Zero-Knowledge-Proof for Moral Hazard Detection in O-RAN without Benchmarks: Let us Play WereWolf Game!

Damla Sarıçelik[†‡], Mohaned Chraiti[*‡], Albert Levi [†], and Ozgur Ercetin[*]
[†]Computer Science and Engineering Department, Sabancı University, Istanbul, Turkey
[*]Electronics Engineering Department, Sabancı University, Istanbul, Turkey
[‡]DGTL X, Kocaeli, Turkey
Emails: damlasaricelik@sabanciuniv.edu, mohaned.chariti@sabanciuniv.edu,
levi@sabanciuniv.edu, and oercetin@sabanciuniv.edu.

***Abstract*—The Open Radio Access Network (O-RAN) paradigm fosters multi-vendor interoperability, allowing modules from different vendors to cooperatively handle network functions, such as temporary data processing or sensor data collection for network operations optimization. However, this integration agility introduces the risk of selecting suboptimal or adversarial modules, leading to moral hazard. Traditional Moral Hazard testing approaches typically rely on a benchmarking data set in addition to historical performance score. However, they deemed impractical, as vendor-supplied modules may not reveal their outputs before deployment, and the network may lack direct access to reference results for validation. This challenge is further compounded by the dynamic nature of network elements and AI-driven models, whose performance can degrade over time due to malicious tampering, obsolescence, or device deterioration, making historical quality assessments ineffective. In this paper, we address the challenge of identifying legitimate vendor-supplied modules among adversarial ones, with respect to a given network functionality/operation, in the absence of benchmarks. We propose a benchmark-free test framework that detects and eliminates adversarial modules using a methodology inspired by the WereWolf game, combined with zero-knowledge proof techniques. Monte Carlo simulations demonstrate that our approach effectively removes adversarial entities while preserving the privacy of legitimate modules.**

***Index Terms*—O-RAN, Zero-Knowledge Proofs, Zero-Trust Architecture, Homomorphic Encryption, Moral Hazard.**

I. Introduction

The integration of modules supplied by different vendors within the same Radio Access Network (RAN) was historically subject of controversial opinion until the emergence of the Open RAN (O-RAN) concept [1]–[3]. Recently, O-RAN has gained recognition within the 3GPP standardization community and is now officially part of the standard [2], [4]. The primary goal of O-RAN is to foster a competitive and dynamic supplier ecosystem by expanding the pool of available modules, thereby enhancing overall network performance, flexibility, and the seamless integration of new functionalities over time [5], [6]. For instance, O-RAN allows service operators to temporarily utilize third-party machine learning models for data processing, optimizing network operations.

In the event where a module can be supplied by multiple vendors, selection and trust become critical challenges. Integrating an adversarial module, temporarily or permanently, can compromise end-to-end network performance, raising issues on fault attribution and conflict resolution. While authentication mechanisms can serve as the first line of defense in verifying the legitimacy of O-RAN modules, they are they are not designed to provide protection against moral hazard. Specifically, an O-RAN module may fail to deliver the expected output, either intentionally (actively malicious) or unintentionally due to outdated models or defective hardware (non-compliant). Furthermore, trust concerns extend to third-party providers offering services to O-RAN. In fact, the RAN Intelligent Controller (RIC) can be honest-but curious and could execute the protocol faithfully but potentially exploiting data. On the other hand, vendor-supplied modules operate on the basis of incentives. Thus, they may be reluctant to share their output during the testing phase due to the risk that the RIC could utilize their data without proper compensation or adherence to contractual agreements. Several real-world scenarios highlight these challenges. For example, multiple AI models hosted in the cloud may offer data processing services, or O-RAN may rely on distributed sensors to provide side information for network optimization. In such cases, *how can the network reliably select trustworthy modules among many options?* Additionally, *how can O-RAN modules show their compliance without disclosing vendor-sensitive data and risking exploitation?*

