When Zero-Trust Meets Federated Learning

Xinran Zhang*, Dan Wang†, Yifei Zhu‡, Weilong Chen*, Zheng Chang*§ and Zhu Han¶‖
*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
†Department of Computing, The Hong Kong Polytechnic University, Hong Kong
‡UM-SJTU Joint Institute, Shanghai Jiao Tong University, Shanghai, China
§Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland
¶Department of Electrical and Computer Engineering, University of Houston, Houston, TX, USA
‖Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea

***Abstract*—Nowadays, Federated Learning (FL) has emerged as a promising and critical machine learning scheme to protect data privacy and reduce communication overhead. As the scale and connectivity expand in the FL system, enhancing the model's robustness against security threats from malicious clients grows ever more critical. An effective defensive solution involves selecting benign clients appropriately, thereby mitigating the vulnerability of the FL system to malicious attacks. However, clients exhibit varying behaviors over time, which complicates the task of accurately modeling their future trustworthiness. Moreover, blindly trusting clients with high trust values poses risks, given the potential for severe losses from betrayal. To tackle these problems, we propose a zero-trust policy in FL aimed at establishing continuous trust in each client while maintaining skepticism towards potential betrayal attacks. Specifically, we develop a Dirichlet-based trust evaluation technique to enable a comprehensive selection of trustworthy participants. This technique leverages the posterior distribution to estimate clients' trust values from their evolving behavior records over time. Then, we anticipate potential betrayal from a selected client and formulate a min-max optimization problem to minimize the worst-case betrayal loss, thereby boosting the system's betrayal-aware robustness. Next, we convert this problem into a convex optimization problem and utilize the interior point method for resolution. We conduct extensive simulations to validate the efficacy of our proposed zero-trust policy in accurately assessing trust and enhancing the model's robustness to betrayal.**

***Index Terms*—federated learning, client selection, zero-trust, trust evaluation, betrayal attack.**

I. Introduction

Federated Learning (FL) [1] is becoming a thriving research area in both academia and industry, primarily attributed to its advantage in protecting data privacy and boosting communication efficiency in machine learning tasks. However, as FL networks expand, their vulnerability to malicious client attacks increases, posing a significant risk to model robustness. To maintain cyber security, numerous studies have focused on developing defensive strategies. Specifically, [2] presents a framework that counters local model poisoning and maintains accuracy by using federated anomaly analytics to identify and assess potentially malicious local models. The work in [3] employs the beta distribution function to model the credibility of FL clients, scheduling trustworthy clients through their continuous trustworthiness. The selection of benign clients or models for FL involvement has proved effective, but the evaluation typically yields multiple levels rather than binary outcomes. Therefore, the adoption of a multi-valued satisfaction scale becomes essential for accurate trust evaluation [4].

As one of the Bayesian approaches, the Dirichlet distribution can be employed to estimate the probabilistic clients' behavior in the future, considering multi-valued satisfaction ratings [5]. Given the dynamic and complex nature of client behaviors, tracking uncertainty over time with the Dirichlet distribution is crucial for assessing their trustworthiness. To mitigate potential attacks from abnormal behaviors in mobile ad hoc networks, [6] uses the Dirichlet distribution for precise reputation modeling of nodes, evaluating trustworthiness by their performance and the peer-assigned reputation credibility. [7] employs the Dirichlet distribution within a blockchain-based system to automatically predict user reliability in an incentive scheme, enhancing the data quality. Although the Dirichlet-based trust management technique accounts for clients' evolving behavior, *the potential loss incurred by the betrayal of highly trustworthy clients is substantial.*

Zero trust can be regarded as an emerging security principle in FL to acknowledge and mitigate the aforementioned betrayal risks. The key idea of zero trust lies in abolishing implicit trust in any client and employing continuous risk-based trust verification [8], [9]. The zero-trust model consists of core components including trust evaluation and policy engine [10], which enables the system to build trust in clients while maintaining skepticism toward everyone. To this end, the system can proactively defend against malicious attacks and unexpected betrayals, protecting cyber security and robustness in heterogeneous and large-scale network systems. Due to these critical advantages, the zero-trust principle has been widely implemented in various network systems. For instance, to protect the satellite networks with increasing data exchange, ZTEI [11] employs continuous authentication and re-evaluation in a multi-dimensional monitoring process. Similarly, MUFAZA [12] aims to secure next-generation networks by employing a moving-horizon dynamic trust evaluation with multiple sources of evidence.

