CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Explainable Graph Attention-Driven Fairness Manipulation for Federated Learning in EdgeIoT

Kai Li*
Jingjing Zheng*
Wei Ni*
Hailong Huang
Pietro Lio
Falko Dressler
Ozgur B. Akan

*CISTER Research Centre

CISTER-TR-250801

2025

Explainable Graph Attention-Driven Fairness Manipulation for Federated Learning in EdgeIoT

Kai Li*, Jingjing Zheng*, Wei Ni*, Hailong Huang, Pietro Lio, Falko Dressler, Ozgur B. Akan

*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, zheng@isep.ipp.pt, wei.ni@cister-labs.pt, hailong.huang@polyu.edu.hk, pl219@cam.ac.uk, dressler@ccs-labs.org, oba21@cam.ac.uk

https://www.cister-labs.pt

Abstract

This paper proposes an innovative adversarial architecture based on Explainable Graph AtTention-embedded autoEncoder (E-GATE), specifically designed to execute fairness manipulation that introduce biasing model updates into the federated learning in edge-based Internet of Things (EdgeIoT). E-GATE aims to generate biasing model updates by maximizing the minimum Kullback-Leibler (KL) divergence between a device 19s local model update and the global model. The E-GATE is trained with attention coefficients to obtain the hidden representations of each data feature in the explainable graph. Additionally, the graph autoencoder is incorporated within the E-GATE architecture to manipulatively reconstruct the correlations among model updates. This approach maximizes the reconstruction loss while keeping the biasing model updates undetected. The EGATE attack is implemented using PyTorch, and experimental results demonstrate that it successfully increases the minimum KL divergence of benign model updates by 70.2%, effectively evading detection by existing defense mechanisms.

Explainable Graph Attention-Driven Fairness Manipulation for Federated Learning in EdgeIoT

Kai Li
CISTER Research Centre
Porto, Portugal
kaili@ieee.org

Jingjing Zheng
CISTER Research Centre
Porto, Portugal
zheng@isep.ipp.pt

Wei Ni
CSIRO
Sydney, Australia
wei.ni@data61.csiro.au

Hailong Huang
The Hong Kong Polytechnical University
Hong Kong
hailong.huang@polyu.edu.hk

Pietro Liò
University of Cambridge
Cambridge, UK
pl219@cam.ac.uk

Falko Dressler
TU Berlin
Berlin, Germany
dressler@tkn.tu-berlin.de

Ozgur B. Akan
University of Cambridge
Cambridge, UK
oba21@cam.ac.uk

***Abstract*—This paper proposes an innovative adversarial architecture based on Explainable Graph AtTention-embedded autoEncoder (E-GATE), specifically designed to execute fairness manipulation that introduce biasing model updates into the federated learning in edge-based Internet of Things (EdgeIoT). E-GATE aims to generate biasing model updates by maximizing the minimum Kullback-Leibler (KL) divergence between a device's local model update and the global model. The E-GATE is trained with attention coefficients to obtain the hidden representations of each data feature in the explainable graph. Additionally, the graph autoencoder is incorporated within the E-GATE architecture to manipulatively reconstruct the correlations among model updates. This approach maximizes the reconstruction loss while keeping the biasing model updates undetected. The E-GATE attack is implemented using PyTorch, and experimental results demonstrate that it successfully increases the minimum KL divergence of benign model updates by 70.2%, effectively evading detection by existing defense mechanisms.**

***Index Terms*—Federated learning, EdgeIoT, fairness manipulation, bias, graph attention networks, graph autoencoder**

I. Introduction

Edge-based Internet of Things (EdgeIoT) represents a critical advancement that merges mobile edge computing with IoT devices, enabling data processing close to the data source [1]. By combining the computational power at the edge with the flexibility of IoT devices, EdgeIoT allows for immediate data analysis and decision-making, thus reducing latency and alleviating bandwidth constraints associated with sending data to cloud servers [2]. This capability is especially valuable for applications requiring prompt responses, such as emergency management, environmental monitoring, and smart city planning. EdgeIoT revolutionizes the way that data is processed, paving the way for agile intelligent systems in metaverse [3].

