AI Control for Trust-based Detection of Attackers in 5G Social Networks

Arjan Durresi[1], Davinder Kaur[1], Suleyman Uslu[1], and Mimoza Durresi[2]

[1]Indiana University Purdue University Indianapolis
[2]Universiteti Europian i Tiranes

October 3, 2023

Abstract

This study presents a framework for detecting and mitigating fake and potentially attacking user communities within 5G social networks. This framework utilizes geo-location information, community trust within the network, and AI community detection algorithms to identify users that can cause harm. The framework incorporates an artificial control model to select appropriate community detection algorithms and employs a trust-based strategy to identify and filter out potential attackers. It adapts its approach by utilizing user and attack requirement data through the artificial conscience control model while considering the dynamics of community trust within the network. What sets this framework apart from other fake user detection mechanisms is its capacity to consider attributes challenging for malicious users to mimic. These attributes include the trust established within the community over time, the geographical location, and the framework's adaptability to different attack scenarios. To validate its efficacy, we apply the framework to synthetic social network data, demonstrating its ability to distinguish potential malicious users from trustworthy ones.

Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

AI Control for Trust-based Detection of Attackers in 5G Social Networks

Davinder Kaur[1] | Suleyman Uslu[2] | Mimoza Durresi[3] | Arjan Durresi[4]

[1]Indiana University Purdue University, Indianapolis, Indiana, USA, Email: davikaur@iu.edu

[2]Indiana University Purdue University, Indianapolis, Indiana, USA, Email: suslu@iu.edu

[3]European University of Tirana, Tirana, Albania, Email: mimoza.durresi@uet.edu.al

[4]Indiana University Purdue University, Indianapolis, Indiana, USA, Email: adurresi@iu.edu

Abstract

This study presents a framework for detecting and mitigating fake and potentially attacking user communities within 5G social networks. This framework utilizes geo-location information, community trust within the network, and AI community detection algorithms to identify users that can cause harm. The framework incorporates an artificial control model to select appropriate community detection algorithms and employs a trust-based strategy to identify and filter out potential attackers. It adapts its approach by utilizing user and attack requirement data through the artificial conscience control model while considering the dynamics of community trust within the network.

What sets this framework apart from other fake user detection mechanisms is its capacity to consider attributes challenging for malicious users to mimic. These attributes include the trust established within the community over time, the geographical location, and the framework's adaptability to different attack scenarios. To validate its efficacy, we apply the framework to synthetic social network data, demonstrating its ability to distinguish potential malicious users from trustworthy ones.

1 | INTRODUCTION

In an era marked by the convergence of two significant technological domains, 5G networks and Artificial Intelligence (AI) systems, our research explores security issues and the innovative methods AI can apply to detect and mitigate attacks within the 5G landscape. The imminent deployment of 5G networks holds the potential to connect countless Internet of Things (IoT) devices, ranging from robots to autonomous vehicles. However, as this intricate ecosystem takes form, it enhances the vulnerability landscape, making it imperative to establish robust security measures.

At the forefront of wireless communication, 5G networks introduce complexities arising from various applications and intricate infrastructure components. While they offer enhancements in speed, connectivity, and reduced latency, their value lies in their capacity to provide security and resilience[1]. The threat exposure landscape for 5G is expected to be much larger than 4G due to more 5G devices and more types of devices that rely on 5G connectivity. Local attacks targeting real-time applications such as self-driving cars, robots, and similar ones are expected to rise. The critical component of 5G infrastructure/technologies is Multi-access Edge Computing (MEC), which will allow the implementation of complex algorithms, including AI, to increase security.

[0]**Abbreviations:** ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

Concurrently, AI systems have become essential to contemporary life and are extensively employed to identify attackers in 5G social networks. Nonetheless, concerns persist about the safety and dependability of AI systems. This is where the necessity for effective control mechanisms becomes evident, and ethical guidelines increasingly emphasize the significance of human agency in governing AI to enhance its adaptability, safety, and reliability.

