypothesize should lead to a stronger attack.

The algorithm is outlined in II-C and employs a breadth-first search (BFS) clustering technique to identify influence regions for each node in the network. Initially, the algorithm calculates these influence regions for all nodes $g \in G$ using BFS. The BFS is run until the number of nodes in a given cluster exceeds the pre-determined cluster area S_{cluster} . This area is calculated by dividing the total number of nodes n_{clients} by the number of adversaries n_{advs} , ensuring that each adversary has a roughly equal area of influence.

Once these influence regions are established, the algorithm enters an iterative phase where it selects adversarial nodes a_{best} in a way that minimizes the overlap between their influence regions. Specifically, for each node still considered ‘honest’, the algorithm calculates the overlap o of its influence region with those of the already selected adversarial nodes. This overlap is quantified as

$$o = \left| \mathcal{C}_G[g] \cap \left(\bigcup_{a \in A} \mathcal{C}_G[a] \right) \right| \quad (2)$$

The node with the least overlap is then selected as the next adversarial node, added to the set A of adversarial nodes, and removed from the set H of honest nodes.

Notice that the algorithm introduces a degree of randomness in the placement of adversarial nodes. The first node is chosen completely at random, providing a starting point for the algorithm. Subsequent nodes are selected based on minimizing overlap, but within that constraint, the placement is influenced

by the current state of the network and the nodes already chosen as adversaries.

Algorithm 1 MaxSpAN-FL Attack Algorithm

```

1: Input: Graph  $G$ , Number of adversaries  $n_{\text{advs}}$ , Number of
   clients  $n_{\text{clients}}$ 
2: Output: List of adversarial nodes  $A$ 
3: Cluster Area  $S_{\text{cluster}} \leftarrow \lfloor n_{\text{clients}}/n_{\text{advs}} \rfloor$ 
4: Cluster dictionary  $\mathcal{C}_G \leftarrow \{g : \{\}\}_{g \in G}$ 
5: for  $g \in G$  do
6:    $\mathcal{C}_G[g] \leftarrow \text{BFS\_Cluster}(G, g, S_{\text{cluster}})$ 
7: end for
8: Set of adversaries  $A \leftarrow \{\}$ 
9: Set of honest nodes  $H \leftarrow G$ 
10: while  $|A| < n_{\text{advs}}$  do
11:    $o_{\text{min}} \leftarrow \infty$ 
12:    $a_{\text{best}} \leftarrow \text{None}$ 
13:   for  $g \in H$  do
14:      $o = \left| \mathcal{C}_G[g] \cap \left( \bigcup_{a \in A} \mathcal{C}_G[a] \right) \right|$ 
15:     if  $o < o_{\text{min}}$  then
16:        $o_{\text{min}} \leftarrow o$ 
17:        $a_{\text{best}} \leftarrow g$ 
18:     end if
19:   end for
20:    $A \leftarrow A \cup a_{\text{best}}$ 
21:    $H \leftarrow H \setminus a_{\text{best}}$ 
22: end while
23: return  $A$ 

```

III. RESULTS AND DISCUSSION

Here, we first discuss the experimental setup in Section III-A, which includes the network architectures, model architectures, and datasets used for evaluation. Then, in Section III-B, we present the experiments conducted and discuss the corresponding results for various hyperparameter settings.

A. Experimental Setup

In our experiments, we consider the decentralized FL model of Section II-A. We evaluate this model for distinct network topologies in which the graph G is drawn at random. We consider two random graph distributions: Directed Geometric (DG) graphs and Erdos-Renyi (ER) graphs. To assess the effectiveness of our proposed MaxSpAN-FL Attack we synthesize networks with varying hyperparameters, such as network connectivity or network size. For our experiments we generate between 20 and 50 network realizations, using different random seeds.

