Journal Paper

Leverage Variational Graph Representation for Model Poisoning on Federated Learning

Kai Li*
Xin Yuan
Jingjing Zheng*
Wei Ni
Falko Dressler
Abbas Jamalipour

*CISTER Research Centre

CISTER-TR-240507

2024/05/03

Leverage Variational Graph Representation for Model Poisoning on Federated Learning

Kai Li*, Xin Yuan, Jingjing Zheng*, Wei Ni, Falko Dressler, Abbas Jamalipour

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: kai@isep.ipp.pt, xin.yuan@data61.csiro.au, zheng@isep.ipp.pt, Wei.Ni@data61.csiro.au, dressler@ccs-labs.org

https://www.cister-labs.pt

Abstract

This article puts forth a new training data-untethered model poisoning (MP) attack on federated learning (FL). The new MP attack extends an adversarial variational graph autoencoder (VGAE) to create malicious local models based solely on the benign local models overheard without any access to the training data of FL. Such an advancement leads to the VGAE-MP attack that is not only efficacious but also remains elusive to detection. VGAE-MP attack extracts graph structural correlations among the benign local models and the training data features, adversarially regenerates the graph structure, and generates malicious local models using the adversarial graph structure and benign models 19 features. Moreover, a new attacking algorithm is presented to train the malicious local models using VGAE and sub-gradient descent, while enabling an optimal selection of the benign local models for training the VGAE. Experiments demonstrate a gradual drop in FL accuracy under the proposed VGAE-MP attack and the ineffectiveness of existing defense mechanisms in detecting the attack, posing a severe threat to FL.

Leverage Variational Graph Representation For Model Poisoning on Federated Learning

Kai Li, *Senior Member, IEEE,* Xin Yuan, *Senior Member, IEEE,* Jingjing Zheng, *Student Member, IEEE,*
Wei Ni, *Fellow, IEEE,* Falko Dressler, *Fellow, IEEE*, and Abbas Jamalipour, *Fellow, IEEE*

***Abstract*—This paper puts forth a new training data-untethered model poisoning (MP) attack on federated learning (FL). The new MP attack extends an adversarial variational graph autoencoder (VGAE) to create malicious local models based solely on the benign local models overheard without any access to the training data of FL. Such an advancement leads to the VGAE-MP attack that is not only efficacious but also remains elusive to detection. VGAE-MP attack extracts graph structural correlations among the benign local models and the training data features, adversarially regenerates the graph structure, and generates malicious local models using the adversarial graph structure and benign models' features. Moreover, a new attacking algorithm is presented to train the malicious local models using VGAE and sub-gradient descent, while enabling an optimal selection of the benign local models for training the VGAE. Experiments demonstrate a gradual drop in FL accuracy under the proposed VGAE-MP attack and the ineffectiveness of existing defense mechanisms in detecting the attack, posing a severe threat to FL.**

***Index Terms*—Federated learning, variational graph autoencoders, data-untethered model poisoning**

I. Introduction

Federated learning (FL) has attracted significant attention recently, and emerged as a distributed deep learning paradigm. With FL, each user device trains its local model with its private data to generate local updates sent to the edge server without sharing the device's private data. The edge server then aggregates the local updates to train a global model, which is sent back to the user devices for the next round of FL training. Based on FL, individual data privacy is protected as no private data is shared [1].

Despite the fact that FL offers a protective measure for the data privacy of user devices, it remains susceptible to cyber-epidemic attacks. In these attacks, malevolent entities, such as compromised user devices, execute model or data poisoning strategies. These tactics are designed to manipulate the FL process and proliferate across other innocuous user devices. Consequently, this leads to the derailment of the training process and a subsequent degradation in the accuracy of the learning outcomes [2]. For the model poisoning attacks, the attacker aims to manipulate the hyperparameters of the benign local model. In contrast, data poisoning attacks involve manipulating the training dataset of benign user devices. To launch effective model or data poisoning attacks [3], the attackers need to access the knowledge of the dataset used for FL training, which helps to minimize the detectability of malicious local models. FL could be manipulated if an attacker launches model poisoning attacks based solely on the benign local and global models overheard without access to the data. Nevertheless, it is challenging for the attacker to achieve effectiveness and undetectability without knowledge of the data. This type of attack is new, has not yet been discussed in the existing literature, and requires further research to develop effective detection and prevention methods. This new attack underscores the importance of securing FL from local and global training threats.

This paper investigates a new adversarial variational graph autoencoder (VGAE)-based model poisoning (VGAE-MP) attack on FL. VGAE-MP is a new data-untethered cyber-epidemic attack, where malicious local models are generated solely based on the benign local models overheard by attackers and the correlation features of the benign local and global models. This attack could be particularly severe in FL systems under wireless settings, due to the broadcast nature of radio. The attacker starts the VGAE-MP attack by overhearing (or eavesdropping on) the transmissions of local model updates from the benign clients in a communication round. The attacker also has the global model that the server shared in the previous communication round. Then, the attacker executes the VGAE-MP model to craftily generate its malicious local model update that, when aggregated, subtly distorts the global model in the current round.

Specifically, the attacker manipulates its malicious model update to introduce erroneous gradients or patterns. This is done by running the adversarial VGAE to capture the correlation of the benign local models and then regenerate the graph structure to create malicious local models that can effectively compromise the global and benign local models while remaining indistinguishable from the benign local models. Over time, this insidious injection of inaccuracies

K. Li is with the Department of Engineering, University of Cambridge, CB3 0FA Cambridge, U.K., and also with Real-Time and Embedded Computing Systems Research Centre (CISTER), Porto 4249–015, Portugal (E-mail: kaili@ieee.org).

J. Zheng is with CyLab Security and Privacy Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA, and also with Real-Time and Embedded Computing Systems Research Centre (CISTER), Porto 4249–015, Portugal (E-mail: zheng@isep.ipp.pt).

X. Yuan and W. Ni are with the Digital Productivity and Services Flagship, Commonwealth Scientific and Industrial Research Organization (CSIRO), Marsfield, NSW 2122, Australia (E-mail: {xin.yuan,wei.ni}@data61.csiro.au).

F. Dressler is with the School of Electrical Engineering and Computer Science, TU Berlin, Germany (E-mail: dressler@ccs-labs.org).

A. Jamalipour is with the School of Electrical and Information Engineering, The University of Sydney, Australia (E-mail: a.jamalipour@ieee.org).

TABLE I: Notation and definition

Notation	Definition
I	number of benign devices
J	number of attackers
$D_i(t)$	datasets of the benign device i at the t-th communication round
D	total datasets of I number of benign devices
$D'(t)$	the claimed data size of the attacker
$\boldsymbol{w}_i(t)$	local model weight parameters of the benign device i
$\boldsymbol{w}'_j(t)$	training parameters of the malicious model at attacker j
$\boldsymbol{w}'_G(t)$	the global model of FL under attack
$\beta'_{i,j}(t)$	the binary indicator for selecting benign local model weights
λ, ρ	the Lagrangian dual variables
$\tau_i(t)$	the training delay of $\boldsymbol{w}_i(t)$ at device i
M	total number of model parameters in $\boldsymbol{w}_i$
$\boldsymbol{w}_i^m(t)$	the m-th feature in $\boldsymbol{w}_i$
$\boldsymbol{\mathcal{A}}$	the adjacency matrix formulated by attacker
$\boldsymbol{\mathcal{F}}$	the feature matrix in VGAE of attacker
$\boldsymbol{\mathcal{L}}$	the Laplacian matrix based on $\boldsymbol{\mathcal{A}}$
$\boldsymbol{\mathcal{L}}_k$	the rank-k SVD approximation of $\boldsymbol{\mathcal{L}}$
η_{loss}	the reconstruction loss of the decoder in VGAE
$\widehat{\boldsymbol{\mathcal{A}}}$	the reconstructed adjacency matrix generated at the decoder of attacker
$\widehat{\boldsymbol{\mathcal{F}}}$	the reconstructed feature matrix at attacker

shifts the global model away from its optimal learning trajectory, leading to a gradual but significant decline in overall FL accuracy.

Since the user devices possessing large datasets could improve the learning accuracy of FL, the server selects a portion of the collected local models for the global aggregation. Likewise, the VGAE-MP, as a white-box attack, also selects the benign local models in the training of the VGAE. For example, the user device selection at the attacker ensures that the selected local models have sufficient data features for retrieving the correlation in the VGAE, while the generated malicious local model is within proximity to the global model in Euclidean distance.

The key contributions of this paper are as follows:

- A new data-untethered model poisoning attack, i.e., VGAE-MP, is proposed to manipulate the correlations of multiple data features in the selected benign local models and maintain the genuine data features substantiating the benign local models;
- A new adversarial VGAE, which is trained together with sub-gradient descent to regenerate the correlations of the local models manipulatively while keeping the malicious local models undetectable.
- The proposed VGAE-MP attack is implemented in PyTorch, showing experimentally that VGAE-MP gradually reduces the accuracy of FL and bypasses the detection of existing poisoning defense mechanisms. This attack can propagate across all benign user devices, which leads to an epidemic infection. The source code of the VGAE-MP attack has been released on GitHub.

The remaining of this paper is structured as follows. Section II introduces the background of adversarial attacks against wireless systems and FL. Section III investigates the FL system model with malicious agents. The proposed VGAE-MP attack is described in Section IV. Section V discusses the performance analysis. Section VI concludes the paper. Table I lists the notation used in the paper.

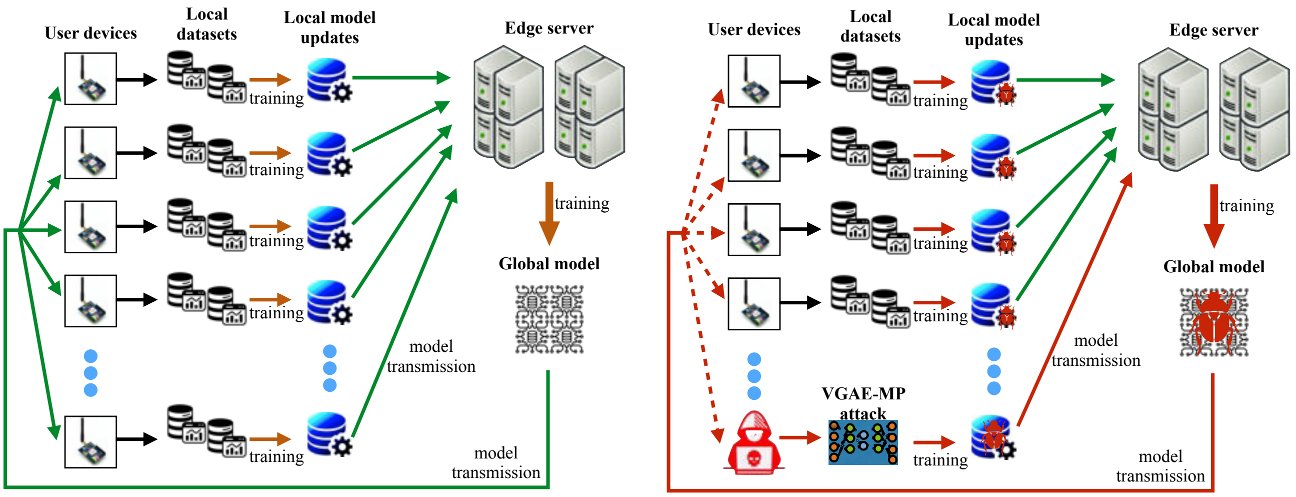
II. Related Work

This section reviews the literature on adversarial attacks and security threats to FL, e.g., model poisoning, data poisoning, inference, and backdoor attacks.