In this paper, we extend the zero-trust principle to FL, introducing a comprehensive zero-trust policy to tackle security threats posed by clients' dynamic behaviors and potential betrayal attacks. The proposed policy aims to boost the security and resilience within FL, establishing a more robust defense against potential attack and betrayal. We first

establish each client's trust for a client selection technique, and then emphasize the zero principle for a *min-max optimization to defend against the worst-case betrayal attacks*. Our main contributions are outlined as follows.

- We propose a zero-trust policy in FL to ensure the reliability and robustness of FL against the evolving risks of clients' attacks and betrayals.
- In particular, we apply the Dirichlet distribution model to evaluate clients' trustworthiness, facilitating an accurate selection technique by incorporating clients' behavior over time based on the multi-valued satisfaction analysis.
- Moreover, in adherence to the principle of maintaining zero trust in all participants, we implement a skepticism mechanism to guard against potential betrayals. To enhance model robustness, we formulate a min-max optimization problem to minimize losses from worst-case betrayals, thus protecting the FL system from unexpected security challenges.
- To validate our proposed zero-trust policy, we conduct extensive simulations. Simulation results show our policy outperforms baseline methods, regarding the effectiveness of trustworthiness and robustness against betrayals.

The structure of this paper is outlined as follows: Section II presents the system model, including the FL model, the Dirichlet-based trust evaluation model, and the threat model. Section III defines the zero-trust policy in FL, formulates a min-max optimization problem, and proposes solutions accordingly. Section IV presents the numerical simulation results. Finally, Section V concludes the paper.

II. System model

A. FL Model

We consider a FL system consisting of a set $\mathcal{C} = \{1, 2, \ldots, C\}$ of clients and a central server. Due to the limited resources and the presence of potentially malicious clients, only a selected subset $\mathcal{K} \in \mathcal{C}$ of clients can engage in the FL task, with their number of participants represented as $K (K < C)$. Each client $k \in \mathcal{K}$ collects its private dataset $\mathcal{D}_k$ with data size $D_k = |\mathcal{D}_k|$ and the size of all training data $\mathcal{D}$ is denoted as $D = \sum_{k\in\mathcal{K}} D_k$.

The primary objective of the FL task is to minimize the global loss function by identifying the optimal model parameter $\overline{\boldsymbol{w}}$ with a dimension of E:

$$\min_{\overline{\boldsymbol{w}}} F(\overline{\boldsymbol{w}}) = \sum_{k\in\mathcal{K}} p_k F_k(\overline{\boldsymbol{w}}), \tag{1}$$

where $\overline{\boldsymbol{w}} \in \mathbb{R}^E$ represents global model parameters, and $p_k = \frac{D_k}{D}$ represents the weight of client k, and we have $\sum_{k\in\mathcal{K}} p_k = 1$. Note that the global loss function is expressed as the weighted aggregation of individual local loss functions. The local loss function $F_k(\overline{\boldsymbol{w}})$ generated by client k can be calculated as the overall discrepancy between the predicted and actual results.

Let $\mathcal{M} = \{1, 2, \ldots, M\}$ denote the set of the global communication rounds and $\mathcal{I} = \{1, 2, \ldots, I\}$ denote the set of the local training rounds. The global model at global round m can be expressed as $\overline{\boldsymbol{w}}_m$ and the local model of client k in local round i at global round m can be represented as $\boldsymbol{w}^i_{k,m}$. The local stochastic gradient descent method is employed with the learning rate η. By aggregating all the local models $\boldsymbol{w}_{k,m}$ after I local rounds, we can obtain the global model

$$\overline{\boldsymbol{w}}_{m+1} = \sum_{k\in\mathcal{K}} p_k \boldsymbol{w}_{k,m}. \tag{2}$$

B. Dirichlet-based Trust Evaluation Model

To enhance the reliability of local updates, a trust-based technique can be utilized to identify and exclude malicious clients. The server selects clients for FL training based on the managed trustworthiness memory for all clients. Clients' future satisfaction levels are estimated by analyzing the distribution of satisfaction levels among client performances. To quantitatively evaluate client contributions, a satisfaction function $S(x)$, ranging from 0 to 1, is employed to quantify the degree of satisfaction with provided performances [13].