Federated learning greatly enhances data processing efficiency and decision-making in EdgeIoT [4]. As illustrated in Fig. 1, federated learning-enabled EdgeIoT involves IoT devices equipped with sensors and computational units that process data locally. Transmitting raw data to a server can be bandwidth-intensive and raise privacy concerns. Instead, federated learning allows each device to upload a shared machine-learning model update with locally processed data. These updates are the only data exchanged between devices and a server or among the IoT devices. Upon receiving the model updates from the devices, the server synthesizes these updates to create a global model. This global model is then transmitted back to the IoT devices, initiating the next round of federated learning. However, the reliance on aggregating updates from multiple devices makes the federated learning vulnerable to attacks [4]. Manipulated devices can inject harmful updates, which may distort the learning model and lead to biasing decisions. This risk is particularly significant in critical applications, such as emergency response, environmental monitoring, and urban surveillance.

To assess fairness on the server, one method involves calculating Kullback-Leibler (KL) divergence between an EdgeIoT device's model update and the global model. This metric quantifies the difference between the probability distributions of the local model updates and the aggregated global model. By measuring this discrepancy, the server can evaluate how much an individual EdgeIoT device's contribution deviates from the collective model, thereby determining the fairness of federated learning.

In this paper, we propose a novel manipulating model attack to compromise federated learning fairness in EdgeIoT. Our approach leverages an Explainable Graph AtTention-

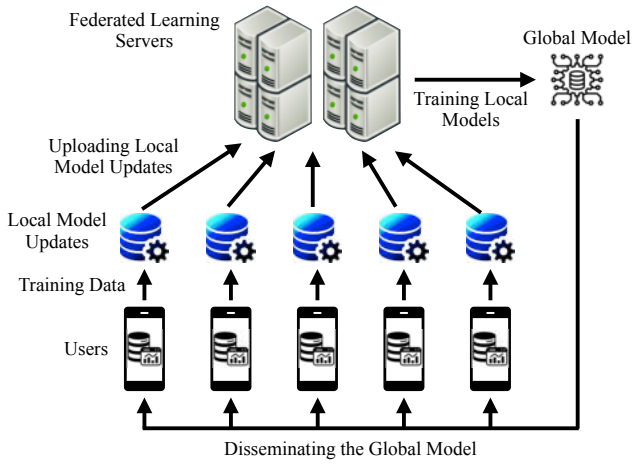

Fig. 1: Federated learning-enabled EdgeIoT, where IoT devices equipped with sensors and computational units process data locally. Model updates are generated at the IoT devices based on their local data, and uploaded to the server. A global model is sent back to the devices to update their neural networks' parameters.

embedded autoEncoder (E-GATE) at the manipulated devices to generate biasing local model updates by analyzing the characteristics of benign local models and global models. Specifically, the manipulated device mimics the behavior of a benign device, accessing the datasets for the training of benign devices and receiving the global models distributed by the server. E-GATE at the manipulated device specializes in interpreting intricate patterns and structures within explainable graph-based model updates. It excels at compressing graph data into a manageable, lower-dimensional space while preserving essential topological features. The manipulated device then reconstructs the explainable graph's structure to retain the local models' structural properties and maximize the federated learning biases. This altered explainable graph structure is used to create harmful local models that align with the benign models' data characteristics. As a result, these biasing local models can disrupt the global model's integrity while remaining consistent with benign models, making the detection of the E-GATE attack challenging.

This paper makes several important contributions:

- Introduction of a new fairness manipulation attack: The manipulated IoT devices conduct the proposed attack which creates biasing model updates based on explainable graphs. This attack aims to manipulate the federated learning fairness in EdgeIoT environments by altering correlations in benign local models while maintaining their original data characteristics.
- Exploration of an E-GATE attack framework: We examine a new E-GATE attack framework, which maximizes the minimum KL divergence of the participating IoT devices' model updates. Moreover, the E-GATE attack is trained to subtly modify explainable correlations within local models, ensuring the manipulation remains undetectable.
- Implementation and evaluation: The proposed E-GATE attack was tested on a Support Vector Machine (SVM) model with PyTorch and Python. Using CIFAR-10 datasets, numerical results show that E-GATE successfully increases the minimum KL divergence of benign model updates by 70.2%. The Cosine similarities between the biasing model updates and the corresponding global models are always below that of the benign local model updates. This underscores the attack's effectiveness in impairing federated learning fairness, as well as the invisibility of the manipulated IoT device.