The periodic model updates in FL bear a discriminative ability that reflects changes in data distribution, including sensitive properties, making it possible for an attacker to infer unintended information. In [4], the authors introduce a poisoning-assisted property inference attack, which injects malicious data into the training dataset to infer a targeted property of the FL model. The attack modifies the training data labels, thereby distorting the decision boundary of the shared global model in FL, resulting in the disclosure of sensitive property information by benign user devices. In [5], the authors present that the attacker can infer the presence or absence of a particular category in the data by carefully crafting a malicious training dataset, despite the secure aggregation methods. A category inference attack is developed, which iteratively generates malicious training data and utilizes them to update the global model in FL. The vulnerability of FL to label inference attacks is presented in [6], where a malicious user device can infer the private labels of other benign devices. With the observed aggregate model updates, three label inference attacks have been developed to infer private labels with the benign devices, including direct, passive, and active label inference attacks.

In [7], a coordinated backdoor attack on FL is designed using model-dependent triggers, where an attacker can inject a backdoor trigger into a target model and then train the models in FL to perform coordinated attacks. The trigger uses the model dependency in FL to activate the backdoor when the target client uses the compromised model. A distributed attacking algorithm is also provided to enable the attackers to select their respective backdoor models for a high attack success rate while maintaining a low impact on the overall training accuracy of FL.

The authors of [8] focus on data poisoning attacks in both sequential and parallel FL settings. These attacks could weaken the performance of trained models by injecting malicious data into the training datasets used in FL. Sequential FL involves user devices training successively, using the output model from the previous device. In contrast, parallel FL involves each user device simultaneously training a local model before sending updates to the server for aggregation. An attacker can later trigger malicious behavior during the prediction phase by modifying specific training inputs using a specific pattern. In [9], it is argued that FL based on weighted averaging and trimmed averaging for mitigating Byzantine faults is still vulnerable to data poisoning attacks. These attacks can lead to considerable reductions in training accuracy, highlighting a critical vulnerability in the current mitigation strategies within FL. A data poisoning attack is



(a) FL with benign user devices

(b) The proposed VGAE-MP attack

Fig. 1: (a) Illustration of FL, where a local model update is trained at each benign user device based on its datasets. The edge server aggregates the benign local model updates to train a global model that will be broadcast to the user devices to update the training parameters of their local models. (b) By eavesdropping on the benign local model updates, the attacker performs the proposed VGAE-MP attack to create a malicious poisoning model that is sent to the server. The malicious model deviates the FL in the opposite direction, thereby falsifying the local model updates of the devices.

studied, which targets the FL system designed to be robust against Byzantine attacks. The data poisoning attack can exploit the characteristics of FL and the Byzantine-robust mechanisms to insert malicious data into the system.

In [10], a model poisoning attack is introduced, which accounts for the characteristics of FL, such as variability in the training data and randomness of the training process. The model poisoning attack uses a transfer learning strategy to improve the attack efficiency. A model poisoning attack on FL based on fake user devices added to the system and operating as legitimate devices is presented in [11]. This fake device can manipulate its data to influence the global model and potentially insert a backdoor or degrade the FL performance. In [12], the authors study a perception poisoning attack in which the attacker manipulates the FL model's perception by altering the training data. The attack can be captured by building a poison perception model for measuring a perception poisoning rate.

In [13], the authors focus on a generative poisoning attack against FL, which generates malicious data using generative adversarial networks (GAN) to target user devices in FL. The attack can introduce bias into the aggregated model by injecting poisoned data generated by the GANs. Another GAN-based poisoning attack against FL is presented in [14]. The GANs-based poisoning attack creates a set of malicious samples by generating poisoned data samples to attack the local models of benign user devices. To degrade the training accuracy of FL, the attacker deceives the aggregation process at the server by strategically altering the models of the benign user devices. The GAN-based poisoning attack is evaluated based on image classification datasets.

The existing data or model poisoning attacks against FL lack the description of the implicit relationship between different local models, which can be detected by recent poisoning defense frameworks based on the probabilistic graph model, e.g., [15], [16]. Additionally, convolutional layers at the aggregator can excessively smooth out the output features of the attacks, resulting in distinguishable discrepancies between the malicious local model and the benign ones. In contrast, the proposed VGAE-MP attack is a new attacking method for model poisoning, which is independent of the data. The VGAE-MP attack manipulates the correlations among multiple data features in selected benign local models while preserving the genuine data features that support those models, thus keeping the malicious local models undetectable.

III. System Model and Problem Statement

This section presents the training of the local and global models of FL in mobile edge computing for image classification as an example. Figure 1(a) presents an FL training process with I benign user devices. Each benign device $i \in [1, I]$ has $D_i(t)$ data samples at the t-th communication round of FL. Let $x(d_i)$ denote a data sample captured at the i-th benign device, and $y(d_i)$ the local model update trained at the i-th benign device, where $d_i \in [1, D_i(t)]$ [17].

The training loss function of a benign device i, denoted by $f(\boldsymbol{w}_i(t); x(d_i), y(d_i))$, measures approximation errors based on the inputs $x(d_i)$ and outputs $y(d_i)$ in the t-th communication round, where $\boldsymbol{w}_i(t) \in \mathbb{R}^{1\times M}$ denotes the local model obtained in the communication round. For example, the loss function can be modeled as linear regression, i.e., $f(\boldsymbol{w}_i(t); x(d_i), y(d_i)) = \frac{1}{2}(\boldsymbol{w}_i^T(t)x(d_i) - y(d_i))^2$, or logistic regression, i.e., $f(\boldsymbol{w}_i(t); x(d_i), y(d_i)) =$

$y(d_i)\log\Big(1+\exp\big(-\boldsymbol{w}_i^T(t)x(d_i)\big)\Big)-(1-y(d_i))\log\Big(1-\frac{1}{1+\exp\left(-\boldsymbol{w}_i^T(t)x(d_i)\right)}\Big)$. Here, $(\cdot)^T$ denotes transpose.

Given $D_i(t)$, the local loss function of the FL at device i for the t-th communication round is

$$F(\boldsymbol{w}_i(t)) = \frac{1}{D_i(t)}\sum_{i=1}^{D_i(t)} f\big(\boldsymbol{w}_i(t); x(d_i), y(d_i)\big) + \alpha\zeta\big(\boldsymbol{w}_i(t)\big), \tag{1}$$

where $\zeta(\cdot)$ is a regularizer function capturing the effect of local training noise; $\alpha \in [0,1]$ is a given coefficient [18].

With the learning rate μ, the local model of device i is updated for T_L local iterations throughout the t-th communication round by

$$\boldsymbol{w}_i(t) \leftarrow \boldsymbol{w}_i(t) - \mu\nabla F(\boldsymbol{w}_i(t)), \tag{2}$$

After the T_L local iterations, all devices upload their local models $\boldsymbol{w}_i(t), \forall i$ to the server. The server aggregates the local models to train a global model denoted by $\boldsymbol{w}_G(t)$ for the t-th communication round. Then, $\boldsymbol{w}_G(t)$ is broadcast to all user devices for their training of $\boldsymbol{w}_i(t+1)$, $\forall i$ in the $(t+1)$-th communication round.

Figure 1(b) shows the FL of the benign user devices under the proposed VGAE-MP attack, where an attacker overhears $\boldsymbol{w}_i(t)$ uploaded by the benign devices. The attacker, who may appear as a legitimate device, can progressively contaminate the global model represented by $\boldsymbol{w}_G(t)$ and the local models of the benign users, i.e., $\boldsymbol{w}_i(t)$, $\forall i \in [1, I]$, by creating and uploading malicious local models during each communication round t. The malicious local model at the attacker's device j is represented by $\boldsymbol{w}'_j(t)$. It is constructed based on the parameters of the benign local models overheard by the attacker during each communication round t. The server aggregates the local models of the user devices, including both benign and malicious models, without realizing the attacker's presence. This creates a contaminated global model, $\boldsymbol{\omega}'_G(t)$. The total size of the local training data reported to the server, $D(t)$, is calculated as the sum of the data size of all devices, $D_i(t)$, and the claimed data size of the attacker, $D'(t)$.

Some FL systems allow the server to select a portion of the collected $\boldsymbol{w}_i(t)$ to train $\boldsymbol{w}_G(t)$. For example, the authors of [19] considered a selection scheme in which the total bandwidth of the selected devices needs to be smaller than the bandwidth capacity. We define a binary indicator $\beta'_{i,j}(t)$ at the attacker. If $\boldsymbol{w}_i(t)$ is selected by the attacker to train its adversarial and contaminating local model, then $\beta'_{i,j}(t) = 1$; otherwise, $\beta'_{i,j}(t) = 0$. Thus, the contaminated global model can be written as

$$\boldsymbol{w}'_G(t) = \sum_{i=1}^{I}\frac{D_i(t)}{D(t)}\beta'_{i,j}(t)\boldsymbol{w}_i(t) + \frac{D'(t)}{D(t)}\boldsymbol{w}'_j(t), \tag{3}$$

where $\boldsymbol{w}'_j(t)$ is the weight parameter of the malicious model trained at attacker j. Then, the server broadcasts $\boldsymbol{w}'_G(t)$ to all I devices.

The FL trains the global model based on the local datasets of all user devices, including the non-existent dataset claimed by the attacker, by minimizing the global loss function:

$$\min_{\boldsymbol{w}'_G(t)} F(\boldsymbol{w}'_G(t)) = \sum_{i=1}^{I}\frac{D_i(t)}{D(t)}\beta'_{i,j}(t)F_i(\boldsymbol{w}_i(t)) + \frac{D'(t)}{D(t)}F'_j(\boldsymbol{w}'_j(t)), \tag{4}$$

where the attacker's claimed local loss function, represented by $F'_j(\cdot)$, is in accordance with (1).