The O-RAN Alliance White Papers [7], [8] underline that the move toward O-RAN calls for a change from static trust models to a Zero Trust Architecture (ZTA), whereby no module is intrinsically trusted and continuous verification is needed. In the context of O-RAN, trust is assessed from two perspectives: security and performance compliance. While the former has been thoroughly investigated, the latter is still emerging, with limited research available. In fact, numerous

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 101108094. This work is supported by Tübitak under grant 122E497.

studies have implemented ZTA to automate the continuous detection of security threats using the RAN Intelligent Controller (RIC). For instance, the authors in [9] proposed a game-theoretic ZTA for automated detection of intelligent jammers in 6G RAN. Liu et al. [10] introduced multi-device anonymous authentication within O-RAN and proposed a ZTA-based local and roaming identity authentication protocol.

While moral hazard detection in O-RAN module selection remains an unexplored subject, techniques from contiguous fields offer potential solutions. However, benchmarking is the main drawback of these techniques. Subjective logic, for instance, has been widely applied for uncertainty quantification and information source selection in areas such as vehicular networks, federated learning, mobile ad hoc networks, and IoT [11]–[18]. However, access to predefined performance criteria and reference datasets is key in subjective logic to evaluating the performance or compliance of a module, model, or device.

Benchmarking has several limitations in the O-RAN environment. First, it requires predefined performance criteria and reference datasets, which are often unavailable in AI-driven systems where inputs and operating conditions change dynamically. Second, adversarial O-RAN modules can be trained to pass a given set of benchmarks, yet fail when different scenarios emerge, making benchmarking vulnerable to gaming and evasion. Third, benchmarking assumes that the O-RAN module discloses vendor-sensitive data during the test, which exposes them to potential exploitation. Assuming a centralized and trustworthy authority is a challenging premise that contradicts the ZTN principle. Testing and filtering compliance among a set of modules without compromising vendor sensitivity or relying on benchmarking data are deemed to be obsolete as network conditions evolve and functionality expands, making compliance testing and selection a significant challenge in O-RAN.

In this paper, we consider the case of O-RAN, where a set of modules, from multi-vendors, assert their ability to perform a specific functionality. The objective is to develop a compliance testing framework to identify and eliminate adversarial modules while retaining legitimate ones, without relying on benchmarking datasets or revealing potentially exploitable outputs. We assume that both the O-RAN evaluator and the modules under test are untrusted at the outset. The proposed framework is inspired by the WereWolf game mechanism to eliminate adversarial modules, and it leverages the principle of Zero-Knowledge Proof (ZKP) to ensure that tested modules do not directly expose their data. It is important to note that, in the absence of the proposed WereWolf game based framework, Zero-Knowledge Proofs are typically used to verify the truth of a specific statement without disclosing any information beyond the verification itself. While ZKPs are generally designed to validate information subject to a benchmark, they are not directly applicable for performance evaluation. To the best of our knowledge, this framework is the first to provide a zero-knowledge-proof and benchmark-free performance testing in O-RAN without exposing potentially exploitable outputs. Specifically, our contribution is as follows.

- Inspired by the WereWolf game, we propose a framework that iteratively eliminates non-compliant nodes via probabilistic trust validation, ensuring a self-regulating, benchmark-free compliance model.
- We prove that data are not exposed through the processes of compliance testing and elimination of adversarial modules.
- Through Monte Carlo simulations, we demonstrate the efficiency, scalability, and adversarial resistance of our framework.

The paper is structured as follows: Sec. II presents the system model, while Sec. III introduces the proposed approach. Sec. IV discusses the simulation results, and Sec. V concludes the paper.

II. System Model and Case Studies

A. System Model

We consider an O-RAN environment in which the RIC aims to test a set of N modules from different vendors for compliance with respect to a given functionality. The RIC is considered to be honest-but-curious, i.e., executing the protocol faithfully but potentially exploiting data. The objective is to identify and eliminate adversarial modules while preserving privacy and avoiding reliance on predefined benchmarks. As illustrated in Fig. 1, different vendors can offer a solution to cover some operations in the O-RAN Distributed Unit (O-DU) and O-RAN Centralized Unit (O-CU). The O-RAN modules are consistently treated as untrusted, i.e., under test. Accordingly, the RIC, specifically the Quality-of-Service Management, is responsible for assessing trust by managing the test execution.