The structure of this paper is as follows: Section II reviews existing research on adversarial attacks and defense mechanisms within EdgeIoT and federated learning. Section III discusses the system model for federated learning in EdgeIoT, including aspects such as device interactions and communication channels. In Section IV, we describe the design and methodology of our E-GATE attack. Section V details the performance evaluation of our approach. Finally, Section VI presents the conclusion and future research.

II. Related Work

This section reviews recent adversarial attacks in EdgeIoT and federated learning.

Unfair coordinating energy resources in smart grids can lead to non-compliance from energy resource owners a grid operator requests. For instance, the grid operator might ask these owners to reduce their solar power output to manage voltage rise issues, offering a tariff reduction in return. As discussed in [5], a false data injection attack was designed to create unfair energy resource coordination by manipulating the measurements sent from EdgeIoT devices to the grid operator. By altering the reported state of charge, the manipulated device undermines the fairness in energy resource coordination with misleading data. An analysis was conducted on eavesdropping and jamming attacks aimed at causing a secrecy outage in fairness-oriented subcarrier allocation within EdgeIoT [6]. In the scenario involving a manipulated device, the secrecy capacity and interception probability were derived based on the maximum achievable rate for legitimate communication.

The authors of [7] proposed a data-agnostic model poisoning attack, which requires no knowledge of training data. An adversarial graph autoencoder was designed to reduce the learning accuracy, where manipulated local models are generated by leveraging the training datasets and capturing the correlation patterns between benign local

and global models. In [8], a study was conducted on fair detection of poisoning attacks to prevent the model from either failing to converge or introducing biased classification outcomes. Their approach focuses on balancing anti-poisoning techniques, resulting in the creation of fairer and increasing inclusive federated learning models.

In [9], a model poisoning attack was developed in which a manipulated device can introduce or worsen algorithmic bias against specific groups of devices, while still preserving a reasonable level of model utility, such as classification accuracy. The attack assumes that a small portion of benign devices (referred to as manipulated devices) have been compromised, allowing the manipulated devices to manipulate the training process on these devices. By solving an optimization problem on a subset of local datasets, the attack adjusts the model parameters within the redundant space, negatively impacting the model's performance for the targeted group. A bias-driven approach was presented in [10] for conducting membership inference attacks on federated learning. A feature amplification technique was studied to capitalize on the rapid increase of the exponential function to magnify the distinctions between member and non-member data. A network representation was presented in [11] to enhance the graph autoencoder by preserving node features and network structure information. This method employed adversarial model learning to increase the mutual information sharing between node features and their representations during the encoding process.

III. System Model of Federated Learning-enabled EdgeIoT

In this section, we begin by outlining a federated learning, using image classification as an example. Additionally, we introduce a defense mechanism that can be deployed at the server to counter adversarial attacks.

We consider a federated learning involving N participants, where I are benign devices, and the remaining $(N-I)$ are legitimate but manipulated devices (or attackers). For each benign device $i \in [1, I]$, the amount of data available at the τ-th iteration is denoted by $D_i(\tau)$, and an input data sample from device i is represented as $s_i \in [1, D_i(\tau)]$. This applies for all $\tau \in [1, T_L]$, where T_L is the total number of iterations in the federated learning. The output produced by the machine learning model for the input s_i is indicated by $y(s_i)$.

At device i, the training loss function for federated learning, expressed as $L(\boldsymbol{\omega}_i(\tau); s_i, y(s_i))$, quantifies the approximation error between the input s_i and its corresponding output $y(s_i)$, where $\boldsymbol{\omega}_i(\tau)$ represents the local model of device i.