The optimization of the proposed VGAE-MP attack aims to construct an optimal $\boldsymbol{w}'_j(t)$ based on the overheard $\boldsymbol{w}_i(t)$ to maximize $F(\boldsymbol{w}'_G(t))$ in (4), while maintaining a reasonably small Euclidean distance between $\boldsymbol{w}'_j$ and $\boldsymbol{w}'_G$. This helps $\boldsymbol{w}'_j(t)$ remain undetectable by the server, because the server could evaluate similarities among local models and eliminate those differing significantly using, e.g., Krum or multi-Krum [20]. Consequently, $\boldsymbol{w}'_G(t)$ deviates the most in the opposite direction that the benign global model would change in the absence of the attack.

The optimization of VGAE-MP launched by the attacker j, $\forall j \in [1, J]$, in the communication round t is formulated as

$$\max_{\boldsymbol{w}'_j(t), \beta'_{i,j}(t)} \quad F(\boldsymbol{w}'_G(t)) \tag{5a}$$

$$\text{s.t.} \quad d(\boldsymbol{w}'_j(t), \boldsymbol{w}'_G(t)) \le d_T, \tag{5b}$$

$$\sum_{i=1}^{I}\beta'_{i,j}(t)d(\boldsymbol{w}_i(t), \bar{\boldsymbol{w}}(t)) \le \Upsilon, \tag{5c}$$

$$\beta'_{i,j}(t) \in \{0, 1\}, \tag{5d}$$

where $d(\cdot,\cdot)$ calculates the Euclidean distance between $\boldsymbol{w}'_j$ and $\boldsymbol{w}'_G$, d_T is a threshold of the Euclidean distance, $\bar{\boldsymbol{w}}(t) = \sum_{i=1}^{I}\frac{D_i(t)}{D(t)}\boldsymbol{w}_i(t)$, and Υ is a predefined upper bound of the overall distance between the selected local models and the aggregated model of the local models.

Constraint (5b) guarantees that the j-th attacker's malicious local model $\boldsymbol{w}'_j$ is in proximity to the global model in terms of Euclidean distance, while constraint (5c) ensures the overall distance between the selected local models and their aggregated model is below the upper bound Υ, i.e., $\sum_{i=1}^{I}\beta'_{i,j}(t)d(\boldsymbol{w}_i(t), \bar{\boldsymbol{w}}(t)) \le \Upsilon$. This is because the defense mechanism at the FL server, e.g., Krum or multi-Krum, may perform local model selection to rule out those dissimilar to the rest. Constraint (5d) defines $\beta'_{i,j}(t)$ as a binary indicator.

IV. Variational Graph Auto-Encoders-Based Model Poisoning Attack on FL

Due to a lack of correlation between the arbitrary features in $\boldsymbol{w}'_j(t)$ and $\boldsymbol{w}_i(t)$, the malicious local model $\boldsymbol{w}'_j(t)$ could be detected by the server. For example, recent graph neural network (GNN)-based FL privacy protection schemes [21], [22] can classify the local model weights based on their features. To tackle this issue, we develop a new adversarial VGAE model in this section to generate $\boldsymbol{w}'_j(t)$ in such a way the individual feature correlation in $\boldsymbol{w}_i(t)$, $\forall i$ is captured in $\boldsymbol{w}'_j(t)$. As a result, the server can hardly detect

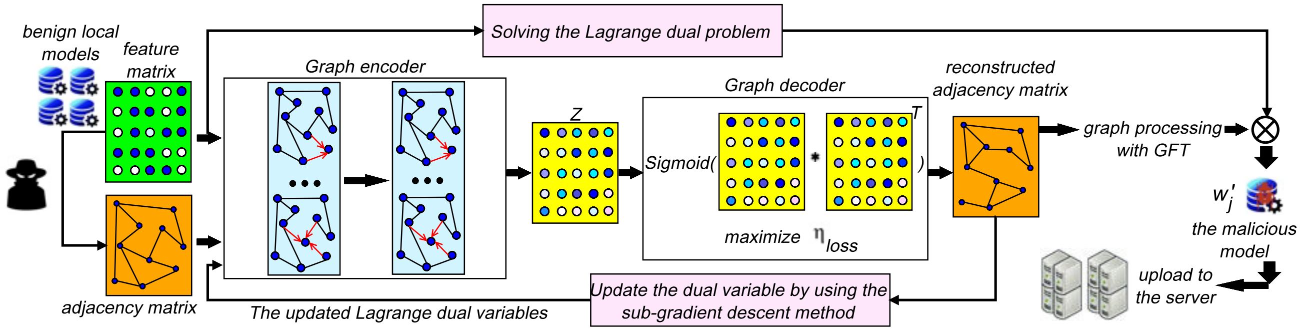

benign local models
feature matrix
Solving the Lagrange dual problem
Graph encoder
Z
Graph decoder
reconstructed adjacency matrix
Sigmoid(
*
)
T
maximize η_{loss}
graph processing with GFT
w'_j
the malicious model
upload to the server
adjacency matrix
The updated Lagrange dual variables
Update the dual variable by using the sub-gradient descent method

Fig. 2: The proposed VGAE-MP attack creates $\boldsymbol{w}'_j(t)$ based on learning the correlation among the parameters of the models being trained in FL, i.e., $\boldsymbol{w}_i(t)$, $\forall i$. A graph encoder trains $\boldsymbol{\mathcal{F}}_j$ and $\boldsymbol{\mathcal{A}}_j$ to build a feature representation matrix $\boldsymbol{\mathcal{Z}}$. The output of the encoder inputs to the decoder for the reconstruction of $\boldsymbol{\mathcal{A}}_j$. The VGAE-MP attack is designed to adjust $\boldsymbol{w}'_j$ to maximize the reconstruction loss η_{loss}, according to (19).

$\boldsymbol{w}'_j(t)$. For the brevity of notation, we omit the subscript $_j$ for the attacker in the following discussions.

The optimization of VGAE-MP in (5) is a non-convex combinatorial problem intractable for conventional optimization techniques. We decouple the VGAE-MP problem in (5) between the model attack and the bandwidth selection using the Lagrangian-dual method [23]. A new approach is developed to iteratively optimize the adversarial local models by running graph autoencoder and subgradient descent, as depicted in Fig. 2.

Let λ and ρ denote the dual variables. The Lagrange function of (5) is given by

$$\mathcal{L}(\beta'_{i,j}(t),\lambda,\rho) = F(\boldsymbol{w}'_G(t)) + \lambda(d_T - d(\boldsymbol{w}'_j(t),\boldsymbol{w}'_G(t))) + \rho\left(\Upsilon - \sum_{i=1}^{I}\beta'_{i,j}(t)d(\boldsymbol{w}_i(t),\bar{\boldsymbol{w}}(t))\right). \tag{6}$$

The Lagrange dual function is

$$\mathcal{D}(\lambda,\rho) = \max_{\boldsymbol{w}'_j(t),\beta'_{i,j}(t)} \mathcal{L}(\beta'_{i,j}(t),\lambda,\rho). \tag{7}$$

The dual problem of the primary problem in (5) is

$$\min_{\lambda,\rho} \mathcal{D}(\lambda,\rho). \tag{8}$$

A. Client Selection

At communication round t, given $\lambda = \lambda(t)$ and $\rho = \rho(t)$, the primary variable $\beta'_{i,j}(t)$ of the bandwidth selection can be optimized by solving

$$\beta'_{i,j}(t)^* = \arg\min_{\beta'_{i,j}(t)} \left\{\sum_{i=1}^{I}\beta'_{i,j}(t)d(\boldsymbol{w}_i(t),\bar{\boldsymbol{w}}(t))\right\}, \text{ s.t. (5d)}, \tag{9}$$

which is a standard 0/1 knapsack problem and can be readily solved using dynamic programming.

B. Generation of Adversarial Local Models

A new adversarial VGAE model, leveraging unsupervised learning on graph-structured data according to the variational auto-encoder [24], is proposed to maximize the Lagrange function (6). For given $\beta'_{i,j}(t)^*$, $\lambda(t)$ and $\rho(t)$, we optimize $\boldsymbol{w}'_j(t)$ by solving

$$\boldsymbol{w}'_j(t)^* = \arg\max_{\boldsymbol{w}'_j(t)}\Big\{F(\boldsymbol{w}'_G(t)) - \lambda(t)d(\boldsymbol{w}'_j(t),\boldsymbol{w}'_G(t)) + \rho(t)\sum_{i=1}^{I}\beta'_{i,j}(t)^* d(\boldsymbol{w}_i(t),\bar{\boldsymbol{w}}(t))\Big\}. \tag{10}$$

An attacker positioned within the effective range of the benign device's wireless signal can overhear the transmitted $\boldsymbol{w}_i(t)$ to the server. The level of access an attacker might have depends on the eavesdropping capabilities. For example, standard wireless signals are broadcast in a spherical radius around the transmitting device. This means that the attacker within that radius can have access to the broadcasted signal, where the attacker can capture the traffic and observe the transmitted information. More advanced attackers might employ methods that allow them to extend the range of their eavesdropping capabilities or to focus on specific directions, allowing them to intercept communications from further away. For example, a highly directional antenna can pick up wireless signals from a much greater distance than a standard antenna.

An attacker, i.e., the j-th attacker, can observe the local model parameters of the benign devices to establish the intrinsic correlation between the different parameters of the local models. A graph can be used to characterize the correlation. The graph is then regenerated manipulatively with the VGAE, and used to produce the malicious local model $\boldsymbol{w}'_j(t)$. By this means, we can maximize (10) while preventing the convergence of $\boldsymbol{w}'_G(t)$. Constraints (5b)-(5d) are satisfied by designing the decoder of the VGAE to reproduce the correlations. This approach reduces structural dissimilarity between $\boldsymbol{w}_i(t)$ and $\boldsymbol{w}'_j(t)$, which invalidates the existing defense mechanisms. The VGAE is tailored to ensure those constraints and hinder the convergence of the global model by extracting the correlation features between benign local models and embedding the correlation features in graphs for malicious local model generation.

1) Graph Construction and Feature Extraction: As illustrated in Fig. 2, the graph represented by $\mathcal{G} = (\mathcal{V}, E, \mathcal{F})$ is utilized to characterize the correlations among the parameters of the models being trained in FL, i.e., $\boldsymbol{w}_i(t)$, $\forall i$ [25]. The vertexes, edges, and feature matrix of the graph are represented by $\mathcal{V}$, E, and $\mathcal{F}$, respectively. The VGAE comprises a graph convolutional network (GCN) encoder and an inner product decoder. The encoder encodes the graph data using its features, and the decoder takes the encoded output as input to reconstruct the original graph $\mathcal{G} = (\mathcal{V}, E, \mathcal{F})$ [26].