Our research embarks on a journey to explore how AI can be harnessed to detect and respond to attacks within the 5G environment. We introduce a novel framework, the "Artificial Conscience and Trust-Based Framework for Fake User Detection in 5G Networks," which leverages AI's artificial conscience model capabilities to adapt dynamically to threats within the 5G network. By integrating AI-driven methods with trust-based techniques and graphical properties, our framework addresses fake user attacks, fortifying security within 5G networks and applications.

Our research navigates through this pioneering framework's theoretical underpinnings, methodological intricacies, and practical applications. By infusing AI control into 5G security, we aim to underscore the significance of controlling AI in this context. By mitigating risks posed by potential attackers, our research seeks to foster heightened security and trust within this rapidly evolving technological landscape. This interdisciplinary approach addresses how AI can be used for fake users and attack detection in 5G. It underscores why control mechanisms are indispensable in safeguarding this critical intersection of technology.

The structure of this paper is as follows: In Section 2, we provide background information and review prior work in the realms of artificial conscience, trust, and fake user detection. Section 3 outlines our novel framework, the "Artificial conscience and trust-based framework to detect fake users." Section 4 offers a practical demonstration of our framework using a synthetic social network dataset. Finally, in Section 5, we present our concluding remarks.

2 | BACKGROUND AND RELATED WORK

This section provides an overview of the historical context and prior research on fake user detection, trust mechanisms, and artificial conscience.

2.1 | AI Control and Artificial Conscience

In the age of pervasive AI integration, the necessity for artificial conscience, or machine conscience, becomes increasingly apparent. This control mechanism plays an essential role in guiding AI behavior ethically, ensuring the reliability and safety of critical decision-making processes, fostering trust and acceptance among users, adapting to dynamic and intricate environments, adhering to evolving regulations, promoting harmonious collaboration between humans and AI, and mitigating biases and discrimination[2]. Artificial conscience is a vital safeguard, aligning AI systems with societal values and ethical norms while enhancing their effectiveness and acceptance in an ever-evolving technological landscape.

Researchers have proposed diverse objectives achievable through artificial conscience, including autonomy, resilience, self-motivation, and information integration[3]. Realizing these objectives necessitates the development of conscious machines capable of emulating specific facets of the human conscious experience. A comprehensive comprehension of consciousness is pivotal in this pursuit. Cognitive neuroscientists like Baars introduced the Global Workspace Theory (GWT) as an analogy to explain consciousness[4], and the concept of the Conscious Turing Machine (CTM) or Conscious AI rooted in theoretical computer science[5]. These theories explore various aspects of consciousness, such as competition among processors and its relation to emotions, contributing to ongoing efforts to integrate elements of human conscience into AI systems[6,7]. All this work collectively on comprehending the human conscience and exploring the potential incorporation of certain facets into artificial intelligence.

The artificial conscience can be a vigilant guardian against deceptive and malicious activities. This heightened ethical sensitivity enables AI algorithms to discern subtle patterns and behaviors indicative of fake user profiles or fraudulent activities. Furthermore, artificial conscience can facilitate adaptive responses, allowing AI to dynamically adjust its detection strategies in response to evolving tactics employed by malicious actors, thereby increasing the security and integrity of 5G networks against the proliferation of fake users and associated threats.

2.2 | Role of Trust

Trust, a fundamental aspect of human relationships, is characterized by the confidence one entity has in the expected behavior of another[8]. In our daily social interactions, trust evolves based on the history of interactions and feedback exchanged among individuals or entities[9]. Greater trust tends to develop between entities with a history of positive interactions, while those with fewer positive and more negative interactions tend to have lower trust. Nonetheless, in the current data-abundant world, establishing trust similar to real life becomes complex. Therefore, a trust framework is necessary to quantify cognitive trust, transforming it into quantifiable metrics that facilitate informed decision-making for entities. Within the realm of fake user detection, this trust framework plays a pivotal role in assessing the trustworthiness of social network users through their interactions with peers[10]. This trust data supplements other characteristics in the identification of potential malicious users.