Given $D_i(\tau)$, the loss function of federated learning in the τ-th iteration is defined by

$$F(\boldsymbol{\omega}_i(\tau)) = \frac{1}{D_i(\tau)} \sum_{s_i=1}^{D_i(\tau)} L(\boldsymbol{\omega}_i(\tau); s_i, y(s_i)) + \beta \cdot f(\boldsymbol{\omega}_i(\tau)), \tag{1}$$

where $f(\cdot)$ is a regularizer function that represents the effect of the local training noise, and $\beta \in [0, 1]$ is a coefficient.

Additionally, the model update for device i at round $\tau+1$ can be defined as follows:

$$\boldsymbol{\omega}_i(\tau+1) \leftarrow \boldsymbol{\omega}_i(\tau) - \eta \nabla F(\boldsymbol{\omega}_i(\tau)), \tag{2}$$

where η is the learning rate assigned to the devices.

During each iteration τ, all devices send their updated models $\boldsymbol{\omega}_i(\tau)$, for every device i, to the server. The server then aggregates these updates to train the global model, denoted as $\boldsymbol{\omega}_G(\tau)$, for the τ-th iteration. The global model $\boldsymbol{\omega}_G(\tau)$ is subsequently broadcast to all devices, allowing them to continue training their local models for the next iteration $\boldsymbol{\omega}_i(\tau+1)$.

Measuring the Cosine similarity can be applied at the server as a defense mechanism to detect manipulated model updates [12]. The Cosine similarity calculates the angular similarity between every two devices' model updates, which is given by

$$\overline{\omega}_{i,i'} = \frac{\boldsymbol{\omega}_i(\tau) \cdot \boldsymbol{\omega}_{i'}(\tau)}{\|\boldsymbol{\omega}_i(\tau)\| \cdot \|\boldsymbol{\omega}_{i'}(\tau)\|}, \tag{3}$$

where i and i' indicate two different devices, $i, i' \in [1, N]$ and $i \neq i'$. $\|\cdot\|$ stands for cardinality of a vector.

IV. The Proposed Fairness Manipulation Attack

In this section, we present the threat model of the proposed fairness manipulation attack, as well as the learning architecture of the E-GATE.

A. Threat Model

Consider a scenario where $(N-I)$ manipulated devices, each with access to their own training data, participate in federated learning, alongside I benign devices, as illustrated in Fig. 1. These manipulated devices, who may pose as legitimate devices, aim to subtly undermine the fairness of the federated learning by generating and submitting manipulated local models during each communication round. Let τ represent the iteration index in the federated learning. Furthermore, it is assumed that the presence of manipulated model updates within the proposed E-GATE architecture remains unknown to the server and the benign devices throughout the training phase. Despite being unaware of any manipulated devices, the server is responsible for continuously monitoring and evaluating the local models

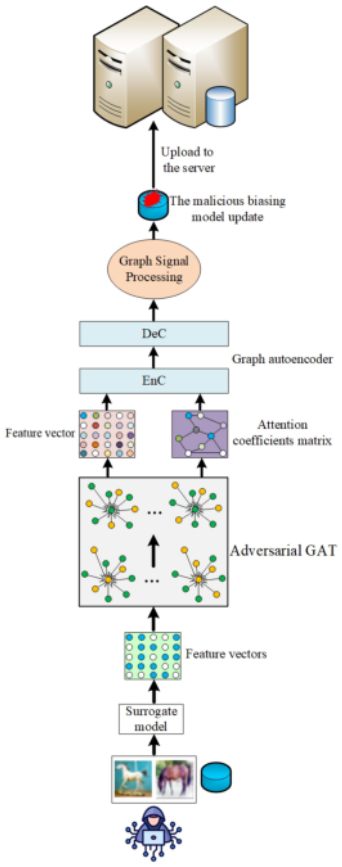

Fig. 2: The architecture of the proposed E-GATE attack, where a surrogate model is generated to extract labels of the manipulated device's biasing data. The adversarial GAT is trained to obtain the hidden representations of each feature in the explainable graph.

submitted by all devices to identify any manipulated or biased contributions.