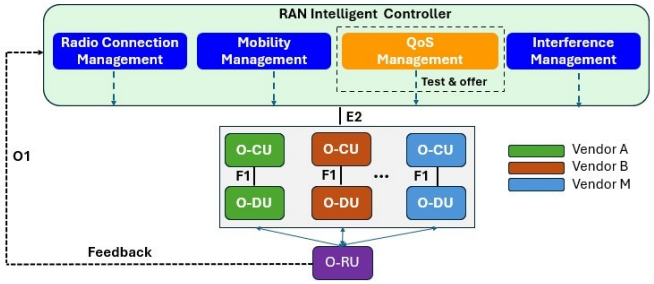

Fig. 1. System architecture of the proposed compliance verification framework in O-RAN, highlighting clients, the central unit, and adversarial modules.

-0.3cm

Let $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$ denote the set of O-RAN modules under evaluation. Among these, a subset $\mathcal{C}_{\text{legit}} \subseteq \mathcal{C}$ of size M consists of legitimate modules, while the remaining $N-M$ modules are adversarial, attempting to pass compliance checks without genuinely meeting the functional requirements.

The RIC conducts L test rounds, where an input $x \in \mathcal{X}$ is randomly selected and sent to the modules for processing. Each module $C_i \in \mathcal{C}$ processes x and produces an output $y_i \in \mathcal{Y}$. The output space $\mathcal{Y}$ is large and discrete, ensuring that random guessing has a negligible probability of yielding the correct output.

We assume that the RIC is not in possession of a labeled data set that can be used for the benchmarking. Moreover, throughout the compliance verification process, neither the RIC nor the modules inherently trust each other. The RIC could exploit legitimate modules by using their outputs alongside inputs to train models or process real data under the guise of testing, introducing a moral hazard. To mitigate this, the proposed framework embeds security measures that constrain the RIC's role to merely facilitating the proper execution of an interactive algorithm for adversarial module elimination.

B. Case studies

Multiple case scenarios in O-RAN require rigorous testing the option provided by multiple vendors, particularly for O-DU functionalities, to ensure compliance, interoperability, and performance consistency across vendors. One example scenario involves validating source and channel coding implementations, where multiple O-DUs from different vendors must comply with 3GPP standards. Ensuring that vendor-specific implementations do not introduce inefficiencies or incompatibilities is essential for seamless network operation.

Another scenario pertains to packet segmentation and reassembly, where O-DUs must correctly process transport blocks according to fronthaul requirements. Variability in vendor implementations can impact latency, packet integrity, and synchronization, necessitating thorough evaluation to ensure alignment with real-time constraints and overall system stability.

A third scenario involves the temporary use of AI models in the cloud for data processing, such as feature extraction, data classification, or compression within the O-DU. Vendor solutions need to be tested before deployment to ensure they do not introduce biased decision making or instability in dynamic network scenarios, thereby maintaining the reliability and efficiency of AI-driven O-DU functionalities.

C. Limitations of intuitive compliance approach

While there is no existing solution pertain testing without benchmarking in O-RAN, this section discusses some intuitive approaches and their limitations. The first approach is random selection, which, in the absence of a benchmarking dataset, may seem intuitive, but is highly inefficient. In the second scenario, multi-vendor modules share their outputs directly. Given the large output space, two matching outputs may be assumed to be correct. However, exposing data poses the risk of exploitation by illegitimate RICs, which could use the service without proper incentives, violating the principles of ZTA.