Let $\boldsymbol{\mathcal{F}}(t) = [\boldsymbol{w}_1(t), \cdots, \boldsymbol{w}_I(t)]^T \in \mathbb{R}^{I \times M}$ be the feature matrix containing all I benign local models at communication round t, where M is the dimension of the local model. Let $\boldsymbol{\omega}^m(t) \in \mathbb{R}^{I \times 1}$ be the m-th column of $\boldsymbol{\mathcal{F}}(t)$. We use $\delta_{m,m'}(t)$ to denote the *cosine similarity* between the $\boldsymbol{\omega}_m(t)$ and $\boldsymbol{\omega}_{m'}(t)$ at communication round t. $m, m' \in [1, M]$. $\delta_{m,m'}(t)$ is defined as [27]

$$\delta_{m,m'}(t) = \frac{(\boldsymbol{\omega}^m(t))^T \boldsymbol{\omega}^{m'}(t)}{\|\boldsymbol{\omega}^m(t)\| \cdot \|\boldsymbol{\omega}^m(t)\|}. \tag{11}$$

The adjacency matrix, denoted by $\boldsymbol{\mathcal{A}}(t) = [\delta_{m,m'}(t)] \in \mathbb{R}^{M \times M}$, is one of the inputs to the encoder of the VGAE model at the attacker. According to $\boldsymbol{\mathcal{A}}(t)$, the topological structure of the graph $\mathcal{G}$ can be constructed at the attacker. The feature matrix $\boldsymbol{\mathcal{F}}(t)$ is the other input to the encoder of the VGAE model at the attacker.

2) Encoder design of the VGAE model: The encoder in the proposed VGAE maps $\mathcal{G}$ to a lower-dimensional representation. We build the encoder based on the GCN architecture, which learns a latent representation that captures the underlying features of $\mathcal{G}$. The encoded representation is then used as input to the decoder to reconstruct the original graph from the lower-dimensional representation to obtain the malicious local model $\boldsymbol{w}'_j(t)$ in (9). For the brevity of notation, we omit the index of communication rounds "t" in the following discussions.

A graph encoder is defined as

$$\boldsymbol{\mathcal{Z}}_1 = f_{\text{relu}}(\boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{A}}, |\boldsymbol{W}_0); \tag{12}$$

$$\boldsymbol{\mathcal{Z}}_2 = f_{\text{linear}}(\boldsymbol{\mathcal{Z}}_1, \boldsymbol{\mathcal{A}}|\boldsymbol{W}_1), \tag{13}$$

where $f_{\text{relu}}(\cdot)$ is the Rectified Linear Unit (ReLU) activation function employed for the first layer, while $f_{\text{linear}}(\cdot)$ is the Linear activation function used for the second layer; and $\boldsymbol{W}_l$ is the learnable parameters specific to the l-th layer of the neural networks.

Since determining the probability distribution of the latent representation of vertexes $\boldsymbol{\mathcal{Z}}$ in $\mathcal{G}$ is difficult and intractable [28], we approximate the true posterior by using a Gaussian distribution $\mathcal{N}(\cdot)$, while the encoder takes $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{A}}$ as its input to an inference model parameterized by a two-layer GCN. Thus, we have

$$q(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}}) = \Pi_{m=1}^{M} q(\boldsymbol{z}_m|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}}), \tag{14}$$

and

$$q(\boldsymbol{z}_m|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}}) = \mathcal{N}(\boldsymbol{z}_m|\boldsymbol{\mu}_m, \text{diag}(\boldsymbol{\sigma}^2)), \tag{15}$$

where $\boldsymbol{\mu} = \boldsymbol{\mathcal{Z}}_2$ builds the matrix of mean vectors $\boldsymbol{\mu}_m$. Likewise, we have $\log \boldsymbol{\sigma} = f_{\text{linear}}(\boldsymbol{\mathcal{Z}}_1, \boldsymbol{\mathcal{A}}|\boldsymbol{W}_1)$ that shares the first-layer parameters $\boldsymbol{W}_0$.

With the identity matrix $\mathcal{I} \in \mathbb{R}^{M \times M}$, we define $\widetilde{\boldsymbol{\mathcal{A}}} = \boldsymbol{\mathcal{A}} + \mathcal{I}$ with the (m, m')-th element $\widetilde{\boldsymbol{\mathcal{A}}}_{m,m'}$, and the (diagonal) degree matrix $\boldsymbol{\mathcal{D}}$ with the (m, m)-th element $\boldsymbol{\mathcal{D}}_{m,m} = \sum_{m'=1}^{M} \widetilde{\boldsymbol{\mathcal{A}}}_{m,m'}$. Each layer of the GCN can be written as

$$f_{\mathcal{G}}(\boldsymbol{\mathcal{Z}}_{l-1}, \boldsymbol{\mathcal{A}}|\mathbf{W}_l) = \phi(\boldsymbol{\mathcal{D}}^{-\frac{1}{2}} \widetilde{\boldsymbol{\mathcal{A}}} \boldsymbol{\mathcal{D}}^{-\frac{1}{2}} \boldsymbol{\mathcal{Z}}_{l-1} \mathbf{W}_l), \tag{16}$$

where $\phi(\cdot)$ is the activation function such as relu$(\cdot)$.

3) Decoder design of the VGAE model: The input to the decoder of the proposed VGAE model is $\boldsymbol{\mathcal{Z}}$, which is the output of the GCN in the encoder. The decoder aims to reconstruct $\boldsymbol{\mathcal{A}}$, denoted by $\widehat{\boldsymbol{\mathcal{A}}}$, predicting whether there is a link between two vertexes by an inner product between latent variables, which is designed as

$$p(\widehat{\boldsymbol{\mathcal{A}}}|\boldsymbol{\mathcal{Z}}) = \sum_{m=1}^{M} \sum_{m'=1}^{M} p(\hat{\delta}_{m,m'}|\boldsymbol{z}_m, \boldsymbol{z}_{m'}); \tag{17}$$

$$p(\hat{\delta}_{m,m'} = 1|\boldsymbol{z}_m, \boldsymbol{z}_{m'}) = \text{sigmoid}(\boldsymbol{z}_m^T \boldsymbol{z}_{m'}), \tag{18}$$

where $\boldsymbol{z}_m \in \mathbb{R}^{M \times 1}$ is the m-th column of $\boldsymbol{\mathcal{Z}}$, and sigmoid$(\cdot)$ is the logistic sigmoid function, i.e., sigmoid$(x) = 1/(1 + \exp^{-x})$. Here, the larger the inner product $(\boldsymbol{z}_m^T \boldsymbol{z}_{m'})$ in the embedding, the more likely vertexes m and m' are connected in the graph, according to $\widehat{\boldsymbol{\mathcal{A}}} = [\hat{\delta}_{m,m'}] \in \mathbb{R}^{M \times M}$ in the autoencoder [29].

We can view (17) as the inverse operation of the encoder for constructing a reconstructed adjacency matrix $\widehat{\boldsymbol{\mathcal{A}}}$ as the output of the decoder. A reconstruction loss function η_{loss} is defined at the decoder to measure the difference between $\boldsymbol{\mathcal{A}}$ and $\widehat{\boldsymbol{\mathcal{A}}}$. Given (14) and (17), η_{loss} is given as

$$\eta_{\text{loss}} = \mathbb{E}_{q(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}})}\left[\log p(\widehat{\boldsymbol{\mathcal{A}}}|\boldsymbol{\mathcal{Z}})\right] - \Phi[q(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}})|p(\boldsymbol{\mathcal{Z}})], \tag{19}$$

where $p(\boldsymbol{\mathcal{Z}}) = \Pi_m p(\boldsymbol{z}_m) = \Pi_m \mathcal{N}(\boldsymbol{z}_m|0, \mathcal{I})$ provides a Gaussian prior, and $\Phi[q(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}})|p(\boldsymbol{\mathcal{Z}})]$ provides the Kullback-Leibler divergence [30] between $q(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{F}})$ and $p(\boldsymbol{\mathcal{Z}})$.

4) Generation of adversarial local models $\boldsymbol{w}'_j(t)$: The Laplacian matrix of $\mathcal{G}$ [31] is built based on the adjacency matrix of the benign models, i.e., $\boldsymbol{\mathcal{A}}$, as given by

$$\mathcal{L} = diag(\boldsymbol{\mathcal{A}}) - \boldsymbol{\mathcal{A}}. \tag{20}$$

By applying singular value decomposition (SVD) [32] to $\mathcal{L}$, i.e., $\mathcal{L} = B \Sigma B^T$, we can obtain a complex unitary matrix $B \in \mathbb{R}^{J \times J}$, also known as graph Fourier transform (GFT) basis, that is used to transform graph data, e.g., $\boldsymbol{\mathcal{F}}$, to its spectral-domain representation. $\Sigma \in \mathbb{R}^{J \times J}$ is a diagonal matrix with the eigenvalues of $\mathcal{L}$ along its main diagonal.

Due to the abundance of local training data at a client, $\boldsymbol{w}_i^m(t)$ typically contains numerous model parameters, i.e., $M \gg 1$, which leads to a large size of $\boldsymbol{\mathcal{A}} = \{\delta_{m,m'}\} \in \mathbb{R}^{M \times M}$. The exact SVD of $\mathcal{L}$ that has an $M \times M$ matrix has time complexity $\mathcal{O}(M^3)$, which is infeasible in the presence of a large $\boldsymbol{\mathcal{A}}$. To reduce the dimensionality of $\boldsymbol{\mathcal{A}}$ while preserving the features, we consider a fast low-rank SVD approximation [33], which retains the k singular values and their corresponding singular vectors, where $k \ll M^3$. In particular, a truncated SVD of $\mathcal{L}$ can be formulated as $\mathcal{L}_k \approx B_k \Sigma_k B_k^T$, where B_k is an $m \times k$ matrix with columns being the first k left singular vectors of $\mathcal{L}$, Σ_k is a $k \times k$ diagonal matrix with entries being the first k singular values of $\mathcal{L}$, and B_k is an $n \times k$ matrix with columns being the first k right singular vectors of $\mathcal{L}$.

With B (or more explicitly, B_k), an attacker, i.e., attacker j, can obtain a matrix S that contains the spectral-domain data features of all $\boldsymbol{\omega}^m(t), \forall m$ by removing the correlations among the models and subsequently focusing on the data features substantiating the local models. S is given by [34]

$$S = B_k^{-1} \boldsymbol{\mathcal{F}}. \tag{21}$$

Likewise, the attacker can produce a Laplacian matrix based on the output of the VGAE, as given by

$$\widehat{\mathcal{L}} = diag(\widehat{\boldsymbol{\mathcal{A}}}) - \widehat{\boldsymbol{\mathcal{A}}}. \tag{22}$$

The corresponding GFT basis, denoted by $\widehat{B}_k$, can be obtained by applying the fast low-rank SVD approximation to $\widehat{L}$. With reference to (21), the malicious local model that follows $\boldsymbol{\mathcal{A}}$ in the VGAE can be determined by

$$\widehat{\boldsymbol{\mathcal{F}}} = \widehat{B}_k S, \tag{23}$$

where $\widehat{\boldsymbol{\mathcal{F}}} \in \mathbb{R}^{I \times M}$. The j-th row vector of $\widehat{\boldsymbol{\mathcal{F}}}$ is selected as the malicious local model $\boldsymbol{w}_j'(t)$ and uploaded by the j-th attacker to the aggregator for global model aggregation in communication round t.