Numerous studies have explored trust modeling and its application in diverse scenarios. For instance, a trust framework proposed by[11], grounded in measurement theory, finds applicability across various domains, including crime detection[12], social networks[13,14], the food-energy sector[15,16,17,18,19,20], healthcare[21,22], edge computing[23], quantum computing[24], and beyond. Recent work[21,25,26,27] has highlighted trust's utility as an acceptance criterion for artificial intelligence algorithms. The widespread use of trust in these applications underscores its potential to identify malicious users within the social network, thereby incorporating community knowledge into the detection process.

2.3 | Geo-Location and Trust-Based Fake User Detection

A significant challenge facing contemporary social networks revolves around identifying fraudulent users. These deceptive individuals establish counterfeit profiles intending to disseminate false information and engage in nefarious activities[28]. Their objective involves creating a genuine and inconspicuous online presence, allowing them to avoid suspicion and gain the trust of fellow users. Detecting these deceptive individuals is crucial for safeguarding the integrity of social networks and preventing potential harm. The motivation driving these deceptive users is rooted in the significant incentives associated with their malicious activities. They endeavor to mimic authentic users by concealing their true identities, expanding their social connections, and engaging with other users. For example, they may utilize authentic user images and profile details to avoid detection[29].

Extensive research has been conducted in fake user detection, with two prevalent approaches being analyzing graph properties and utilizing machine learning algorithms. One graph-based approach involves examining various social graph characteristics[30]. Some researchers have proposed machine learning techniques, such as classification and clustering, which rely on profile attributes for fake user identification[31]. While these methods offer valuable insights for detection, they often overlook the adaptability and determination of fraudulent account creators. Incorporating community trust information becomes invaluable to capture the adaptive nature of fake users. This trust data can be computed using a trust framework, pivotal in summarizing users' relationships within social networks. There is a need for methods that consider community knowledge, geo-location of the users, and graphical properties. Therefore, we have proposed a geo-location and trust-based mechanism to capture the geographical location and community knowledge to detect fake users. This framework considers the geographical location of users and their trust values calculated using community knowledge for the detection. The geo-location information complements various malicious user detection techniques. With the introduction of 5G networks, which are widely being adapted, we have precise geo-location information[32] that can be used in fake user detection techniques. 5G networks include many advances in wireless networking[33,34,35,36,37,38]. Furthermore, 5G systems provide ubiquitous geo-location information with 1-meter accuracy utilizing a multitude of satellite and ground support systems[32]. Because of this, today's communication is moving more towards location-aware communication, and this information can be utilized to improve fake user detection techniques.

3 | ARTIFICIAL CONSCIENCE AND TRUST-BASED FRAMEWORK TO DETECT FAKE USERS

This section introduces a security system founded on artificial conscience and trust principles to enhance the efficacy of fake user detection. Within this framework, the artificial conscience control module and trustworthiness, computed through community-based knowledge, are integral to the detection process. To facilitate evaluation, we introduce a metric for categorizing social network user communities as fake or genuine. Section 3.1 outlines the trust framework, which computes user trust based on interactions. Section 3.2 explores various community detection algorithms, while Section 3.3 delves into the artificial conscience control module. Lastly, Section 3.4 explains the evaluation metrics employed for community classification.

3.1 | Trust Framework

Within social networks, the trust framework plays a pivotal role by providing a means to consolidate trust information gleaned from users' historical interactions. As an illustration, Twitter's retweets, likes, and comments are used to gauge trustworthiness, whereas Facebook indicators like likes, comments, and shares are used for trust assessment [39]. The trust framework [11] is used to quantify user trust. The two-component approach is used to calculate trust, comprising Impression and Confidence.