Specifically, a manipulated device $j \in [1, N-I]$ generates a biased model update $\boldsymbol{\omega}_j^a(\tau)$ by exploiting the parameters of the benign local models observed during iteration τ. The server, unaware of the manipulated device's actions, aggregates the model updates from all devices, including both benign and manipulated ones. The total size of the training data reported to the server, denoted as $D_G(\tau)$, is computed as the sum of the data sizes from all benign devices, $D_i(\tau)$, along with the data size contributed by the manipulated device, $D_j^a(\tau)$. This aggregation leads to a compromised global model $\boldsymbol{\omega}_G^a(\tau)$ that generates

$$\boldsymbol{\omega}_G^a(\tau) = \sum_{i=1}^{I}\sum_{j=1}^{N-I}\frac{D_i(\tau)}{D_G(\tau)}\alpha_{i,j}^a(\tau)\boldsymbol{\omega}_i(\tau) + \sum_{j=1}^{N-I}\frac{D_j^a(\tau)}{D_G(\tau)}\boldsymbol{\omega}_j^a(\tau). \tag{4}$$

Specifically, $\alpha_{i,j}^a(\tau)$ is a binary variable indicating whether manipulated device j is able to intercept the model update $\boldsymbol{\omega}_i(\tau)$. This value, $\alpha_{i,j}^a(\tau)$, is known to the manipulated device. In other words, if the manipulated device j can extract $\boldsymbol{\omega}_i(\tau)$ based on the shared dataset for use in adversarial training to craft a biased model update, then $\alpha_{i,j}^a(\tau) = 1$. Otherwise, $\alpha_{i,j}^a(\tau) = 0$. The global model $\boldsymbol{\omega}_G^a(\tau)$, after being aggregated, is then broadcast by the server to all N devices.

The KL divergence between $\boldsymbol{\omega}_i(\tau)$ and $\boldsymbol{\omega}_G^a(\tau)$ can be formulated to measure the fairness of federated learning [13], which is given by

$$d_{\mathrm{KL}}(\boldsymbol{\omega}_i(\tau), \boldsymbol{\omega}_G^a(\tau)) = \sum_{\tau'=1}^{\tau} P(\boldsymbol{\omega}_i(\tau')) \log\Big(\frac{P(\boldsymbol{\omega}_i(\tau'))}{P(\boldsymbol{\omega}_G^a(\tau'))}\Big), \tag{5}$$

where $P(\cdot)$ is a probability density function, and $d_{\mathrm{KL}}(\cdot,\cdot)$ calculates the KL divergence between $\boldsymbol{\omega}_i(\tau)$ and $\boldsymbol{\omega}_G^a(\tau)$.

B. E-GATE Attack

Based on (4) and (5), the loss function with regard to the federated learning fairness is defined as

$$\delta_{\mathrm{Loss}} = \min_{i\in[1,N]} d_{\mathrm{KL}}(\boldsymbol{\omega}_i(\tau), \boldsymbol{\omega}_G^a(\tau)). \tag{6}$$

Fig. 2 presents the architecture of the proposed E-GATE attack, which aims to maximize δ_{Loss} in (6). A surrogate model $\tilde{g}(D_j^a(\tau))$ is used at the manipulated device to extract a set of biasing feature vectors corresponding to its adversarial training data. The feature vector representing $\tilde{g}(D_j^a(\tau))$, along with $\boldsymbol{\omega}_i(\tau)$, is expressed as $\boldsymbol{h} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, ..., \overrightarrow{h_e}\}$, where e refers to the size of the graph and h_e is constrained by the number of features associated with each vertex.

E-GATE is trained to obtain the hidden representations of each feature in the explainable graph. Based on the input of $\boldsymbol{h}$, E-GATE calculates attention coefficients for each of the vertices and features. Specifically, the encoder is constructed using an architecture based on the M layers of graph convolutional networks (GCN). This explainable design enables the encoder to learn a representation that effectively captures the essential characteristics of the model updates, which can be formulated as

$$\mathcal{Z}^M = f_G(\mathcal{Z}^{M-1}, \mathcal{A}|\mathbf{w}^M), \tag{7}$$

where $f_G(\cdot,\cdot|\cdot)$ represents a spectral convolution operation, while $\mathbf{w}^M$ signifies the weight matrix corresponding to the M-th layer within the GCN [14].