III. Proposed Approach: WereWolf Game for Adversarial Elimination

A. Backgorund in WereWolf Game

The proposed approach is inspired by the WereWolf game, which is a social deduction party game where players are assigned secret roles as either villagers or werewolves. The game revolves around the conflict between the villagers, who must identify and eliminate the werewolves, and the werewolves, who work covertly to eliminate the villagers without revealing their identities. One key dynamic is that the werewolves know each other's identities, allowing them to strategize and focusing on eliminating villagers excursively at each time that get a chance. In contrast, the villagers are at a disadvantage as they do not have any direct information about who the werewolves are. This creates an inherent asymmetry where the werewolves have a higher probability of winning. They avoid eliminating each other and instead focus on weakening the village team, while the villagers are left to eliminate players randomly without knowledge of their roles, making it difficult to make informed decisions.

The game continues until one side achieves its goal: either the villagers successfully identify and eliminate all the werewolves, or the werewolves eliminate enough villagers to outnumber them, securing a victory.

B. Proposed approach

The proposed approach builds upon a modified version of the WereWolf game, where the primary key success is that the werewolves know each other's identities. In the context of our framework, the "werewolves" represent the legitimate modules, while the "villagers" refer to the non-legitimate modules. A critical element of this design is ensuring that the legitimate devices, similar to the werewolves, are aware of each other's identities without directly disclosing their output data y_i. To achieve this, we introduce a privacy-preserving mechanism based on zero-knowledge proofs, whereby each module employs homomorphic encryption to protect its data. These encrypted values are then shared with the RIC, which are then broadcasted to all modules.

As described here within in more details, homomorphic encryption enables the modules to compare their results without decrypting the data, thereby maintaining privacy while allowing for identification of matching outputs. Moreover, it guarantees that neither the non-legitimate modules nor the RIC have knowledge of the modules with matching outputs. Following this, the process of elimination begins. In each round, the RIC randomly selects a surviving module and grants it the authority to eliminate another module. When a legitimate module is selected, it eliminates a non-legitimate module that produces a different output. Conversely, when a non-legitimate module is chosen, it eliminates a module at random, as all other modules appear to yield different results. The homomorphic encryption ensures that the identity of legitimate modules (the "werewolves") remains concealed, as the encrypted data prevent the non-legitimate devices from identifying those with matching outputs.

In subsequent sections, we provide a detailed explanation of the key elements and operations underpinning the proposed approach.

1) Privacy Preserving: To ensure privacy, each O-RAN module encrypts its data using randomized homomorphic encryption, where encrypting the same value multiple times or by multiple devices yields different ciphertexts. This prevents

adversaries from correlating encrypted outputs. Homomorphic encryption enables secure computation and comparison of encrypted data without decryption [19].

We use the Paillier cryptosystem. In particular, we assume that all modules share the same Paillier key pair, which is not shared with the RIC. The encryption of x_i is given by:

$$E(x_i) = g^{x_i} r_i^N \mod N^2 \tag{1}$$

Here, g is an element of the multiplicative group $\mathbb{Z}_{N^2}^*$ and must be chosen such that $g^N \mod N^2 \neq 1$. In many implementations, g is conveniently set to $N+1$. N is defined as the product of two large primes, p and q (i.e., $N = p \cdot q$). The security of the scheme fundamentally relies on the computational difficulty of factoring N. The random value r_i is chosen from $\mathbb{Z}_N^*$, meaning it must be coprime to N. The randomization introduced by r_i ensures that ciphertexts remain indistinguishable, even for identical plaintexts.

The comparison occurs at the RIC. Since the RIC does not have access to the secret key, it can only perform comparisons over encrypted values without decryption. We leverage Paillier's homomorphic subtraction property:

$$E(x_1 - x_2) = E(x_1) \cdot E(x_2)^{-1} \mod N^2 \tag{2}$$

The RIC raises the homorphocally encrypted difference to a random exponent $\alpha_{1,2}$, yielding $E(\alpha_{1,2}(x_1 - x_2))$. Even if a module has access to the secret key and one of the plaintexts (e.g., x_1), it can only infer $\alpha_{1,2}x_2$, which is random. Meanwhile, if $x_1 = x_2$, decryption yields $\alpha_{1,2}(x_1 - x_2) = 0$. Thus, similarity verification can be performed without exposing the data.