Let a discrete random variable X quantify the client satisfaction levels, drawn from a set $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ of N ($N \geq 2$) finite rival events [14]. Each $x_n \in [0, 1]$ defines the various levels of satisfaction and satisfies $x_{n+1} > x_n$. The weight τ_n assigned to each satisfaction level x_n increases as x_n increases, and the total sum of weights for all levels equals 1. Let the vector $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_n\}$ denote the probability distribution of X where $\sum_{n=1}^{N} \pi_n = 1$. The likelihood of X taking each possible satisfaction level x_n can be defined as

$$\mathrm{P}\{X = x_n\} = \pi_n, \tag{3}$$

where $\mathrm{P}[\cdot]$ is the probability function.

Let $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ denote the cumulative observations and initial beliefs of clients, serving as the foundation for understanding the probability distribution of X. To assign greater significance to recent observations compared to older ones, a forgetting factor γ ($\in [0, 1]$) is incorporated into the observation vector $\boldsymbol{\alpha}$ in the following manner

$$\boldsymbol{\alpha}^t = \gamma Y^{t-1} + Y^t = \sum_{i=1}^{t} \gamma^{t-i} \boldsymbol{Y^i} + b_0 \gamma^t Y^0, \tag{4}$$

where t represents the times of the observations. Y^t is the satisfaction level of the t-th observation, and it be expressed as a N dimensional tuple $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$, where a 1 in the n-th element indicates the satisfaction level is equivalent to x_n. The parameter $b_0 > 0$ reflects the prior belief, and Y^0 serves as the initial setting for the probability distribution.

By integrating prior knowledge with observations to quantify decision uncertainty, the Dirichlet distribution is leveraged to proactively update trust estimations for clients. It represents initial beliefs about uncertain events, evolving into a posterior distribution when merged with sample data. This facilitates dynamic trust updates based on interaction history, leading to the following definition.

Definition 1. *Suppose the satisfaction level x_n is obtained from N numbers, the probability for x_n is π_n, and the*

observation for x_n is α_n. The posterior Dirichlet distribution of P can be expressed as

$$f(\boldsymbol{\pi}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{n=1}^{N} {\pi_n}^{\alpha_n - 1}. \tag{5}$$

Let $\alpha_0 = \sum_{n=1}^{N} \alpha_n$ and $B(\boldsymbol{\alpha})$ can be expressed in terms of the gamma function as follows

$$B(\boldsymbol{\alpha}) = \frac{\prod_{n=1}^{N} \Gamma(\alpha_n)}{\Gamma(\alpha_0)}. \tag{6}$$

Then, the expectation of the probability π_n, considering the historical observations in $\boldsymbol{\alpha}$, can be expressed as follows [15]:

$$\mathbb{E}[\pi_n|\boldsymbol{\alpha}] = \frac{\alpha_n}{\alpha_0}. \tag{7}$$

Therefore, the mean of the posterior expected values of $\{\pi_1, \pi_2, \ldots, \pi_N\}$ are applied to represent the trustworthiness of client k, which can be expressed as:

$$V_k = \sum_{n=1}^{N} \tau_n \mathbb{E}[\pi_n|\boldsymbol{\alpha}] = \frac{\sum_{n=1}^{N} \tau_n \alpha_n}{\alpha_0}. \tag{8}$$

C. Threat Model

Although trustworthy clients can participate in FL training, there's a risk of them initially gaining trust and betraying the system later. To simplify the further analysis, we assume a single malicious client, who adds random noise from a Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_E)$ to the uploaded parameters as the attack approach [1]. The betrayer employs subtle noise attacks to degrade the model's accuracy while evading detection by the trust evaluation system.

To further derive the loss of betrayed attacks, we first introduce the definition of (ϵ, ζ)-Potential Attack Under Zero Trust, as follows.

Definition 2. *(ϵ, ζ)-Potential Attack Under Zero Trust states that when the betrayer applies the attack model Y, for all measurable output sets $\mathcal{R}$, and any pair of neighboring datasets X and X', the following inequality holds:*

$$\frac{\mathrm{P}[Y(\mathcal{X}) \in \mathcal{R}] - \zeta}{\mathrm{P}\left[Y\left(\mathcal{X}'\right) \in \mathcal{R}\right]} \leq e^{\epsilon}, \tag{9}$$

where $\epsilon > 0$ represents the attack impact, and $\zeta \in [0, 1]$ denotes the probability of failure.