The network hyperparameters we consider are as follows:

- Connectivity parameters for 25-node networks for both graph types. For the DG graphs, we randomly assign nodes’ locations in a 2D unit square and vary the connection radius parameter r to 0.2, 0.4, and 0.6 to simulate different levels of network density. Correspondingly, for

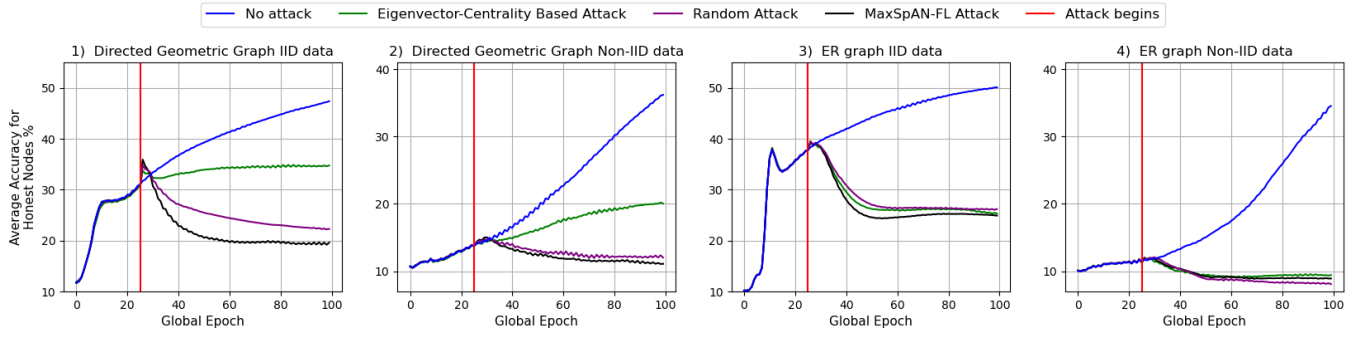


Fig. 2: Average testing accuracy of honest nodes in 25-node networks, comparing Directed Geometric graphs with connection radius $r = 0.2$ and Erdős–Rényi graphs with edge probability $p = 0.5$, for both IID and Non-IID data distributions. The effects of various attack placement strategies on the network’s performance are illustrated. Adversarial percentage is 20%.

ER graphs, we adjust the edge creation probability p , setting it to 0.1, 0.3, and 0.5.

- Size parameters for the DG graph with $r = 0.2$. We consider network sizes of 10, 25, 50, and 100, and for each network we consider different adversarial percentages, ranging from 5% up to 20%.

Following network generation, each node trains an image classification model. We use the Fashion-MNIST (FMNIST) dataset for evaluation, distributing it across nodes in both IID and Non-IID manners, with the latter having 3 classes per node. Tailored Convolutional Neural Network (CNN) model architecture serves as our classifier. The model architecture consists of two convolutional layers, each succeeded by a dropout layer, a ReLU activation function, and a max-pooling operation, culminating in two dense layers for classification.

Each node executes a predefined number of local training iterations on their respective model. Upon completion, the models are distributed amongst neighboring nodes via the previously mentioned S-AB algorithm. We use Adam optimizer for training.

In all cases, we evaluate the 2 baseline placement attacks, as well as the MaxSpAN-FL attack for selecting the group of adversarial nodes that run FGSM as described in II-A. Adversarial nodes always ignore models from their in-neighbors.¹

B. Experimental Results

1) *Node Placement:* Figure 2 illustrates that the adversarial placement attack potency depends on both the network’s graph structure and the data distribution among clients. In directed geometric (DG) graphs with IID data, our proposed MaxSpAN-FL attack method surpasses alternative strategies, achieving a minimal final test accuracy of 19.6%, outperforming the 22.3% and 34.8% final accuracies achieved by random node and eigenvector-centrality based attacks, respectively. However, the performance disparity narrows in other scenarios, with the MaxSpAN-FL attack outperforming the next best method by only 0.93% in DG graphs with Non-IID data and 0.63% in ER graphs with IID data. In ER graphs with Non-IID data, our method trails slightly behind the most effective attack, with a marginal performance gap of 0.82%.

¹The source code supporting our experiments is publicly available at https://github.com/Adampi210/MaxSpANFL_atck_code_data.git.

Intuitively, the similar performance observed across placement attack strategies in ER graphs follows from the high degree of randomness in ER: with large probability, all nodes in an ER graph have similar centrality scores and any two nodes have a small shortest path length. Thus, our three placement attack strategies are nearly equivalent in ER graphs.

Conversely, in DG graphs with IID data, our MaxSpAN-FL algorithm not only demonstrates superior final performance but also exhibits a faster rate of decreasing test accuracy. The algorithm reaches its lowest accuracy point around the 60th global epoch, a point not achieved by other methods even by the 100th epoch. This highlights the significance of strategic adversarial placement in DG graphs.