The security of this scheme is based on the Decisional Composite Residuosity Problem, which states that given $y \in \mathbb{Z}_{N^2}^*$, it is computationally infeasible to determine whether there exists an integer x such that:

$$x^N \equiv y \mod N^2. \tag{3}$$

This hardness assumption, similar to integer factorization, ensures the encryption remains secure against chosen-plaintext attacks.

C. Protocol Design and Communication Exchange

The test consists of L rounds. In each round l, each network module i receives a random input x from the RIC and computes its encrypted output as $E(y_i)$, where $y_i = f(x)$ is the module's response to x. The encrypted output $E(y_i)$ is then transmitted to the RIC. Recall that all modules use the same Paillier key pair.

Upon receiving the encrypted outputs, the RIC assigns a random alias $a_i^{(l)}$ to each module, ensuring that these aliases change in every testing round. Since the RIC does not possess the secret key, it can only compute encrypted differences by leveraging the homomorphic property. Furthermore, it scales each encrypted difference by a random exponent $\alpha_{i,j}$, giving:

$$E(\alpha_{i,j}(y_i - y_j)), \quad \forall i, j \in \{1, \ldots, N\},\ i \neq j$$

The RIC then sends to node C_i the set of alias-output pairs:

$$\mathcal{S}_i = \{(a_j^{(l)}, E(\alpha_{i,j}(y_i - y_j))) \mid j \neq i,\ j = 1, \ldots, N\}$$

Each node decrypts the received values. The result equals zero when the compared outputs are identical, and a random nonzero value otherwise. In this way, legitimate modules can identify the aliases corresponding to peers with matching outputs.

At each iteration t of the game, the RIC designates a randomly selected surviving module j to choose an alias for elimination. The decision process follows the structure of the WereWolf game:

- If C_j is a legitimate module, it knows the aliases of other legitimate modules and exclusively targets non-legitimate modules for elimination.
- If C_j is a non-legitimate module, it lacks any similarity with other outputs and cannot infer the identities of the legitimate group. Consequently, it selects an alias uniformly at random for elimination.

The game proceeds iteratively until only legitimate modules remain, or a single non-legitimate module is left. If all adversarial modules are eliminated, the selected module by the RIC announces the game's conclusion, declaring all remaining participants as legitimate. However, this declaration must be verified through consensus among the surviving modules, ensuring that no adversarial device prematurely terminates the game. At the conclusion of each game:

- If a non-legitimate module remains, it receives a score of 1, while all other modules receive a score of 0.
- If legitimate modules successfully eliminate all adversarial modules, the last designated legitimate module must disclose the aliases of all remaining legitimate devices. This claim is validated through confirmation requests from the group, preventing adversarial modules from falsely concluding the game early.

In the case where legitimate modules mark a win, each of them receives a score of one, including those that were eliminated during gameplay. Otherwise, a surviving adversarial module receives a score of one, while the remaining modules receive a score of zero. After L tests, the normalized score accumulated by a module y_i, denoted as R_i, corresponds to the total score accumulated over all games. As the number of tests increases,

$$R_i \underset{L\to\infty}{\to} P_{\text{win}}, \quad \forall C_i \in \mathcal{C}_{\text{legit}}$$

and

$$R_i \underset{L\to\infty}{\to} \frac{1}{N-M}(1 - P_{\text{win}}), \quad \forall C_i \notin \mathcal{C}_{\text{legit}}.$$

After L rounds, modules with the highest R_i values are considered legitimate.