C. Update of Dual Variables

Given the attack model $\boldsymbol{w}_j'(t)$, with the obtained $\beta_{i,j}'(t)^*$, the sub-gradient descent method can be taken to update $\lambda(t)$ and $\rho(t)$ by solving the dual problem (8). Specifically, $\lambda(t)$and $\rho(t)$ are updated by [35]

$$\lambda\left(t+1\right) = \left[\lambda(t) - \varepsilon\left(d(\boldsymbol{w}_j'(t), \boldsymbol{w}_G'(t)) - d_T\right)\right]^+; \tag{24a}$$

$$\rho\left(t+1\right) = \left[\rho(t) - \varepsilon\left(\sum_{i=1}^{I} \beta_{i,j}'(t)^* d(\boldsymbol{w}_i(t), \bar{\boldsymbol{w}}(t)) - \Upsilon\right)\right]^+, \tag{24b}$$

where ε is the step size, and $[x]^+ = \max\left(0, x\right)$. At initialization, $\lambda(t)$ and $\rho(t)$ are non-negative, i.e., $\lambda(1) \geq 0$ and $\rho(1) \geq 0$, to ensure (24) converges.

Since the attacker aims to generate the malicious local models to disorient FL, the proposed VGAE is constructed and trained to maximize η_{loss}. As a consequence, $\boldsymbol{w}_j'(t)$ progressively and increasingly contaminates the FL training, as global model aggregations increase, i.e., $t = 1, 2, 3, \cdots$.

D. Algorithm Design of The VGAE-MP Attack

According to the design of the new VGAE-MP attack in Figure 2, Algorithm 1 is developed along with the FL training of the benign user devices and the FL server. Specifically, the FL server broadcasts $\boldsymbol{w}_G'$ in every communication round. Each benign node i $(1 \leq i \leq I)$ applies the LocalTraining_start($\boldsymbol{w}_G'$) function for training the local model $\boldsymbol{w}_i$. Each attacker, i.e., the j-th attacker $(1 \leq j \leq J)$, overhears the global model $\boldsymbol{w}_G'$ and the local model $\boldsymbol{w}_i$ from the benign nodes. The GAE is trained to maximize the reconstruction loss with $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{F}}$. At the output of the GAE, the attacker achieves the optimal malicious local model, i.e., $\boldsymbol{w}_j'$. Then, $\boldsymbol{w}_j'$ is uploaded to the FL server for aggregation. As $\boldsymbol{w}_j'$ is highly correlated with $\boldsymbol{w}_i$ from the benign user devices, the FL server is unlikely to detect and identify the attacker.

Algorithm 1 The proposed VGAE-MP attack algorithm

```
1. Initialize: $\mathcal{G} = (\mathcal{V}, E, \mathcal{F})$, $T_L$, $I$, $J$, $d_T$, $\boldsymbol{w}_G'(t)$,
   $\boldsymbol{w}_i^m(t)$, and $\lambda(1) \geq 0$.
   % Adversarial FL:
for round $t = 1, 2, 3, \cdots$ do
    for Local iteration number $l = 1, \cdots, T_L$ do
        All benign user devices train their benign local
        model $\boldsymbol{\omega}_i(t)$, $i = 1, \cdots, I$.
    end for
    All benign user devices upload their benign local
    models $\boldsymbol{w}_i(t)$, $i = 1, \cdots, I$ to the server, and the
    attackers overhear the benign local models.
    The attacker $j$ carries out the proposed VGAE, i.e.,
    VGAE($\boldsymbol{\omega}^m(t), \forall m, \boldsymbol{\mathcal{F}}, \lambda(t)$), and obtains $\boldsymbol{w}_j'(t)$:
        · Build the adjacency matrix $\boldsymbol{\mathcal{A}} = [\delta_{m,m'}] \in$
        $\mathbb{R}^{M \times M}$ according to (11), and input $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{F}}$
        into the VGAE.
        · Train the VGAE to maximize the reconstruction
        loss $\eta_{\text{loss}}$ to obtain $\widehat{\boldsymbol{\mathcal{A}}}$.
        · Obtain $S$ based on (20) and (21), next obtain
        $\widehat{\boldsymbol{\mathcal{F}}}$ based on (22) and (23), and then determine
        $\boldsymbol{\omega}^m(t)$ based on $\widehat{\boldsymbol{\mathcal{F}}}$.
    Update $\lambda(t)$, according to (24).
    The attacker uploads the malicious local model $\boldsymbol{w}_j'(t)$
    to the server.
    The server aggregates selected local models to obtain
    the global model under attack $\boldsymbol{w}_G'(t)$ by (3), and
    broadcasts $\boldsymbol{w}_G'(t)$.
    All benign user devices update their local models
    with the global model, i.e., $\boldsymbol{w}_i(t) \leftarrow \boldsymbol{w}_G'(t)$, $\forall i$.
end for
```

Note that an attacker positioned in proximity to benign devices and equipped with radio transceivers has the potential to passively eavesdrop on the transmitted local models from one or more benign devices. This allows the attacker to discern their characteristics and subsequently devise a malicious local model. The more benign local models are overheard, the more profound the exploration into the feature correlation between the benign local and global models, and the more unlikely the malicious local models are detected by the server. The VGAE-MP attack remains operational even if only a single benign local model is overheard, though its effectiveness is diminished compared to scenarios where multiple benign local models are eavesdropped upon.

Although cryptography can prevent eavesdropping attacks to some extent, existing techniques, such as those

TABLE II: The setting of parameters in PyTorch

Parameters	Values
number of benign devices (I)	5 $\sim$ 30
number of attackers (J)	1 $\sim$ 5
communication rounds of the FL	100
number of local iterations (T_L)	10
model parameters in $\boldsymbol{w}_i(t)$ (M)	100, 200, or 300
1st hidden layer size of the VGAE	32
2nd hidden layer size of the VGAE	16
learning rate of the VGAE	0.01
batch size of the SVM	30
learning rate of the SVM	0.001
regularization of the SVM loss function	0.01
k-Fold cross-validation	5

developed in [36] and [37], have demonstrated the possibility of deciphering encrypted information with limited initial data. This risk is even more threatening with the rapid advancement of Quantum computing. The proposed data-untethered VGAE-MP attack could potentially work in compiling with these attack techniques to evade cryptographic protection of the benign local models and poison the training of FL.

V. Performance Evaluation

This section demonstrates the implementation of the proposed new VGAE-MP attack in PyTorch. Based on MNIST handwritten digits [38], FashionMNIST and CIFAR-10 datasets [39], the training accuracy of the local and global models under the attack is evaluated. The detection rate of the VGAE-MP attack is also presented, which is measured according to the Euclidean distance between the malicious local model and the benign one. The source code of the proposed VGAE-MP attack is available on GitHub: https://github.com/jjzgeeks/VGAE-based_Model_Poisoning_Attack_FL.

A. Implementation of The VGAE-MP Attack

The benign FL is designed to improve image classification accuracy, while the proposed VGAE-MP attack aims to reduce accuracy and cause label misclassification. The number of benign devices I increases from 5 to 30, while the number of attackers J increases from 1 to 5. The global model $\boldsymbol{w}'_G(t)$ in FL is trained with 100 communication rounds, and training of the local model $\boldsymbol{w}_i(t)$ is carried out in 10 iterations. For building the adjacency matrix $\boldsymbol{\mathcal{A}}$ at the attacker, the number of selected model parameters in $\boldsymbol{w}_i(t)$, i.e., M, is set to 100, 200, or 300. The VGAE encoder is a two-layer GCN network with a dropout layer to prevent overfitting. The VGAE decoder is an inner product. The Adam optimizer with a learning rate of 0.01 is adopted to optimize the network. For all datasets, we use the same encoder, decoder and SVM models. Table II lists the setting of parameters in PyTorch.

The proposed VGAE-MP attack is implemented on an SVM model using PyTorch 1.12.1, Python 3.9.12 on a Linux workstation with an Intel(R) Core(TM) i7-9700K CPU@3.60GHz (8 cores) and 16 GB of DDR4 memory@2400 MHz. The experiments are carried out on three datasets:

- The standard MNIST dataset, comprising 60,000 training and 10,000 testing grayscale images of handwritten digits from 1 to 10;
- The FashionMNIST dataset, comprising Zalando's article grayscale images with the size of 28 × 28 in ten classes, including 60,000 and 10,000 images for training and testing, respectively;
- The CIFAR-10 dataset, consisting of 60,000 images with the size of 32 × 32 in ten classes (6,000 per class), 50,000 for training and 10,000 for testing.