Lastly, the data distribution among clients significantly influences the performance gap between placement attack strategies. In Non-IID scenarios, this gap notably reduces, potentially due to the added complexity of class distribution in adversarial node placement and the varying degrees of model convergence at the time of attack. Notably, the test accuracy of attacked clients in Non-IID settings is substantially lower than in IID scenarios, further underscoring the role of data distribution in shaping attack outcomes.

2) *Effects of Connectivity:* Figure 3 demonstrates the influence of network connectivity on the potency of different placement attacks, considering both IID and Non-IID data distributions in networks with 25 nodes. The analysis of this setting is based on the Attack Accuracy Loss (AAL) measure, calculated as

$$AAL = \sum_{i=t_{\text{attack}}}^{N_{\text{epochs}}} [(\% \text{ acc no attack})[i] - (\% \text{ acc with attack})[i]]$$

as it measures both by how much and how fast the attack decreases the testing accuracy. Higher AAL indicates better attack potency. Attack performance advantage is computed as:

$$\frac{AAL_{\text{best attack}} - AAL_{\text{next best attack}}}{AAL_{\text{next best attack}}} \times 100\%$$

In Directed Geometric (DG) graphs, connectivity is varied with radius values $r \in \{0.2, 0.4, 0.6\}$. Our MaxSpAN-FL attack demonstrates superior performance in the $r = 0.2$ and $r = 0.6$ scenarios under IID conditions, outperforming the next best attack by 14.0% and 21.8%, respectively. However,

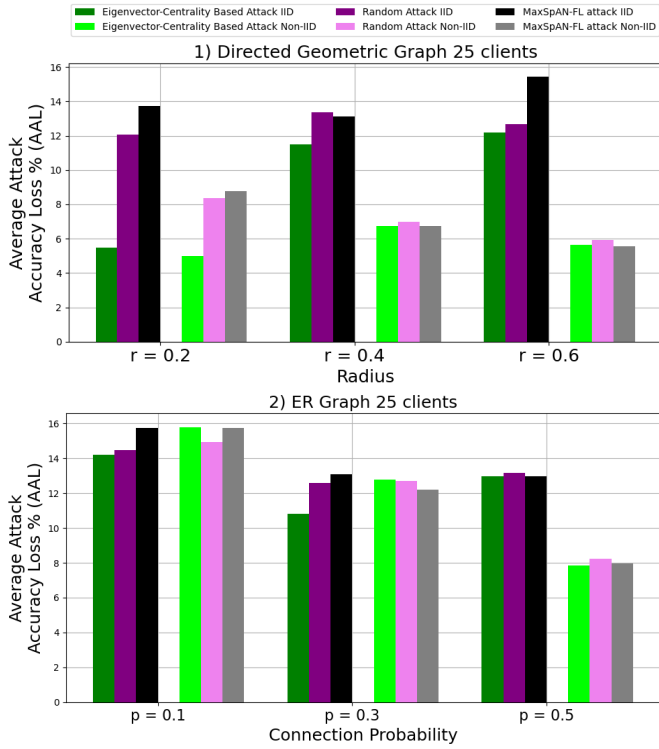


Fig. 3: Average Attack Accuracy Loss for Directed Geometric and ER graphs with 25 nodes and 20% adversaries for different connectivity parameters.

in the $r = 0.4$ scenario, the performance margin narrows, with our attack trailing by a slight margin of 1.8%. These findings indicate that MaxSpAN-FL maintains a competitive edge across varying connectivity levels in DG graphs.

For Erdős-Rényi (ER) graphs, we assess the impact of different connection probabilities $p \in \{0.1, 0.3, 0.5\}$. The MaxSpAN-FL attack shows its greatest advantage in the $p = 0.1$ case, surpassing the next best attack by approximately 9%. As the connection probability increases, the performance gap narrows, aligning with intuition that more sparse ER graphs exhibit greater variance in edge distribution, making network data more informative.

These results not only affirm the performance edge and competitiveness of the MaxSpAN-FL attack across diverse network configurations but also offer several intriguing insights. Notably, in Non-IID scenarios, the peak Attack Accuracy Loss consistently decreases with increased graph connectivity, a trend not as pronounced in IID settings. This is likely caused by enhanced information exchange among honest clients in denser networks. In ER graphs, as connectivity approaches that of a fully connected network, the placement attack performance differences go to 0, reinforcing the notion that node placement becomes less critical in highly connected networks. Additionally, the Eigenvector-Centrality Based attack improves its performance by 123.4% in DG graphs with IID data as connectivity increases, suggesting better connectivity to most central nodes, even though it still underperforms MaxSpAN-FL for all radius values.