A smaller ϵ results in a reduced distinguishability between neighboring datasets due to the betrayer introducing more intense noise. Inspired by [16], the Gaussian noise standard deviation σ_k can be represented as

$$\sigma_k = \frac{2\rho\sqrt{2M\log(1.25/\zeta)}}{p_k \epsilon_k D}, \tag{10}$$

where ρ is defined as the upper bound of the model parameters $\|w\|$, and $p_k D$ is the data size for client k.

[1]We assume a Gaussian noise attack in our threat model, but other attacks can also be employed given their similar goals and impacts on the FL system.

We consider the general assumptions of the non-convex loss function similar to other works [17] as follows:

Assumption 1. *$F_k(\boldsymbol{w})$ is L-smooth and lowered bounded for all $k \in \{1, 2, \ldots, K\}$. For all $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$: $F_k(\boldsymbol{w}_2) - F_k(\boldsymbol{w}_1) \leq (\boldsymbol{w}_2 - \boldsymbol{w}_1)^T \nabla F_k(\boldsymbol{w}_1) + \frac{L}{2} \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2^2$, where $F_k(\boldsymbol{w}) \geq \underline{F} \geq -\infty$.*

Assumption 2. *$F_k(\boldsymbol{w})$ is β-Lipschitz: for all $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$: $\|F_k(\boldsymbol{w}_2) - F_k(\boldsymbol{w}_1)\| \leq \beta \|\boldsymbol{w}_2 - \boldsymbol{w}_1\|$.*

To measure the impacts of the noise attack on the performance of FL, we provide an upper bound of $\mathbb{E}[F(\overline{\boldsymbol{w}}_T)] - F(\boldsymbol{w}^*)$ by deriving Theorem 1 as follows.

Theorem 1. *Let Assumptions 1 and 2 hold. Let $\Delta_0 = F(\overline{\boldsymbol{w}}_0) - F(\overline{\boldsymbol{w}}^*)$, where $F(\overline{\boldsymbol{w}}_0)$ and $F(\overline{\boldsymbol{w}}^*)$ represent the loss functions with the initial parameters $\boldsymbol{w}_0$ and the optimal parameter $\boldsymbol{w}^*$. When only client k betrays the system, we can derive the following upper bound for $\mathbb{E}[F(\overline{\boldsymbol{w}}_M)] - F(\boldsymbol{w}^*)$:*

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}[\|\nabla F(\overline{\boldsymbol{w}}_{m-1})\|^2] \leq \frac{\Delta_0 + A_1 p_k \sigma_k + A_2 (p_k \sigma_k)^2}{\eta M I (1 - L\eta I)}, \tag{11}$$

where $A_1 = \sqrt{\frac{2}{\pi}} \beta M E$ and $A_2 = LME^2$.

Proof. Since only one betrayal attacker k utilizes the noise attack with the standard deviation σ_k, and the local data size is up to the weight p_k, we can get the result by substituting (10) into [16]. □

Theorem 1 establishes a connection between the Gaussian noise attacks, denoted by σ_k, and the convergence of FL, quantified by $\mathbb{E}[F(\overline{\boldsymbol{w}}_T)] - F(\boldsymbol{w}^*)$. Stronger attacks with higher standard deviation σ increase the upper bound, consequently leading to decreased accuracy. Moreover, a higher weight p_k further degrades the FL convergence performance since betrayers can deteriorate the global aggregation more significantly. Therefore, assigning appropriate weights is crucial for maintaining the FL system's robustness by balancing the global trustworthiness and potential attack impacts.

III. Our Proposed Zero-Trust Policy in FL

In this section, we integrate the Dirichlet-based trust evaluation technique and the betrayal-aware defense mechanism into FL, and propose our zero-trust policy, intending to mitigate evolving security threats and worst-case betrayal attacks, as shown in Fig. 1. Firstly, we design a client selection mechanism to guarantee the performance of FL considering complex and dynamic clients' behavior. Then, we formulate a min-max optimization problem to model the maximal betrayal loss minimization. Next, we propose a solution to solve the problem effectively with the given algorithm.