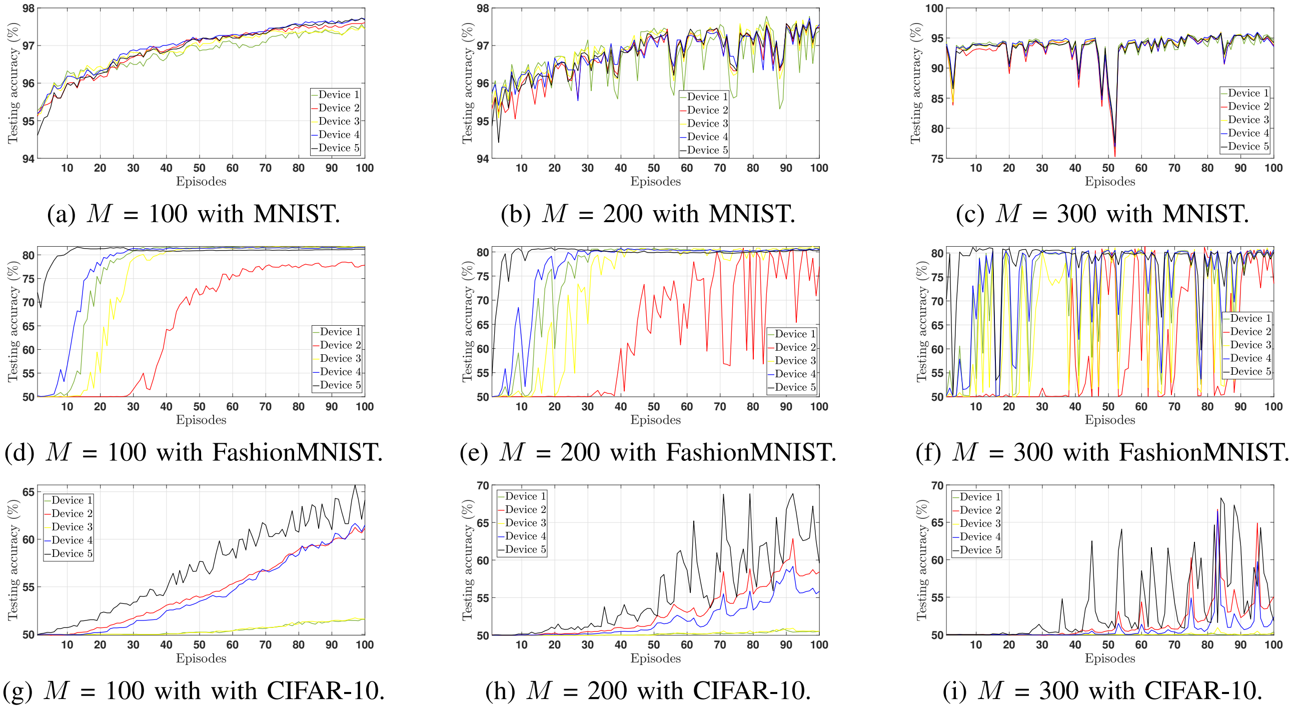
At each benign user device, a standard quadratic optimization algorithm is utilized to train the SVM models with the three datasets. The loss function of the SVM models is $F_i(\boldsymbol{w}_i(t)) = \frac{1}{2}\|\boldsymbol{w}_i(t)\|_2^2 + \frac{1}{D_i}\sum_{i=1}^{I}\max\left\{0, 1-y_i^{d_i}(\beta_i+\boldsymbol{\omega}_i^T(t)x_i^{d_i})\right\}$, where β_i is a parameter that can be obtained based on $\boldsymbol{w}_i(t)$.

In addition, the proposed VGAE-MP attack is compared with an existent data-agnostic model poisoning (MP) attack that produces malicious local models by mimicking other benign devices' training samples to degrade the learning accuracy. The MP attack considered for comparison has been used in several existing studies, e.g., [40], [41], where the attacker manipulates the training process by injecting a fake device and sending fake local models to the server. Moreover, we implemented another existing attack on FL, i.e., a random MP (RMP) attack considered in [9], [11]. Specifically, RMP generates the malicious local model by injecting a Gaussian random noise into the received global model, which can enlarge the magnitudes of the random local model updates using a scaling factor.

B. Attacking Performance

In Fig. 3, we plot the local model's testing accuracy with 100 FL communication rounds under the proposed VGAE-MP attack on the MNIST, FashionMNIST, and CIFAR-10. When M of VGAE-MP increases from 100 to 300, the FL accuracy fluctuates dramatically, successfully restraining the convergence of the testing accuracy. Using FashionMNIST as an example, the FL accuracy of the five benign devices converges to 80% in 3(d) under the VGAE-MP attack with M = 100. Once M increases to 300 in 3(f), the FL accuracy of the five benign devices consistently experiences fluctuations between 50% and 80%. This confirms that M determines the size of features in $\boldsymbol{\omega}^m(t)$ whose correlation in $\boldsymbol{\mathcal{A}}$ is learned to generate the malicious poisoning model $\boldsymbol{w}'_j(t)$. Therefore, a large M leads to a more complete graph trained by the VGAE model.

In Figs. 3(a) to 3(i), we interestingly observe that VGAE-MP demonstrates more prominent attacking performance on the FL with the FashionMNIST and CIFAR-10 than the one with the MNIST. This might be attributed to the variances in the MNIST, FashionMNIST, and CIFAR-10. MNIST comprises grayscale images of handwritten digits, whereas FashionMNIST houses grayscale images of apparel and accessories. CIFAR-10, on the other hand, has 10 distinct categories of objects, including animals, vehicles, among others. The simplicity of the MNIST's handwritten digits

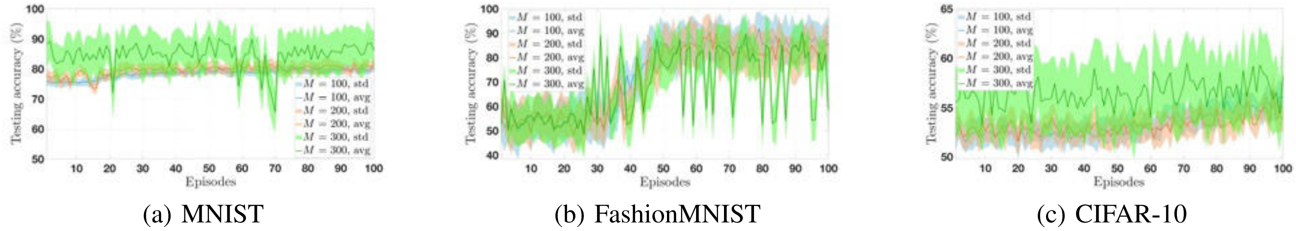


(a) M = 100 with MNIST. (b) M = 200 with MNIST. (c) M = 300 with MNIST.

(d) M = 100 with FashionMNIST. (e) M = 200 with FashionMNIST. (f) M = 300 with FashionMNIST.

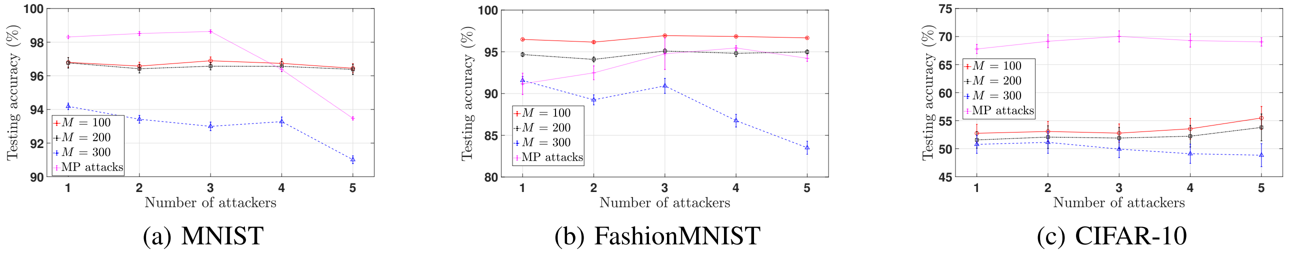
(g) M = 100 with with CIFAR-10. (h) M = 200 with CIFAR-10. (i) M = 300 with CIFAR-10.

Fig. 3: Given 100 FL communication rounds, I = 5 and J = 2, we study the local model's testing accuracy under the proposed VGAE-MP attack on the MNIST, FashionMNIST, and CIFAR-10 datasets.



(a) MNIST (b) FashionMNIST (c) CIFAR-10

Fig. 4: The global model's testing accuracy ("avg" means the average value and "std" stands for the standard deviation) under the VGAE-MP attack on the MNIST, FashionMNIST, and CIFAR-10 datasets.



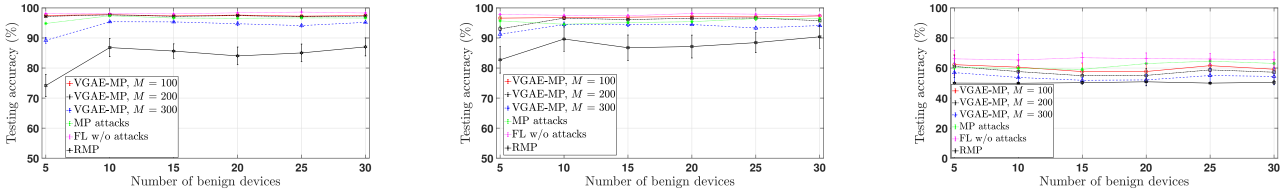
(a) MNIST (b) FashionMNIST (c) CIFAR-10

Fig. 5: Given the MNIST, FashionMNIST, and CIFAR-10 datasets, the average testing accuracy of the local models under the VGAE-MP attack when J increases from 1 to 5.

may make them more easily classified by FL compared to the more complex images found in FashionMNIST or CIFAR-10.

Fig. 4 shows the global model's testing accuracy measured at the server based on the MNIST, FashionMNIST, and CIFAR-10. Under the VGAE-MP attack, the steady convergence of FL accuracy is inhibited. In particular, for the CIFAR-10 with M = 300, the FL accuracy maintains around 58% under the VGAE-MP attack. Moreover, the VGAE-MP attack doesn't lead to a considerable decrease in the testing accuracy of the global model. This is because that a significant performance dip could potentially reveal the presence of the attacker.

Fig. 5 plots the average testing accuracy of the local



(a) MNIST (b) FashionMNIST (c) CIFAR-10

Fig. 6: Given $J = 5$, the average testing accuracy under the VGAE-MP attack on the MNIST, FashionMNIST, and CIFAR-10 datasets, where I increases from 5 to 30.

models under the VGAE-MP attack when J increases from 1 to 5. Since the number of benign devices is fixed at 3, the FL accuracy falls with the growth of the number of attackers. This is because the proposed VGAE-MP attack hinders the training convergence of FL. In particular, when M = 300, the average testing accuracy under the VGAE-MP attack drops about 5%, 12%, and 4% according to the MNIST, FashionMNIST, and CIFAR-10, respectively. When J = 5, the VGAE-MP attack outperforms the MP attack 10% and 20% given the FashionMNIST and CIFAR-10, respectively. The reason is the new VGAE-MP attack reconstructs the adversarial adjacency matrix according to the individual features of the devices. Consequently, the attacker falsifies the local models to maximize the FL loss.

Fig. 6 depicts the average testing accuracy of FL without the attack and FL under the VGAE-MP, MP or RMP attack, where J is set to 5 and I ranges from 5 to 30. As the benign devices increase, the FL accuracy under the VGAE-MP, MP and RMP attacks improves, given that the FL can quickly converge when the ratio of $\boldsymbol{w}_i(t)$ to $\boldsymbol{w}'_j(t)$ is heightened. On the three considered datasets, the FL accuracy is 6%, 5%, and 5% under the VGAE-MP attack lower than it is under the MP attack, respectively, when M = 300 and I = 5. The RMP attack has lower FL accuracy than the VGAE-MP and MP attacks. This is because the malicious local model update of the RMP attack is generated according to a Gaussian random noise, which is not correlated with any benign local models. However, this can make the malicious local models more easily detected and subsequently eliminated, as will be shown in Fig. 7.

Existing MP attacks in FL result in a high training loss of the FL model. One way to detect these malicious attacks is to compare the distance between the malicious local and global models with the distance between the benign local and global models. Suppose the distance between the malicious local and global models is larger. In such case, it can indicate a malicious attack, and the server can detect it accordingly.