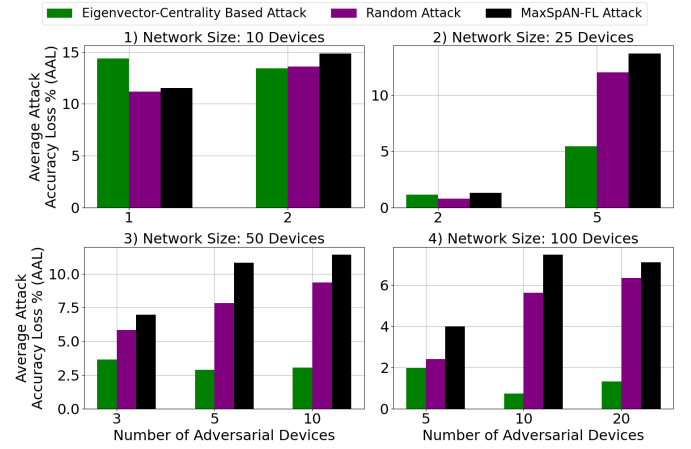


Fig. 4: Average Attack Accuracy Loss for Directed Geometric graphs with $r = 0.2$ and IID data distribution, for different network sizes and number of adversaries

3) Effects of Network Size and Adversarial Percentage: The impact of network size and adversarial percentage on the performance differential among various attack strategies is illustrated in Figure 4.

We consider the Directed Geometric (DG) graphs with radius $r = 0.2$, since ER graphs inherently are generated using IID process, therefore their performance across various sizes is consistent. Additionally, we focus on IID data distribution, as Non-IID scenarios did not yield significant variances in attack performance to discern clear trends.

In nearly all examined cases, the MaxSpAN-FL attack surpasses other methods, particularly notably at adversarial percentages of 10% and 20% in larger networks. The attack achieves performance advantage of up to 66.5% improvement over the closest competing attack in a 100-node network with 5 adversaries. Even in the smallest network scenario of 10 clients and 2 adversaries, MaxSpAN-FL maintains a 9.4% advantage, which is its lowest performance difference that's non-negative. The sole exception occurs in a 10-client network with a single adversary, where the Eigenvector-Centrality based attack excels, outperforming MaxSpAN-FL by approximately 20.1%. This is an expected outcome, as MaxSpAN-FL's initial node placement is random, equating to a random node attack in single-adversary scenarios. Conversely, the Eigenvector-Centrality approach targets the most influential node, thereby maximizing attack potency. This exception aside, MaxSpAN-FL consistently outperforms other strategies across varying network sizes in DG graphs, underscoring the effectiveness of the proposed placement attack method.

4) Effects of Attack Deployment Time: Our final experiments investigate the impact of varying attack deployment times during global aggregation. We continue to focus on the Directed Geometric (DG) graph with 25 clients and radius $r = 0.2$, for previously stated reasons. The findings are depicted in Figure 5.

Across various attack timings, the MaxSpAN-FL attack consistently demonstrates superior performance in terms of both final model accuracy and adversarial accuracy loss.

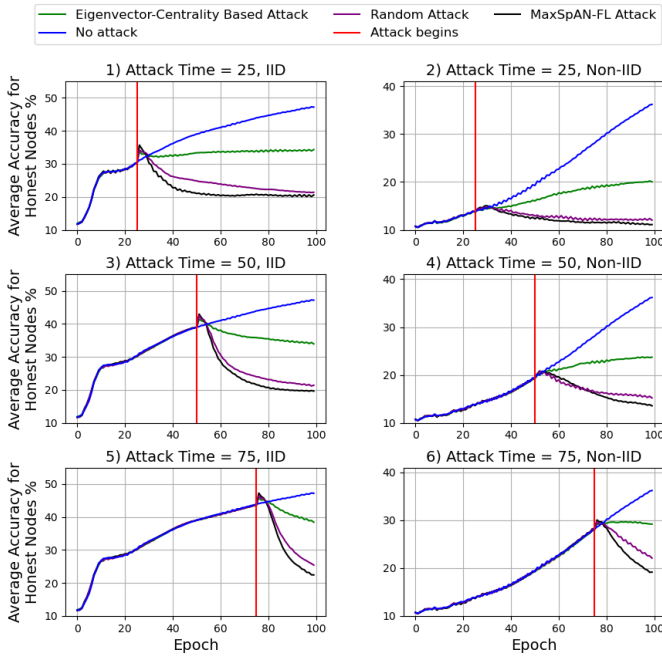


Fig. 5: Average testing accuracy over honest nodes in 25-node Directed Geometric Graph with $r = 0.2$, for both IID and Non-IID data distributions for different attack deployment times. Adversarial percentage is 20%.