D. Game success probability

At iteration t, the remaining modules are $N - t$, with M_t legitimate and $A_t = N - t - M_t$ adversarial. Define $P_{\text{win}}(M_t, A_t)$ as the probability that the legitimate group wins

given that M_t legitimate and A_t adversarial modules remain. At each round, a module is chosen uniformly at random for elimination. The probability of selecting a legitimate module is:

$$P_L(t) = \frac{M_t}{N-t}. \tag{4}$$

The probability of selecting an adversarial module is:

$$P_A(t) = \frac{A_t}{N-t}. \tag{5}$$

If a legitimate module is chosen, it eliminate an adversarial module, it is removed, reducing A_t by 1. The transition is:

$$(M_t, A_t) \to (M_t, A_t - 1). \tag{6}$$

If an adversarial module is chosen, it must eliminate another module. Since there are $N - t - 1$ remaining modules, it eliminates a legitimate module with probability:

$$P_{L|A}(t) = \frac{M_t}{N-t-1}, \tag{7}$$

or an adversarial module with probability:

$$P_{A|A}(t) = \frac{A_t - 1}{N-t-1}. \tag{8}$$

Using these probabilities, the probability that the legitimate group wins satisfies the recurrence:

$$\begin{aligned} P_{\text{win}}(M_t, A_t) &= P_L(t) P_{\text{win}}(M_t, A_t - 1) \\ &+ P_A(t) \left[P_{L|A}(t) P_{\text{win}}(M_t - 1, A_t) \right. \\ &\left. + P_{A|A}(t) P_{\text{win}}(M_t, A_t - 1) \right]. \end{aligned} \tag{9}$$

Substituting the transition probabilities:

$$\begin{aligned} P_{\text{win}}(M_t, A_t) &= \frac{M_t}{N-t} P_{\text{win}}(M_t, A_t - 1) \\ &+ \frac{A_t}{N-t} \left(\frac{M_t}{N-t-1} P_{\text{win}}(M_t - 1, A_t) \right. \\ &\left. + \frac{A_t - 1}{N-t-1} P_{\text{win}}(M_t, A_t - 1) \right). \end{aligned} \tag{10}$$

Since the game ends when $A_t = 0$, we can treat the state $(M_t, 0)$ as an absorbing state, i.e., a state once entered, the system cannot transition out of it. They can be defined as:

$$P_{\text{win}}(M_t, 0) = 1 \quad \text{if} \quad M_t > 0 \quad \text{(i.e., the legitimates won)},$$

and

$$P_{\text{win}}(0, A_t) = 0 \quad \text{if} \quad M_t = 0 \quad \text{(i.e., the adversarial won)}.$$

By utilizing the transition states and absorbing states, a Markov chain can be constructed. Accordingly, one can calculate the steady-state probabilities or the probability of reaching an absorbing state (win or lose) starting from an initial state $(M, N - M)$. Alternative approaches include using dynamic programming or Monte Carlo simulations to analyze the system. Given the limitations on paper length, the probability is empirically estimated using Monte Carlo simulations instead of being derived through theoretical analysis.

Considering the normalized scoring system described at the end of Sec. III-C, legitimate devices are selected when

$$\max_{C_i \in \mathcal{C}_{\text{legit}}} R_i \geq \max_{C_i \notin \mathcal{C}_{\text{legit}}} R_i.$$

For sufficiently large L, legitimacy is ensured when

$$P_{\text{win}} \geq \frac{1}{N-M}(1 - P_{\text{win}}).$$

IV. Simulation Results

The experimental setup consists of a total of $N = 10$ modules under test, with the number of malicious devices $N - M$ varying between 0 and 10. We consider $L = 100$ test iterations. Recall that the used the randomized homomorphic encryption and the designed protocol ensure that legitimate modules identify each other with 100% certainty at the begining of each game. Moreover, the adversarial modules will not be able discover the alias set of legitimate ones. Thus, the results depend solely on game performance and are independent of the encrypted data. The results are presented in Table I. Recall that the legitimate set is identified when its normalized score over L tests is higher than that of the adversarial set.