Given an identity matrix $I \in \mathbb{R}^{J\times J}$, $\widetilde{\mathcal{A}}$ can be formulated as $\widetilde{\mathcal{A}} = \mathcal{A} + I$, and we have $\overline{\mathcal{A}}_{xy} = \sum_{j'} \widetilde{\mathcal{A}}_{jj'}$. To generate a feature representation of the graph, the encoder can be written as

$$f_G(\mathcal{Z}^{M-1}, \mathcal{A}|\mathbf{w}^M) = \Phi^M(\overline{\mathcal{A}}^{-\frac{1}{2}}\widetilde{\mathcal{A}}\overline{\mathcal{A}}^{-\frac{1}{2}}\mathcal{Z}^{M-1}\mathbf{w}^M), \tag{8}$$

where $\Phi^M(\cdot)$ denotes a nonlinear activation function, for instance, $\tanh(\cdot)$ or ReLU$(\cdot)$; meanwhile, $\overline{\mathcal{A}}^{-\frac{1}{2}}\widetilde{\mathcal{A}}\overline{\mathcal{A}}^{-\frac{1}{2}}$ represents the symmetrically normalized adjacency matrix.

The output produced by the graph decoder is the reconstructed adjacency matrix, denoted as $\widehat{\mathcal{A}}$, which is given by

$$\widehat{\mathcal{A}} = \text{sigmoid}\left(\mathcal{Z}^M \left(\mathcal{Z}^M\right)^T\right), \tag{9}$$

where the sigmoid function is specified by $\text{sigmoid}(x) = 1/(1+\exp(-x))$. This formulation suggests that the likelihood of correlation between model updates within the graph increases with the magnitude of the inner product $(\mathcal{Z}^M(\mathcal{Z}^M)^T)$.

The discrepancy between $\mathcal{A}$ and its reconstructed counterpart $\widehat{\mathcal{A}}$ is quantified through a reconstruction loss function, which is defined as

$$\phi_{\text{loss}} = \mathbb{E}_{f_G(\mathcal{Z}^{M-1},\mathcal{A}|\mathbf{w}^M)}\Big[\log\ p(\ \widehat{\mathcal{A}}\ |\ \mathcal{Z}^M\ )\Big], \tag{10}$$

where the probability $p(\widehat{\mathcal{A}}\ |\ \mathcal{Z}^M)$, determined by the decoder, reflects the degree of correlation among the model updates.

V. Numerical Analysis

This section presents the KL divergence of the EdgeIoT devices based on the CIFAR-10 datasets. The detection efficacy of the E-GATE attack is examined through the metric of Cosine similarity between the local models and the global one, as depicted in (3).

For $I = 5$, Fig. 3 illustrates the KL divergence $d_{\text{KL}}(\boldsymbol{\omega}_i(\tau), \boldsymbol{\omega}_G^a(\tau))$ in (5) under the E-GATE attack. We carried out an evaluation using the CIFAR-10 dataset, focusing on two scenarios: one with a single attacker and another with five attackers participating in the federated learning. In each case, the figure illustrates the KL divergence for every device over 100 communication rounds. When five attackers are present, the KL divergence is approximately three times higher than in the scenario with just one attacker. This is expected, as a greater number of attackers results in more manipulated model updates, i.e., $\boldsymbol{\omega}_j^a(\tau)^*$, being incorporated into the federated learning. As a result, the maximum loss function related to federated learning fairness, as shown in (6), increases accordingly.

In particular, the five benign devices in both scenarios exhibit similar KL divergence values. This consistency is reasonable because the biased model updates generated by E-GATE are aggregated into the global model, which is used to update all devices. The incorporation of these biased updates into the global model affects all devices uniformly, leading to comparable KL divergence among the benign devices.