A. Zero-Trust Policy

To filter potentially malicious clients, we set a threshold ψ based on the trust values V calculated in (8) for client

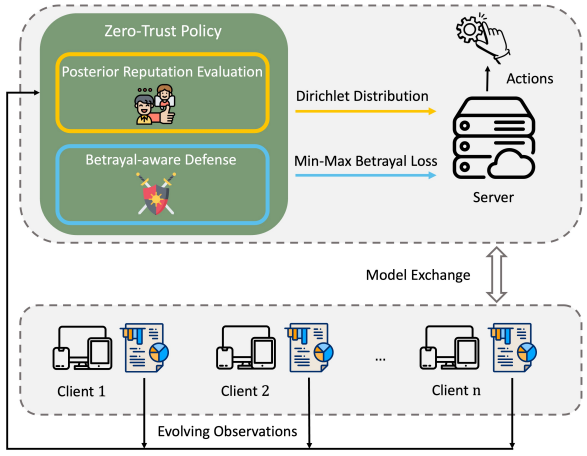

Fig. 1: The zero-trust policy including the posterior trust evaluation technique and betrayal-aware defense mechanism in FL.

selection. We schedule reliable users satisfying:

$$a_k = \begin{cases} 1, & V_k \geq \psi, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

where ψ is a threshold on the trustworthiness scores. If $a_k = 1$, client k is scheduled for FL training; otherwise, clients are not selected.

While the above client selection method effectively excludes highly untrustworthy clients, it overlooks the subtle difference in trustworthiness among the remaining clients. In the context of diverse trust levels, simply equalizing the weight of each client's contribution risks exposing the model to vulnerabilities. It is crucial to allocate different weights $p_k (k \in \mathcal{K})$ based on the trust scores of clients during model aggregation to enhance robustness. To this end, we introduce a new threshold θ to set a lower boundary for the accumulated impact of weighted trust evaluation, which can be written as

$$\sum_{k \in \mathcal{K}} a_k p_k V_k \geq \theta. \tag{13}$$

In addition, to fully embrace the zero-trust policy, we maintain suspicion toward all clients even those who are highly trustworthy. If highly trustworthy clients betray, the losses can be more severe than those from less trustworthy clients, posing a significant threat to the robustness of the system. By tackling the worst-case betrayal attack directly, we aim to enhance the overall security and effectiveness of our strategy. Therefore, we formulate a min-max optimization problem to acknowledge and address the potential severity of the most adverse betrayal in the following subsection.

B. Min-Max Problem Formulation

Here, we aim to minimize potential worst-case betrayal losses without compromising the integrated stability based on clients' trustworthiness. To achieve the zero-trust goal, the objective to be minimized can be simplified as $A_1 a_k p_k \sigma_k + A_2 (a_k p_k \sigma_k)^2$ from (11) by removing the constant values. Moreover, we restrict the lower bound of the sum of weighted trust values to ensure the FL performance. We formulate the following min-max optimization problem by optimizing $\boldsymbol{a} = [a_1, a_2, \ldots, a_K]$ to select clients, and the aggregation weights $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$ to allocate aggregation weights to each client:

$$\min_{\boldsymbol{a}, \boldsymbol{p}} \max_{k \in \mathcal{K}} \{A_1 a_k p_k \sigma_k + A_2 (a_k p_k \sigma_k)^2\}, \tag{14a}$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{K}} a_k p_k V_k \geq \theta, \tag{14b}$$

$$\sum_{k \in \mathcal{K}} a_k p_k = 1, \tag{14c}$$

$$p_k \in (0, 1], \forall k \in \mathcal{K}, \tag{14d}$$

$$a_k \in \{0, 1\}, \forall k \in \mathcal{K}, \tag{14e}$$

where the constraint in (14b) indicates that the accumulated trustworthiness should exceed a threshold θ, as explained in (13). The constraints (14c) and (14d) limit the feasible values of the optimization variable p_k. The constraint in (14e) restricts that the selection index is a binary choice. Hence, it is of great significance to strike a balance between global trustworthiness and potential betrayal attacks by optimizing the selection indices $\boldsymbol{a}$ and clients' weight $\boldsymbol{p}$.