To evaluate the invisibility of the proposed VGAE-MP attack, we study the distance between the local and the global models based on the CIFAR-10 datasets in Fig. 7, where I = 5 and J = 3. As shown in Figs. 7(a), 7(b), and 7(c), the Euclidean distances between the malicious local models generated by the new VGAE-MP attack and the corresponding global models are below that of the benign local models. This makes it difficult for the server to detect and defend against the attacker. In contrast, as shown in Figs. 7(d) and 7(e), the MP attack and the RMP attack result in a significantly larger distance between the malicious local and global models, making them easier to detect. This highlights the key strength of the proposed VGAE-MP attack, that is, VGAE-MP generates malicious local models based on the feature correlation between the benign local and global models, and hence makes the differences between the malicious and benign local models indistinguishable.

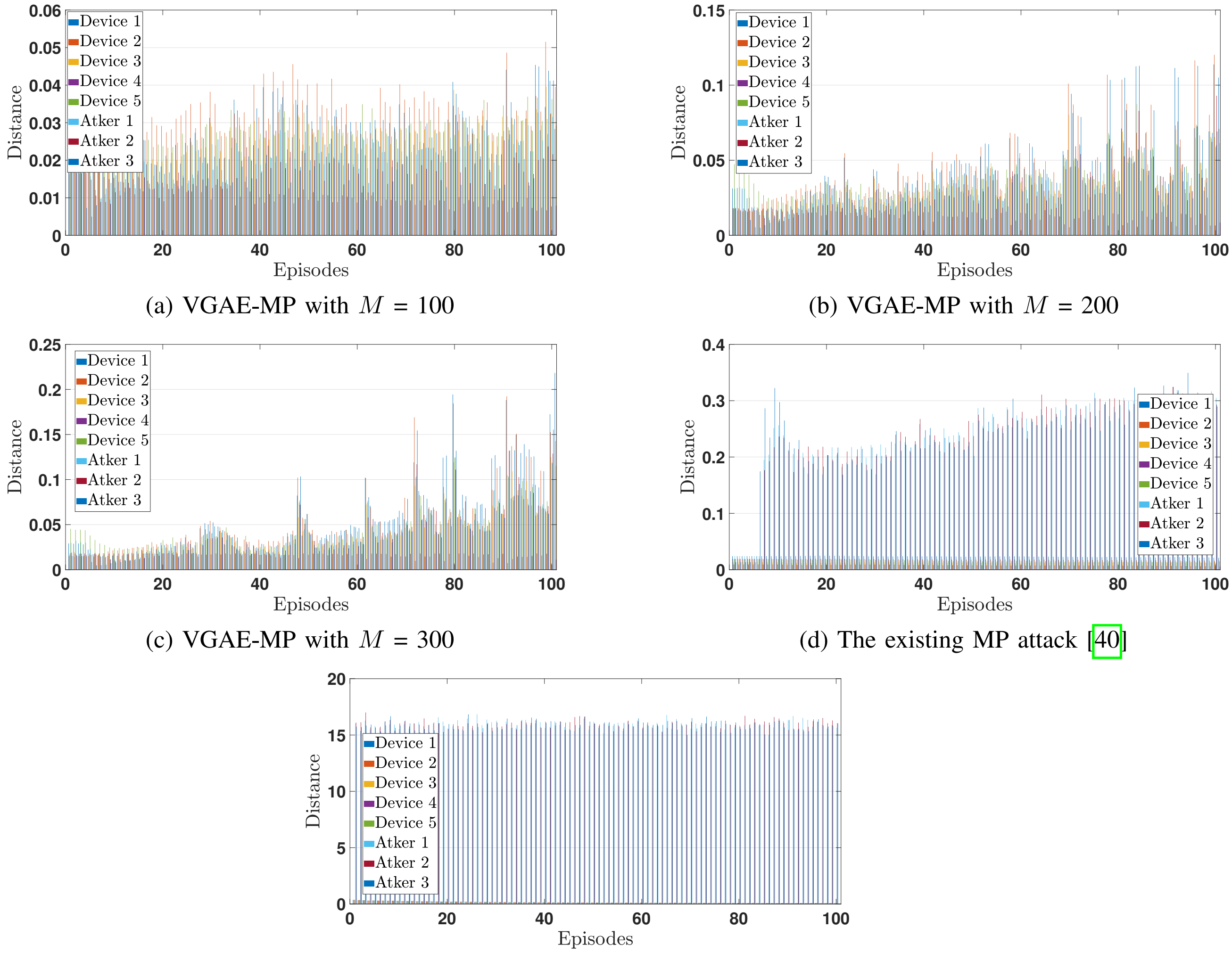
Fig. 8 shows the average testing accuracy under the proposed VGAE-MP attack on the MNIST, FashionMNIST, or CIFAR-10 dataset. This is observed as the attacker eavesdrops on an increasing number of benign user devices, ranging from 1 to 25. Generally, a noticeable fall in the local model updates' average accuracy is observed as the number of eavesdropped benign devices escalates. This trend is attributed to the attacker's ability to intercept more benign local models, thereby acquiring a broader range of correlation features. Such extensive data aids in crafting a more potent malicious model for effective system poisoning. In particular, the average accuracy on the MNIST, FashionMNIST, and CIFAR-10 datasets drops about 27.4%, 32.3%, and 24.9%, respectively.

VI. Conclusion and Future Work

In this paper, a new data-untethered VGAE-MP attack against FL was proposed, where the adversarial VGAE was developed to create malicious local models based solely on the benign local models overheard without access to the training data of FL. The proposed adversarial VGAE allows the attacker to extract the common underlying data features of the benign local models and their correlations to generate the malicious model with which the FL training loss is maximized. The VGAE-MP attack maintains the feature correlation between the benign local and global models, making the differences between the malicious and benign local models indistinguishable. The VGAE-MP attack on the FL was implemented using PyTorch with the source code released on GitHub. The performances were evaluated using the MNIST, fashionMNIST, and CIFAR-10 datasets.

The proposed data-untethered VGAE-MP attack involves a single poisoning objective, which aims to degrade the

2023.



(a) VGAE-MP with $M = 100$

(b) VGAE-MP with $M = 200$

(c) VGAE-MP with $M = 300$

(d) The existing MP attack [40]

(e) The RMP attack [11]

Fig. 7: Based on the CIFAR-10 training datasets, the Euclidean distances of the local models are measured at the server in order to detect a poisoning attack, where we set $I = 5$ and $J = 3$.

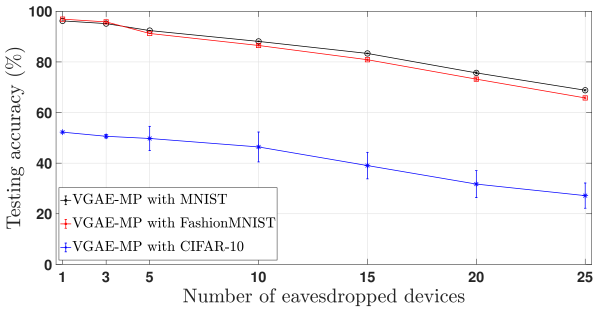


Fig. 8: The number of eavesdropped benign local model updates increases from 1 to 25, based on the MNIST, FashionMNIST, or CIFAR-10 datasets.

training accuracy of FL. In our future work, multiple performance metrics of FL will be considered in the poisoning, such as training fairness, robustness, and model utility. A multi-objective optimization will be formulated while the VGAE will be further studied to extract the graph representation.

Acknowledgements

This work was supported by the CISTER Research Unit (UIDP/UIDB/04234/2020) and project ADANET (PTDC/EEICOM/3362/2021), financed by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology).

References

[1] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, “Towards personalized federated learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[2] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang, and S. Y. Philip, “Privacy and robustness in federated learning: Attacks and defenses,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[3] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, “Manipulating machine learning: Poisoning attacks and countermeasures for regression learning,” in *IEEE symposium on security and privacy (SP)*. IEEE, 2018, pp. 19–35.

2023.

[4] Z. Wang, Y. Huang, M. Song, L. Wu, F. Xue, and K. Ren, “Poisoning-assisted property inference attack against federated learning,” *IEEE Transactions on Dependable and Secure Computing*, 2022.

[5] J. Gao, B. Hou, X. Guo, Z. Liu, Y. Zhang, K. Chen, and J. Li, “Secure aggregation is insecure: Category inference attack on federated learning,” *IEEE Transactions on Dependable and Secure Computing*, 2021.

[6] C. Fu, X. Zhang, S. Ji, J. Chen, J. Wu, S. Guo, J. Zhou, A. X. Liu, and T. Wang, “Label inference attacks against vertical federated learning,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1397–1414.

[7] X. Gong, Y. Chen, H. Huang, Y. Liao, S. Wang, and Q. Wang, “Coordinated backdoor attacks against federated learning with model-dependent triggers,” *IEEE network*, vol. 36, no. 1, pp. 84–90, 2022.

[8] F. Nuding and R. Mayer, “Data poisoning in sequential and parallel federated learning,” in *Proceedings of the 2022 ACM on International Workshop on Security and Privacy Analytics*, 2022, pp. 24–34.

[9] M. Fang, X. Cao, J. Jia, and N. Z. Gong, “Local model poisoning attacks to byzantine-robust federated learning,” in *Proceedings of the 29th USENIX Conference on Security Symposium*, 2020, pp. 1623–1640.

[10] V. Shejwalkar and A. Houmansadr, “Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning,” in *NDSS*, 2021.

[11] X. Cao and N. Z. Gong, “Mpaf: Model poisoning attacks to federated learning based on fake clients,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3396–3404.

[12] K.-H. Chow and L. Liu, “Perception poisoning attacks in federated learning,” in *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2021, pp. 146–155.

[13] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, “Poisongan: Generative poisoning attacks against federated learning in edge computing systems,” *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3310–3322, 2020.

[14] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, “Poisoning attack in federated learning using generative adversarial nets,” in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2019, pp. 374–380.

[15] X. Li, Z. Qu, S. Zhao, B. Tang, Z. Lu, and Y. Liu, “Lomar: A local defense against poisoning attack on federated learning,” *IEEE Transactions on Dependable and Secure Computing*, 2021.

[16] A. Qayyum, M. U. Janjua, and J. Qadir, “Making federated learning robust to adversarial attacks by learning data and model association,” *Computers & Security*, vol. 121, p. 102827, 2022.

[17] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, “Federated learning for online resource allocation in mobile edge computing: A deep reinforcement learning approach,” in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[18] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, “Model pruning enables efficient federated learning on edge devices,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[19] J. Zheng, K. Li, N. Mhaisen, W. Ni, E. Tovar, and M. Guizani, “Exploring deep reinforcement learning-assisted federated learning for online resource allocation in privacy-preserving edgeIoT,” *IEEE Internet of Things Journal*, 2022.

[20] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, “Machine learning with adversaries: Byzantine tolerant gradient descent,” *Advances in neural information processing systems*, vol. 30, 2017.