Impression (m): Impression, also termed trustworthiness, represents the extent of trust that one entity places in another. It comprehensively summarizes all direct and indirect interactions among entities within social networks. In this context, impression is computed as the mean of measurements obtained from interactions among entities, following the formula in Equation 1.

$$m = \frac{\sum_{i=1}^{i=N} m_i}{N} \tag{1}$$

Confidence (c): Confidence quantifies an entity's level of certainty concerning its perceptions of another entity's trustworthiness. It measures an entity's confidence in its judgments of others' trustworthiness. This component is instrumental in accounting for errors during impression calculations and is closely linked to the variability of these measurements. In cases where measurement variance is low, an entity exhibits high confidence in its impressions of others, as depicted in Equation 2.

$$c = 1 - 2 * e \text{ where } e = \sqrt{\frac{\sum_{i=1}^{i=N} (m_i - m)^2}{N * (N-1)}} \tag{2}$$

Trust encompasses impression and confidence, serving as a metric to quantify user trust within social networks. When direct connections between users are limited, trust inference is carried out using aggregation and transitivity operators [11].

Trust Aggregation: When multiple pathways exist between two entities, trust aggregation efficiently consolidates trust. For instance, if there are two paths (A-B-D and A-C-D) to reach entity D starting from A, the aggregated trust is calculated as a weighted combination of trust measurements from both paths, as delineated in Equations 3 and 4.

$$m_D^{A:B} \oplus m_D^{A:C} = \frac{A_1 * m_D^{A:B} + w_2 * m_D^{A:C}}{\sum w_i} \tag{3}$$

$$e_D^{A:B} \oplus e_D^{A:C} = \sqrt{\frac{1}{(\sum w_i)^2}(w_1^2 * (e_D^{A:B})^2 + w_2^2 * (e_D^{A:C})^2))} \tag{4}$$

Trust Transitivity: Trust transitivity comes into play when direct connections between two entities are absent and is employed to compute indirect trust. For instance, if there is no direct path from entity A to C but a path exists from A to B and another from B to C, transitive trust utilizes the trust values between A-B and B-C to determine the trust from A to C, as outlined in Equations 5 and 6.

$$m_B^A \otimes m_C^B = m_{min} = min(m_B^A, m_C^B) \tag{5}$$

$$e_B^A \otimes e_C^B = min(e_i \text{ where } m_i = m_{min}) \tag{6}$$

This unique trust framework serves as a valuable tool for assessing user trust within online social networks, aiding in decision-making processes, and reinforcing the reliability and trustworthiness of interactions.

3.2 | Community Detection in 5G social network

Real-world social networks inherently exhibit a graph-like structure, which can be formally represented as a graph G(V, E) where nodes (V) correspond to entities or users within the social network, while edges (E) symbolize the relationships between these entities. These social network graphs often encompass a multitude of nodes, necessitating the development of efficient methods for information retrieval. An optimistic approach to analyzing these networks involves partitioning the network into communities characterized by densely interconnected nodes within each community and sparser connections with nodes in other communities. Much work has been done to detect such communities effectively and rapidly. These community detection algorithms are broadly categorized into two types: Agglomerative methods, where edges are incrementally added to the graph from strongest to weakest, and Divisive methods, where edges are removed one by one based on edge weights.

Agglomerative Community Detection Algorithms:

Clauset-Newman-Moore Algorithm: This algorithm operates by incrementally merging nodes into communities to optimize modularity, a measure of community quality[40]. It starts with each node as an individual community and iteratively combines them to form larger communities, effectively building the hierarchy of communities within a network. Evaluating the modularity score at each step identifies communities that contribute positively to modularity, resulting in meaningful network partitioning. The algorithm exhibits a computational complexity of O(ldlogn), where l represents the number of edges, n denotes the number of nodes, and d signifies the depth of the dendrogram.