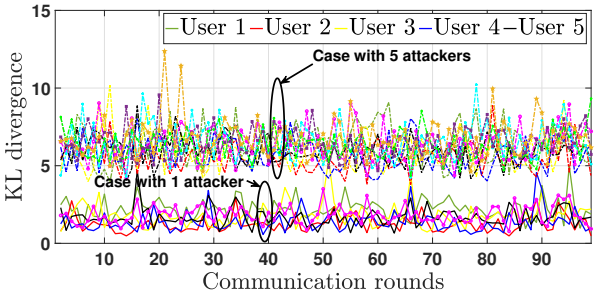

Fig. 3: When I = 5, the KL divergence $d_{\text{KL}}(\boldsymbol{\omega}_i(\tau), \boldsymbol{\omega}_G^a(\tau))$ under attacks with one or five attackers.

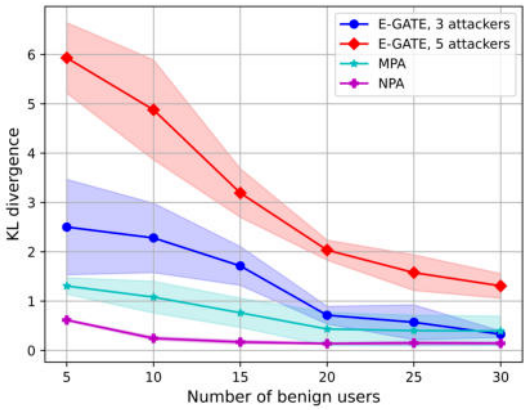

Fig. 4: When I increases from 5 to 30, the KL divergence $d_{\text{KL}}(\boldsymbol{\omega}_i(\tau), \boldsymbol{\omega}_G^a(\tau))$ given three or five attackers.

In Fig. 4, we perform a comparative assessment of the average KL divergence of the local models influenced by the proposed E-GATE attack versus those impacted by the existing model poisoning attack (MPA) and noise poisoning attack (NPA) methods [15], [16]. Specifically, MPA produces altered models with the goal of decreasing the accuracy of federated learning, whereas NPA creates models by adding Gaussian random noise to the global model received from the server. Moreover, the performance analysis examines scenarios with an increasing number of benign devices, I, ranging from 5 to 30, and varying numbers of attackers, from 1 to 5. Notably, when five attackers are involved, Fig. 4 demonstrates that the KL divergence under the E-GATE attack shows a significant rise, being 70.2% and 85.4% higher than the KL divergences observed under MPA and NPA, respectively.

To assess the stealthiness of the proposed E-GATE attack, we examine the Cosine similarity between local models and the global model, as defined in (3). Fig. 5 shows that the Cosine similarities between the manipulated model updates produced by the E-GATE attack and the global models consistently remain lower than those of benign

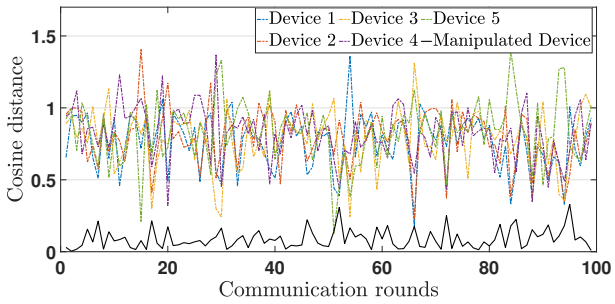

Fig. 5: Given 100 FL communication rounds and $I = 5$, the Cosine similarities of the local models are measured at the server in order to detect a fairness manipulation attack.

local updates. This is because E-GATE attack is trained to subtly modify explainable correlations within local models, ensuring that the manipulation remains undetectable. This makes it challenging for the server to detect and mitigate fairness biases, as a result, the manipulated models can be seamlessly merged with the benign ones.

VI. Conclusion and Future Research

This paper examines the impact of model fairness manipulation attacks on federated learning within EdgeIoT environments, where machine learning models are trained locally on individual devices and aggregated by a server to refine a global model. We propose a novel model fairness manipulation attack, i.e., E-GATE, which creates biasing model updates based on explainable graphs. E-GATE aims to maximize the minimum KL divergence between a device's local model update and the global model. E-GATE is skilled at detecting and interpreting structural correlations within the graph representations of these benign models, as well as the features of the data that support them. By reconstructing these explainable graph structures, E-GATE maximizes the reconstruction loss while keeping the biasing model updates undetected.