C. Our Proposed Solution

It is challenging to solve the aforementioned problem directly due to the uncertainty of the maximal loss and the integer constraints on a_k. To provide an effective solution, we first introduce an auxiliary variable ξ. Then, we eliminate the parameter $\boldsymbol{a}$ by setting $p_k = 0$ when $a_k = 0$, and $p_k > 0$ when $a_k = 1$. Subsequently, the problem can be reformulated as follows

$$\min_{\xi, \boldsymbol{p}} \quad \xi, \tag{15a}$$

$$\text{s.t.} \ A_1 p_k \sigma_k + A_2 (p_k \sigma_k)^2 \leq \xi, \forall k \in \mathcal{K}, \tag{15b}$$

$$\sum_{k \in \mathcal{K}} p_k V_k \geq \theta, \tag{15c}$$

$$\sum_{k \in \mathcal{K}} p_k = 1, \tag{15d}$$

$$p_k \in [0, 1], \forall k \in \mathcal{K}. \tag{15e}$$

Since only the constraint in (15b) is convex and the objective and other constraints are linear, the problem (15) is a convex optimization problem. Thus, we apply the interior point method to solve it. The complete zero-trust process in FL is illustrated in Algorithm 1. The major complexity lies in solving (15), which involves complexity $\mathcal{O}(K^{3.5} \cdot \log(1/\delta))$ with accuracy δ by using the interior point method.

IV. Simulation Analysis

To visualize the probability density function (pdf) of the Dirichlet distribution, we consider $N = 3$ satisfaction levels: {dissatisfied, neutral, satisfied} across 10 rounds. Initially,

Algorithm 1: The zero-trust process in FL

$m \leftarrow 0$
while FL does not converge **do**
  Calculate V_k of clients using (8);
  Select clients satisfying (12);
  Broadcast the global model and assign weights p_k to the selected clients by solving (14);
  Clients utilize their $p_k D$ local data to train the model and upload it to the server;
  Aggregate the local models into a new global model by following (2);
  $m \leftarrow m + 1$;
end while

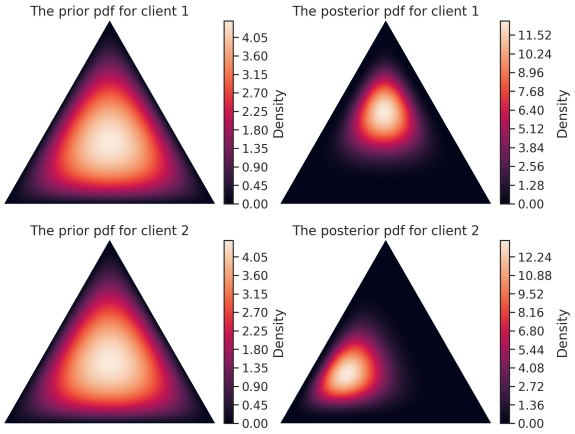

Fig. 2: The prior and posterior Dirichlet pdf for 2 clients.

with no prior information about the probability distribution, we assume a uniform prior distribution with $Y^0 = (2, 2, 2)$. The weights for the satisfaction levels are $\tau = \{0.01, 0.14, 0.85\}$, and the forgetting factor γ is set to 0.9.

Fig. 2 depicts the dynamic Dirichlet pdfs for two clients across two rows. The left column shows the prior Dirichlet pdf, which is the same for both users. In the right column, the posterior Dirichlet pdf is shown after accumulating 10 new observations, with the peak narrowing compared to the prior pdf. For Client 1, satisfaction level x_3 dominates across all rounds, resulting in a posterior pdf biased away from the base of the triangle, which corresponds to the π_3 axis. Conversely, Client 2 gathers more evaluations at satisfaction level x_1, biasing the posterior pdf away from the right side of the triangle, representing the π_1 axis.

The figure in Fig. 3 illustrates the changes of trust values over observation rounds, with three lines corresponding to different initial belief weights (b_0) of 10, 20, and 30. The first two curves represent users exhibiting continual benign behaviors, while the third curve demonstrates a user's shift towards malicious actions. Altering the initial belief weights (b_0) leads to varied convergence times for trust values. Notably, when b_0 is lower and benign behavior persists, the trust values rapidly stabilize. Conversely, a dissatisfied evaluation,

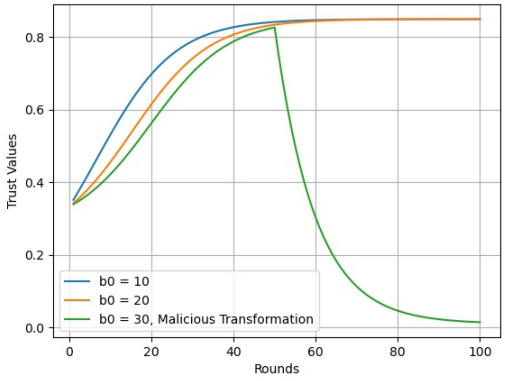

Fig. 3: The trust values under different settings.

depicted by the green curve, results in a swift decline in trust. Therefore, while subsequent satisfactory performances can alleviate the impact of this malicious transformation, rebuilding trust requires significantly more time.