Label Propagation Algorithm: The Label Propagation Algorithm begins with each node assigned a unique label[41]. In each iteration, nodes update their labels based on the majority label among their neighboring nodes. This process continues until stable labeling is achieved, with nodes predominantly sharing labels with their neighbors. The resulting labels correspond to communities within the network. The computational complexity of this algorithm is O(ldlogn).

Community Detection Using Modularity Optimization (Louvain Algorithm): The Louvain Algorithm is an unsupervised agglomerative approach that iteratively optimizes modularity[42]. It begins by assigning each node to its community and then merges communities to maximize the modularity score. The modularity optimization and community aggregation phases work in tandem to identify communities that contribute positively to modularity. It is well-suited for large networks due to its lower computational complexity of O(nlogn).

Divisive Community Detection Algorithms:

Girvan-Newman Algorithm: The Girvan-Newman Algorithm is a divisive approach that starts with the entire network as a single community[43]. It iteratively removes edges with the highest betweenness centrality, effectively breaking bridges between communities. The process continues until the network is fragmented into distinct communities. It is particularly effective at detecting communities that are well-connected internally but have limited connections with other communities. The computational complexity of this algorithm scales as $O(l^2n)$, rendering it impractical for large social networks. This algorithm becomes ineffective when the number of nodes surpasses a few thousand.

Communities detected using the above algorithms are evaluated using modularity, coverage, and performance measures to assess the quality of communities detected by various algorithms.

- Modularity: Modularity gauges the quality of partitioning nodes into communities within a network. It quantifies the discrepancy between the count of edges within communities and the anticipated number of edges in a random network possessing the same degree distribution. Higher modularity values indicate a more favorable community structure.
- Coverage: Coverage appraises the proportion of edges contained within communities relative to the total number of edges in the network. A greater coverage implies a more successful assignment of edges to communities.
- Performance (P): Performance evaluates communities' quality by considering intra-community edges and non-edges. It measures the ratio of the sum of edges within communities and non-edges between communities to the total potential edges. Elevated performance values signify a well-defined community structure.

3.3 | Artificial Conscience Control Module

The Artificial Conscience Control Module is vital for controlling AI systems in alignment with user-defined requirements. Users' distinct expectations and demands shape the operation of AI systems, injecting meaning into otherwise algorithmic processes[44]. This framework assumes a decision-making task for the AI system, involving deploying multiple machine-learning algorithms and a range of metrics to evaluate potential solutions. These metrics, referred to as "agents," negotiate with one another based on user-assigned weights and trust derived from their peers to compute an "Artificial Feeling" (AF) as a weighted average among agents. Figure 1 explains our artificial conscience control model.

The process unfolds through the following stages:

- User Specification: Users articulate their expectations and demands concerning AI system performance. They allocate weights to evaluation metrics, imparting significance and context to the decision-making process.
- Agent Initialization: Each evaluation metric, acting as an "agent," is initialized to achieve the most suitable solution based on its unique criteria. These criteria may relate to machine learning algorithms, metrics, or parameters.
- Algorithm Selection: Agents can have diverse priorities, including selecting machine learning algorithms that best suit specific attack detection requirements. Different agents may emphasize the importance of distinct algorithms within the overall solution.

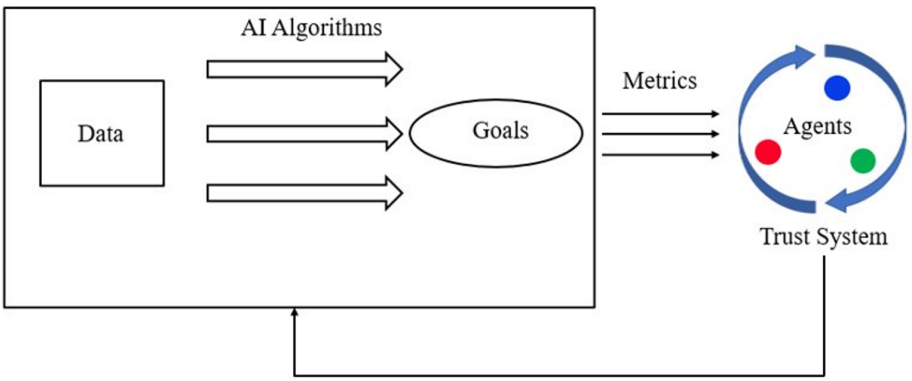

FIGURE 1 Artificial Conscience Control Module to Control AI.