[21] B. Wang, A. Li, M. Pang, H. Li, and Y. Chen, “Graphfl: A federated learning framework for semi-supervised node classification on graphs,” in *IEEE International Conference on Data Mining*. IEEE, 2022, pp. 498–507.

[22] X. Yuan, J. Chen, J. Yang, N. Zhang, T. Yang, T. Han, and A. Taherkordi, “Fedstn: Graph representation driven federated learning for edge computing enabled urban traffic flow prediction,” *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[23] D. P. Palomar and M. Chiang, “A tutorial on decomposition methods for network utility maximization,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[24] T. Cemgil, S. Ghaisas, K. Dvijotham, S. Gowal, and P. Kohli, “The autoencoding variational autoencoder,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 077–15 087, 2020.

[25] K. Li, X. Yuan, J. Zheng, W. Ni, and M. Guizani, “Exploring adversarial graph autoencoders to manipulate federated learning in the internet of things,” in *International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2023, pp. 898–903.

[26] Y. Wang, B. Xu, M. Kwak, and X. Zeng, “A simple training strategy for graph autoencoder,” in *International Conference on Machine Learning and Computing*, 2020, pp. 341–345.

[27] D. Zhu, Y. Ma, and Y. Liu, “Anomaly detection with deep graph autoencoders on attributed networks,” in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–6.

[28] A. Hasanzadeh, E. Hajiramezanali, K. Narayanan, N. Duffield, M. Zhou, and X. Qian, “Semi-implicit graph variational auto-encoders,” *Advances in neural information processing systems*, vol. 32, 2019.

[29] S. Pan, R. Hu, S.-f. Fung, G. Long, J. Jiang, and C. Zhang, “Learning graph embedding with adversarial training methods,” *IEEE transactions on cybernetics*, vol. 50, no. 6, pp. 2475–2487, 2019.

[30] J. M. Joyce, “Kullback-leibler divergence,” in *International encyclopedia of statistical science*. Springer, 2011, pp. 720–722.

[31] J. J. Molitierno, *Applications of combinatorial matrix theory to Laplacian matrices of graphs*. CRC Press, 2016.

[32] K. Lange, “Singular value decomposition,” in *Numerical analysis for statisticians*. Springer, 2010, pp. 129–142.

[33] A. K. Menon and C. Elkan, “Fast algorithms for approximating the singular value decomposition,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 2, pp. 1–36, 2011.

[34] B. Shan, W. Ni, X. Yuan, D. Yang, X. Wang, and R. P. Liu, “Graph learning from band-limited data by graph fourier transform analysis,” *Signal Processing*, vol. 207, p. 108950, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168423000245

[35] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[36] S. Hebrok, S. Nachtigall, M. Maehren, N. Erinola, R. Merget, J. Somorovsky, and J. Schwenk, “We really need to talk about session tickets: A {Large-Scale} analysis of cryptographic dangers with {TLS} session tickets,” in *Proceedings of 32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 4877–4894.

[37] D. Diaz-Sanchez, A. Marín-Lopez, F. A. Mendoza, P. A. Cabarcos, and R. S. Sherratt, “Tls/pki challenges and certificate pinning techniques for iot and m2m secure communications,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3502–3531, 2019.

[38] L. Deng, “The mnist database of handwritten digit images for machine learning research [best of the web],” *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[39] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.

[40] M. T. Hossain, S. Islam, S. Badsha, and H. Shen, “Desmp: Differential privacy-exploited stealthy model poisoning attacks in federated learning,” in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*. IEEE, 2021, pp. 167–174.

[41] X. Cao, Z. Zhang, J. Jia, and N. Z. Gong, “Flcert: Provably secure federated learning against poisoning attacks,” *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3691–3705, 2022.

2023.

Kai Li (S'09–M'14–SM'20) received the B.E. degree from Shandong University, China, in 2009, the M.S. degree from The Hong Kong University of Science and Technology, Hong Kong, in 2010, and the Ph.D. degree in computer science from The University of New South Wales, Sydney, NSW, Australia, in 2014. Currently, he is a Visiting Research Scientist with the Division of Electrical Engineering, Department of Engineering, University of Cambridge, U.K., and a Senior Research Scientist with the CISTER Research Centre, Porto, Portugal. He is also a CMU-Portugal Research Fellow, jointly supported by Carnegie Mellon University (CMU), Pittsburgh, PA, USA, and the Foundation for Science and Technology (FCT), Lisbon, Portugal. In 2022, he was a Visiting Research Scholar with the CyLab Security and Privacy Institute, CMU. Prior to this, he was a Post-Doctoral Research Fellow with the SUTD-MIT International Design Centre, Singapore University of Technology and Design, Singapore, from 2014 to 2016. He has also held positions as a Visiting Research Assistant with the ICT Centre, CSIRO, Brisbane, QLD, Australia, from 2012 to 2013, and a full-time Research Assistant with the Mobile Technologies Centre, The Chinese University of Hong Kong, Hong Kong, from 2010 to 2011. He has been an Associate Editor of journals, such as *Internet of Things* (Elsevier) since 2024, *Nature Computer Science* (Springer) since 2023, *Computer Communications* (Elsevier) and *Ad Hoc Networks* (Elsevier) since 2021, and IEEE ACCESS from 2018 to 2024.

Xin Yuan (M'19–SM'24) received the B.E. degree in Communication Engineering from Taiyuan University of Technology, Shanxi, China, in 2013, and the dual Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, and the University of Technology Sydney (UTS), Sydney, Australia, in 2019 and 2020, respectively. She is currently a Senior Research Scientist at CSIRO, Sydney, NSW, Australia. She is also an Adjunct Senior Lecturer at the University of New South Wales (UNSW). Her research interests include machine learning and optimization, and their applications to the integrity, efficiency, and security of intelligent systems and networks. She has been an Editor for IEEE Transactions on Vehicular Technology since 2023.

Jingjing Zheng (S'22) is currently near to completion of pursuing the Ph.D. degree in electrical and computer engineering with the University of Porto, Porto, Portugal. He is a Student Researcher with CISTER Research Center, Porto, Portugal. In 2022, he was a Visiting Research Scholar with the CyLab Security and Privacy Institute, CMU. His main research interests include federated learning, machine learning security, and edge computing.

Wei Ni (SM'15–F'24) received the B.E. and Ph.D. degrees in Electronic Engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. He is a Principal Research Scientist at CSIRO, Sydney, Australia. He is also a Conjoint Professor at the University of New South Wales, an Adjunct Professor at the University of Technology Sydney, and an Honorary Professor at Macquarie University. He also serves as a Technical Expert at Standards Australia in support of the ISO standardization of AI and Big Data. He was a Postdoctoral Research Fellow at Shanghai Jiaotong University from 2005 to 2008; Deputy Project Manager at Bell Labs, Alcatel/Alcatel-Lucent from 2005 to 2008; and Senior Researcher at Devices R&D, Nokia from 2008 to 2009. He has co-authored one book, ten book chapters, more than 300 journal papers, more than 100 conference papers, 26 patents, ten standard proposals accepted by IEEE, and three technical contributions accepted by ISO. His research interests include 6G security and privacy, machine learning, stochastic optimization, and their applications to system efficiency and integrity.

Dr. Ni has been an Editor for IEEE Transactions on Wireless Communications since 2018, an Editor for IEEE Transactions on Vehicular Technology since 2022, and an Editor for IEEE Transactions on Information Forensics and Security and IEEE Communications Surveys and Tutorials since 2024. He served first as the Secretary, then the Vice-Chair and Chair of the IEEE VTS NSW Chapter from 2015 to 2022, Track Chair for VTC-Spring 2017, Track Co-chair for IEEE VTC-Spring 2016, Publication Chair for BodyNet 2015, and Student Travel Grant Chair for WPMC 2014.

Falko Dressler (F'17) received the M.Sc. and Ph.D. degrees from the Department of Computer Science, University of Erlangen, in 1998 and 2003, respectively. He is currently a Full Professor and the Chair of Telecommunication Networks with the School of Electrical Engineering and Computer Science, TU Berlin. He has been the Associate Editor-in-Chief of IEEE TRANSACTIONS ON MOBILE COMPUTING and *Computer Communications* (Elsevier) and an Editor of journals, such as IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, *Ad Hoc Networks* (Elsevier), and *Nano Communication Networks* (Elsevier). He has been chairing conferences, such as IEEE INFOCOM, ACM MobiSys, ACM MobiHoc, IEEE VNC, and IEEE GLOBECOM. He has authored the textbooks *Self-Organization in Sensor and Actor Networks* (Wiley & Sons) and *Vehicular Networking* (Cambridge University Press). He has been an IEEE Distinguished Lecturer and an ACM Distinguished Speaker. He is an ACM Distinguished Member. He is a member of the German National Academy of Science and Engineering (acatech). He has been serving on the IEEE COMSOC Conference Council and the ACM SIGMOBILE Executive Committee. His research objectives include adaptive wireless networking (sub-6GHz, mmWave, visible light, and molecular communication) and wireless-based sensing with applications in ad-hoc and sensor networks, the Internet of Things, and cyber-physical systems.

Abbas Jamalipour (S'86–M'91–SM'00–F'07) received the Ph.D. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1996. He is currently a Professor of ubiquitous mobile networking with The University of Sydney. He has authored nine technical books, 11 book chapters, over 550 technical papers, and five patents, all in the area of wireless communications and networking. He is a fellow of the Institute of Electrical, Information, and Communication Engineers (IEICE) and the Institution of Engineers Australia, an ACM Professional Member, and an IEEE Distinguished Speaker. Since 2014, he has been an elected member of the Board of Governors of the IEEE Vehicular Technology Society. He was a recipient of the number of prestigious awards, such as the 2019 IEEE ComSoc Distinguished Technical Achievement Award in Green Communications, the 2016 IEEE ComSoc Distinguished Technical Achievement Award in Communications Switching and Routing, the 2010 IEEE ComSoc Harold Sobol Award, the 2006 IEEE ComSoc Best Tutorial Paper Award, and over 15 best paper awards. He has been the General Chair and the Technical Program Chair of several prestigious conferences, including IEEE ICC, GLOBECOM, WCNC, and PIMRC. He was the President of the IEEE Vehicular Technology Society, from 2020 to 2021. Previously, he held the positions of the Executive Vice-President and the Editor-in-Chief of VTS Mobile World. He was the Vice President-Conferences and a member of the Board of Governors of the IEEE Communications Society. He sits on the editorial board of IEEE ACCESS and several other journals. He is a member of the Advisory Board of IEEE INTERNET OF THINGS JOURNAL. Since January 2022, he has been the Editor-in-Chief of IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was also the Editor-in-Chief of IEEE WIRELESS COMMUNICATIONS.