To examine the performance of the zero-trust policy, we conduct a FL simulation involving 50 clients and 1 central server, employing the MNIST dataset and a 3-layer deep neural network. Each client conducts 2 local epochs, while the global training spans 100 epochs, utilizing a learning rate of 0.01 and the Adam optimizer. We introduce the noise attack from a Gaussian distribution to simulate malicious clients' behavior. To highlight the superiority of the zero-trust policy, we consider the following baselines for comparison:

- *Benign*: All clients are benign and no defense actions are taken in the system.
- *Betrayal*: One of the trustworthy clients turns malicious and attacks the model by adding subtle noise to the uploaded model, and no defense actions are taken.

The loss functions of the three methods over FL iterations are displayed in Fig. 4. We observe that all three methods converge, even in scenarios involving betrayal. This resilience can be attributed to the continuous trust evaluation that effectively filters out malicious clients. The loss is highest in the Betrayal scenario, indicating the substantial negative impacts of betrayal on the FL performance. This is because potential betrayers remain hidden within trustworthy user groups, and they may attempt to attack the model while strategically evading detection. Furthermore, our zero-trust policy results in lower losses compared to the Betrayal scenario, demonstrating its efficacy in mitigating losses from trustworthy client betrayals. This is attributed to consistent skepticism towards each participant and the allocation of optimized weights to minimize the worst-case betrayal loss, thereby ensuring betrayal-aware robustness. Moreover, the loss incurred from the zero-trust policy exceeds that in the Benign scenario, attributable to the presence of potential betrayers.

Fig. 5 shows the loss function values under three methods versus the attack intensity, represented by the Gaussian noise standard deviation σ. From the figure, it can be observed that

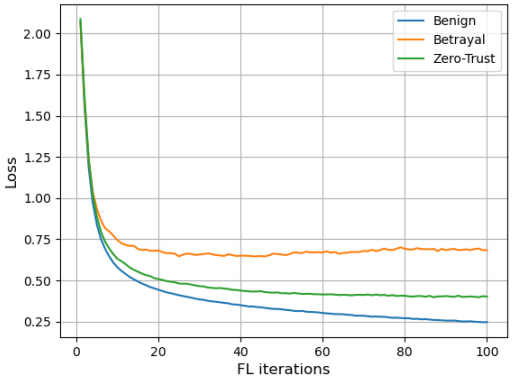

Fig. 4: The training loss with different methods.

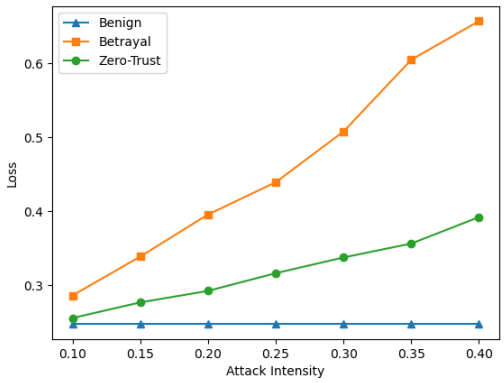

Fig. 5: The training loss under different attack intensities.

both Benign and Zero-trust policies experience an increase in loss as the attack intensity rises, indicating decreased model accuracy in the presence of a betrayal attack. This observation aligns with Theorem 1, where a higher attack intensity leads to increased interference by the attacker on the global model. We can also find that the loss remains constant in the Benign scenario, where no betrayal occurs.

V. Conclusion

In this paper, we emphasized the importance of integrating a multi-value trust evaluation technique and a betrayal-aware mechanism within a zero-trust-enabled FL system. We proposed a Dirichlet-based trust evaluation technique with a multi-value assessment over time, enabling the continuous selection of participants. Recognizing the risks of unconditional trust in highly reputed clients due to the potential for significant betrayal impacts, we adopted the zero-trust principle. Our zero-trust policy involved a consistent skepticism toward all clients, proactively guarding against the possibility of betrayal from even highly trustworthy participants. Following that, we introduced a min-max formulation aimed at mitigating the worst-case loss of betrayal. Through extensive simulations, we verified the effectiveness of our zero-trust policy, demonstrating its capacity to model accurate trust and significantly enhance the betrayal-aware robustness of the model.