Future research into employing E-GATE for manipulating federated learning in EdgeIoT holds promise for advancing offensive and defensive strategies. E-GATE's explainable ability to model intricate relationships and dependencies in data positions it as a powerful tool for developing fairness manipulation attacks tailored to the unique topological structures of EdgeIoT. On the defensive front, there is growing interest in creating new models to detect E-GATE-based attacks by examining explainable graph properties for anomalies or signs of manipulation.

Acknowledgements

This Paper was supported by the CISTER Research Unit (UIDP/UIDB/04234/2020), financed by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology), and by project Aero.Next Portugal (ref. C645727867-00000066), funded by the EU/Next Generation, within call n.º 02/C05-i01/2022 of the Recovery and Resilience Plan (RRP).

References

[1] 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*, vol. 9, no. 21, pp. 21 099–21 110, 2022.

[2] L. Fotia, F. Delicato, and G. Fortino, "Trust in edge-based Internet of Things architectures: state of the art and research challenges," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–34, 2023.

[3] K. Li, Y. Cui, W. Li, T. Lv, X. Yuan, S. Li, W. Ni, M. Simsek, and F. Dressler, "When Internet of Things meets Metaverse: Convergence of physical and cyber worlds," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 4148–4173, 2022.

[4] K. Li, Z. Zhang, A. Pourkabirian, W. Ni, F. Dressler, and O. B. Akan, "Towards resilient federated learning in cyberedge networks: Recent advances and future trends," *arXiv preprint arXiv:2504.01240*, 2025.

[5] Y. Hu, X. Xian, Y. Jin, and S. Wang, "Fairness-guaranteed DER coordination under false data injection attacks," *IEEE Internet of Things Journal*, vol. 10, no. 21, pp. 19 043–19 053, 2023.

[6] B. Ahuja, D. Mishra, and R. Bose, "Fairness-aware subcarrier allocation to combat full duplex eavesdropping and jamming attacks in IoT," in *IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[7] K. Li, J. Zheng, X. Yuan, W. Ni, O. B. Akan, and H. V. Poor, "Data-agnostic model poisoning against federated learning: A graph autoencoder approach," *IEEE Transactions on Information Forensics and Security*, 2024.

[8] A. K. Singh, A. Blanco-Justicia, J. Domingo-Ferrer, D. Sánchez, and D. Rebollo-Monedero, "Fair detection of poisoning attacks in federated learning," in *Proceedings of IEEE international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2020, pp. 224–229.

[9] S. I. A. Meerza and J. Liu, "EAB-FL: Exacerbating algorithmic bias through model poisoning attacks in federated learning," *arXiv preprint arXiv:2410.02042*, 2024.

[10] L. Zhang, L. Li, X. Li, B. Cai, Y. Gao, R. Dou, and L. Chen, "Efficient membership inference attacks against federated learning via bias differences," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, 2023, pp. 222–235.

[11] D. Li, D. Li, and G. Lian, "Variational graph autoencoder with adversarial mutual information learning for network representation learning," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 3, pp. 1–18, 2023.

[12] J. Liu, X. Li, X. Liu, H. Zhang, Y. Miao, and R. H. Deng, "Defendfl: A privacy-preserving federated learning scheme against poisoning attacks," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[13] Z. Xie and S. Song, "FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 1227–1242, 2023.

[14] K. Li, X. Yuan, J. Zheng, W. Ni, F. Dressler, and A. Jamalipour, "Leverage variational graph representation for model poisoning on federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[15] S. Li, E. Ngai, and T. Voigt, "Byzantine-robust aggregation in federated learning empowered industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1165–1175, 2021.

[16] K. Li, J. Zheng, W. Ni, H. Huang, P. Liò, F. Dressler, and O. B. Akan, "Biasing federated learning with a new adversarial graph attention network," *IEEE Transactions on Mobile Computing*, 2024.