This suggests that the effectiveness of the proposed attack is largely unaffected by its deployment time. Notably, in non-IID scenarios, the final accuracy discrepancy widens with later attack times. This trend could intuitively be attributed to the degree of model convergence at the point of attack, although this requires further investigation. Additionally, while attacks initiated at later stages do not reach convergence within 100 epochs, their trajectories seem to be aligning towards final accuracy levels similar to those observed in the 25-epoch attack scenario.

C. Key Takeaways

- The impact of adversarial node placement is significantly influenced by the network's graph structure and the data distribution, as shown in Figure 2.
- In most scenarios, MaxSpAN-FL either surpasses or matches baseline approach performances. MaxSpAN-FL performance improvement reaches up to around 66.5%.
- Network connectivity plays an important role in determining the effectiveness of different adversary placement strategies and their relative performance.
- Varying graph sizes maintain the performance disparity between adversarial placement strategies. The proportion of adversarial nodes notably influences this difference, particularly in scenarios with fewer adversarial nodes.
- The timing of the attack consistently maintains performance difference trend between different placement strategies. Networks with extended training rounds exhibit larger variations in the degree of performance disparity among these strategies.

IV. CONCLUSION AND FUTURE WORK

The rapid advancement of decentralized processing frameworks introduces a range of novel applications, while simultaneously presenting unique security challenges. In this paper we delved into the impact of adversarial node placement within FL environments, a critical aspect often influenced by network topological factors. Our comprehensive analysis reveals that the effect of adversarial node placement on attack potency in FL depends on a variety of elements, including network type, data distribution, connectivity, size, and the timing of the attack. We also developed a novel attack placement algorithm, and demonstrated superior performance in attack potency across diverse conditions. The attack performance suggests that under various network conditions, there is a combination of nodes that performs better on average. In future work we plan to explore additional graph types, centrality measures, and the role of data heterogeneity in attack strategies.

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [3] T. Ranathunga, A. McGibney, S. Rea, and S. Bharti, "Blockchain-based decentralized model aggregation for cross-silo federated learning in industry 4.0," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4449–4461, 2023.
- [4] M. Raynal, D. Pasquini, and C. Troncoso, "Can decentralized learning be more robust than federated learning?" 2023.
- [5] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," 2020.
- [6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.
- [7] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," 2017.
- [8] S. Wang, R. Sahay, and C. G. Brinton, "How potent are evasion attacks for poisoning federated learning-based signal classifiers?" 2023.
- [9] S. Tang and B. L. Mark, "Analysis of virus spread in wireless sensor networks: An epidemic model," in *2009 7th International Workshop on Design of Reliable Communication Networks*, 2009, pp. 86–91.
- [10] Z. Zhang, H. Wang, C. Wang, and H. Fang, "Modeling epidemics spreading on social contact networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 410–419, 2015.
- [11] M. Mirchev, I. Mishkovski, and L. Kocarev, "Selective attack in virus propagation processes," in *ICT Innovations 2010: Second International Conference, ICT Innovations 2010, Ohrid Macedonia, September 12-15, 2010. Revised Selected Papers 2*. Springer, 2011, pp. 286–295.
- [12] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," in *Proceedings of the ACM SIGMETRICS*, ser. SIGMETRICS '10. Association for Computing Machinery, 2010, p. 203–214.
- [13] G. Yar, C. Nita-Rotaru, and A. Oprea, "Backdoor attacks in peer-to-peer federated learning," *arXiv preprint arXiv:2301.09732*, 2023.
- [14] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1142–1154, 2022.
- [15] X. Cao, J. Jia, and N. Z. Gong, "Provably secure federated learning against malicious clients," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6885–6893, May 2021.
- [16] R. Xin, A. K. Sahu, U. A. Khan, and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 8353–8358.