VI. Acknowledgement

This work is partly supported by the National Natural Science Foundation of China under Grant 62071105, NSF CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, US Department of Transportation, Toyota, Amazon, JST ASPIRE JPMJAP2326, RGC GRF 15200321, 15201322, 15230624, RGC-CRF C5018-20G, ITC ITF-ITS/056/22MX, PolyU 1-CDKK and the National Key R&D Program of China under Grant 2023YFB2704400.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. Int. Conf. AI and Statist.*, Fort Lauderdale, FL, Apr. 2017, pp. 1273–1282.

[2] S. Shi, C. Hu, D. Wang, Y. Zhu, and Z. Han, "Federated Anomaly Analytics for Local Model Poisoning Attack," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 596–610, Feb. 2022.

[3] Z. Song, H. Sun, H. H. Yang, X. Wang, Y. Zhang and T. Q. S. Quek, "Reputation-Based Federated Learning for Secure Wireless Networks," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1212-1226, Jan. 2022.

[4] T. Nguyen, D. Hoang, and A. Seneviratne, "Dirichlet-Based Initial Trust Establishment for Personal Space IoT Systems," in *IEEE Int. Conf. Commun.*, Kansas City, MO, May 2018.

[5] C. J. Fung, J. Zhang, I. Aib, and R. Boutaba, "Dirichlet-Based Trust Management for Effective Collaborative Intrusion Detection Networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 8, no. 2, pp. 79–91, Jun. 2011.

[6] E. Chiejina, H. Xiao, B. Christianson, A. Mylonas, and C. Chiejina, "A Robust Dirichlet Reputation and Trust Evaluation of Nodes in Mobile Ad Hoc Networks," *Sensors*, vol. 22, no. 2, Art. no. 2, Jan. 2022.

[7] C. Zhang, M. Zhao, L. Zhu, W. Zhang, T. Wu, and J. Ni, "FRUIT: A Blockchain-Based Efficient and Privacy-Preserving Quality-Aware Incentive Scheme," *IEEE J. Select. Areas Commun.*, vol. 40, no. 12, pp. 3343–3357, Dec. 2022.

[8] Y. Ge, T. Li, and Q. Zhu, "Scenario-Agnostic Zero-Trust Defense with Explainable Threshold Policy: A Meta-Learning Approach," in *IEEE Conf. Comput. Commun. Workshops*, Hoboken, NJ, USA, May 2023, pp. 1–6.

[9] A. Wylde, "Zero trust: Never trust, always verify," in *Int. Conf. Cyber Situational Awareness Data Anal. Assess.*, Virtual, Jun. 2021, pp. 1–4.

[10] S. Rose, O. Borchert, S. Mitchell, S. Connelly, "Zero trust architecture," *NIST Special Publication*, 800:207, Aug. 2020.

[11] P. Fu, J. Wu, X. Lin, and A. Shen, "ZTEI: Zero-Trust and Edge Intelligence Empowered Continuous Authentication for Satellite Networks," in *Proc. IEEE Glob. Commun. Conf.*, Rio de Janeiro, Brazil, Dec. 2022, pp. 2376–2381.

[12] Y. Ge and Q. Zhu, "MUFAZA: Multi-Source Fast and Autonomous Zero-Trust Authentication for 5G Networks," in *IEEE Mil. Commun. Conf.*, National Capital Region, USA, Nov. 2022, pp. 571–576.

[13] C. J. Fung, J. Zhang, I. Aib, and R. Boutaba, "Dirichlet-Based Trust Management for Effective Collaborative Intrusion Detection Networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 8, no. 2, pp. 79–91, Jun. 2011.

[14] N. T. Nguyen, G. Zheng, Z. Han, and R. Zheng, "Device fingerprinting to enhance wireless security using nonparametric Bayesian method," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 1404–1412.

[15] Kotz, Samuel, Narayanaswamy Balakrishnan, and Norman L. Johnson, *Continuous multivariate distributions, Volume 1: Models and applications*, vol. 1. John Wiley and Sons, 2004.

[16] T. Liu, B. Di, P. An, and L. Song, "Privacy-Preserving Incentive Mechanism Design for Federated Cloud-Edge Learning," IEEE Trans. Netw. Sci. Eng., vol. 8, no. 3, pp. 2588–2600, Jul. 2021.

[17] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *Proc. Int. Conf. Learn. Representations*, May 2020, pp. 1-26.