TABLE I
Simulation Results Summary

N	M	L	Legit. score	Non-Legit. score	Success
10	10	100	1	0	Yes
10	9	100	0.988	0.012	Yes
10	8	100	0.982	0.009	Yes
10	7	100	0.968	0.012	Yes
10	6	100	0.908	0.023	Yes
10	5	100	0.840	0.023	Yes
10	4	100	0.648	0.058	Yes
10	3	100	0.496	0.072	Yes
10	2	100	0.224	0.097	Yes
10	1	100	0.093	0.1	No
10	0	100	0	0.1	No

While it is common to assume that the number of adversarial devices is significantly lower than the number of legitimate ones for any effective security technique, the proposed approach demonstrates high performance even when the proportion of legitimate devices is low. The results show that reliable devices consistently win when the number of adversarial modules is below eight. For instance, the results in Table I show that the legitimate set can still be identified through testing even when adversarial devices constitute 70% of the total population. This highlights the effectiveness of the WereWolf game-based approach. It is important to note that although the test is conducted 100 times, the RIC has no knowledge of the responses, as the information remains encrypted using homomorphic encryption. The results outperfom by far the random selction where the sucess rate is in the range M/N.

The convergence time is proportional to the average number of iterations before the game ends. Fig. 2 illustrates the relationship between the number of malicious devices and the average number of iterations per game. The results indicate that convergence time increases as the proportion of legitimate devices grows relative to adversarial ones. For instance, the

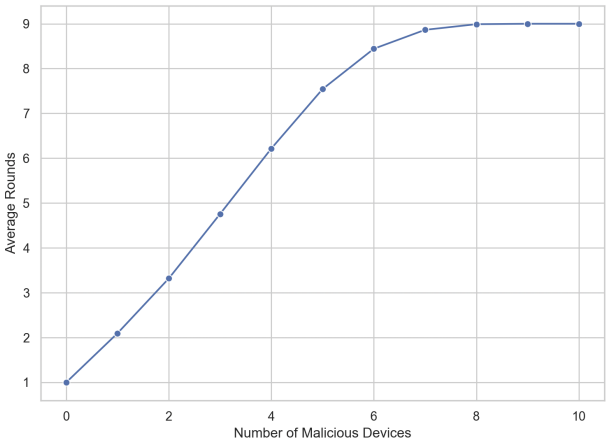

Fig. 2. Average Rounds vs. Number of Malicious Devices

average number of iterations per game is approximately five for $M = 7$ and three for $M = 8$, reaching its peak when $N_{\text{malicious}} = 9$.

In comparison, random selection converges in a single iteration, which provides an advantage in terms of speed. However, as shown in the figure, the increase in the number of iterations remains moderate. Overall, as the proportion of adversarial modules rises, the system requires more iterations to reliably converge to a compliant state.

V. Conclusion

This paper introduced a privacy-preserving compliance verification mechanism using a game-theoretic approach to distinguish legitimate modules from adversarial ones. By integrating homomorphic encryption with a probabilistic elimination strategy, the system ensures compliance verification without relying on predetermined benchmarks or centralized trust authorities. The proposed method allows legitimate nodes to iteratively identify and eliminate adversarial entities while preserving privacy.

Simulation results demonstrate high detection accuracy and resilience in adversarial conditions, showing that the system effectively identifies legitimate modules even when adversarial presence is significant. Furthermore, the convergence analysis indicates that the approach remains computationally efficient, with the number of required iterations scaling predictably with the adversarial proportion.

Future research will focus on optimizing the computational efficiency of HE and extending the framework to adaptive adversarial environments. Additionally, exploring cryptographic optimizations, such as post-quantum secure ZKPs and fully homomorphic encryption, will further enhance security and scalability. This work contributes to the development of a benchmark-free, decentralized trust enforcement mechanism, offering a robust security solution for next-generation networks.

Acknowledgment

The authors would like to thank Prof. Erkay Savas for the technical insights, which helped shape some of the ideas presented in this paper.

References

[1] S. K. Singh, R. Singh, and B. Kumbhani, "The evolution of radio access network towards open-RAN: Challenges and opportunities," in *Proc. IEEE WCNCW*, 2020.

[2] V. Dixit, J. Plachy, K. Sun, A. Ikami, E. Obiodu, and K. Lee, "O-RAN Towards 6G," *O-RAN next Generation Research Group (nGRG) Research Report*, pp. 1–28, 2023.

[3] L. M. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, "The evolution of mobile network operations: A comprehensive analysis of open RAN adoption," *Computer Networks*, p. 110292, 2024.

[4] M. Alavirad, U. S. Hashmi, M. Mansour, A. Esswie, R. Atawia, G. Poitau, and M. Repeta, "O-ran architecture, interfaces, and standardization: Study and application to user intelligent admission control," *Frontiers in Communications and Networks*, vol. 4, p. 1127039, 2023.

[5] A. S. Abdalla and V. Marojevic, "End-to-end O-RAN security architecture, threat surface, coverage, and the case of the open fronthaul," *IEEE Communications Standards Magazine*, vol. 8, no. 1, pp. 36–43, 2024.

[6] A. Arnaz, J. Lipman, M. Abolhasan, and M. Hiltunen, "Toward integrating intelligence and programmability in open radio access networks: A comprehensive survey," *IEEE ACCESS*, vol. 10, pp. 67 747–67 770, 2022.

[7] D. Telekom, "Open ran security white paper," 2022. [Online]. Available: https://www.o-ran.org/oran-ecosystem-resources/open-ran-security-white-paper-march-2022

[8] O.-R. ALLIANCE, "Zero trust architecture for secure o-ran," 2024. [Online]. Available: https://mediastorage.o-ran.org/white-papers/O-RAN.WG11.ZTA%20for%20Secure%20O-RAN%20White%20Paper-2024-05.pdf

[9] H. Sedjelmaci, N. Kaaniche, and K. Tourki, "Secure and resilient 6G RAN networks: A decentralized approach with zero trust architecture," *Journal of Network and Systems Management*, vol. 32, no. 2, pp. 1–23, 2024.

[10] H. Liu, M. Ai, R. Huang, R. Qiu, and Y. Li, "Identity authentication for edge devices based on zero-trust architecture," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 23, p. e7198, 2022.

[11] I. Lotfi, M. Qaraqe, A. Ghrayeb, and D. Niyato, "Vmguard: Reputation-based incentive mechanism for poisoning attack detection in vehicular metaverse," *arXiv preprint arXiv:2412.04349*, 2024.

[12] N. Oren, T. J. Norman, and A. Preece, "Subjective logic and arguing with evidence," *Artificial Intelligence*, vol. 171, no. 10-15, pp. 838–854, 2007.

[13] A. Jøsang, "A logic for uncertain probabilities," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 03, pp. 279–311, 2001.

[14] A. Josang, R. Hayward, and S. Pope, "Trust network analysis with subjective logic," in *Conference Proceedings of the Twenty-Ninth Australasian Computer Science Conference (ACSW 2006)*. Australian Computer Society, 2006, pp. 85–94.

[15] Y. Liu, K. Li, Y. Jin, Y. Zhang, and W. Qu, "A novel reputation computation model based on subjective logic for mobile ad hoc networks," *Future Generation Computer Systems*, vol. 27, no. 5, pp. 547–554, 2011.

[16] A. Jøsang, "The consensus operator for combining beliefs," *Artificial Intelligence*, vol. 141, no. 1-2, pp. 157–170, 2002.

[17] L. M. Kaplan, M. Şensoy, Y. Tang, S. Chakraborty, C. Bisdikian, and G. De Mel, "Reasoning under uncertainty: Variations of subjective logic deduction," in *Proceedings of the 16th International Conference on Information Fusion*. IEEE, 2013, pp. 1910–1917.

[18] S. Dietzel, R. Van Der Heijden, H. Decke, and F. Kargl, "A flexible, subjective logic-based framework for misbehavior detection in V2V networks," in *Proceeding of IEEE international symposium on a world of wireless, mobile and multimedia networks 2014*. IEEE, 2014, pp. 1–6.

[19] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2009.