- Negotiation Rounds: Agents engage in a series of negotiation rounds (typically n rounds), the exact number being contingent on the application or user needs.
- Artificial Feeling (AF) Computation: The negotiation process culminates in the computation of an "Artificial Feeling" (AF), which represents a consensus or amalgamation of solutions from all agents. This composite solution balances all participating agents' preferences, trust levels, and algorithm selection choices.
- Final Decision: The AI system embraces the AF as the ultimate decision or solution. This outcome reflects a collective agreement accommodating diverse user requirements, including selecting machine learning algorithms tailored to specific attack detection demands.

In summary, the Artificial Conscience Control Module facilitates the alignment of AI systems with user expectations, considering the choice of machine learning algorithms and other parameters crucial for effective attack detection. This process ensures meaningful human involvement in AI-driven decision-making, promoting transparency and adaptability in complex environments.

3.4 | Community Evaluation Metrics

Various assessment criteria are available to detect fake communities within the social network. The analytical process is vital in distinguishing legitimate user communities from counterfeit users. In devising these metrics, we operate under the premise that counterfeit users are prone to having few or no connections with genuine users. It is improbable for a genuine user to establish connections with unfamiliar individuals. Counterfeit users, however, endeavor to simulate authentic user behavior by amassing connections, often with fellow counterfeit users. Building on these assumptions, we have formulated evaluation metrics characterized by three key attributes: Density, Time to Create Community, and Trust.

- Density: This attribute is pivotal in identifying communities or clusters of counterfeit users. Counterfeit user communities tend to exhibit a higher concentration of connections than communities of genuine users, primarily because genuine users are less likely to connect with counterfeit profiles. Counterfeit users often foster numerous connections among themselves. The density of a community is represented as a ratio, quantifying the number of edges within the community relative to the maximum possible edges it could contain, as defined by Eq. 7. Here, 'E' signifies the number of edges, and 'V' represents the number of nodes or users within the community.

$$d = \frac{2 * |E|}{|V|(|V| - 1)} \tag{7}$$

  Notably, counterfeit user communities tend to exhibit markedly higher density values than those comprising genuine users, reflecting their inherent interconnectedness, whereas genuine user communities are typically more dispersed. The

density attribute is leveraged with other attributes, such as trust and time, to create communities to detect counterfeit user communities.

- Time to Create Community: Counterfeit users establish connections more rapidly than their genuine counterparts. Their motivation to rapidly amass connections leads to the formation of high-density communities in a relatively short time frame. Both density and the time taken to create a community are pivotal factors in distinguishing communities of genuine users from those of counterfeit users.

- Trust: Trust is crucial in identifying counterfeit user communities, enabling informed decision-making grounded in community insights. Trust serves as a summary of user relationships, encompassing both intra-community and external connections.

  It can be applied to counterfeit user detection in two ways: Average Trust and Trust Over Time. Average Trust consolidates trust values from all users within the community. This is particularly pertinent because counterfeit user communities endeavor to maintain elevated trust values to evade detection when disseminating malicious content. They achieve this by engaging in more positive interactions among themselves. Trust Over Time, however, captures the temporal fluctuation of trust. Genuine users tend to experience more pronounced trust fluctuations due to their many positive and negative interactions. Conversely, counterfeit users tend to exhibit a consistent increase in trust over time.

These attributes, collectively incorporated within the framework of evaluation metrics, equip us with a comprehensive approach to detecting counterfeit user communities.

4 | IMPLEMENTATION

In this section, we provide an overview of the dataset utilized in our experiment, the implementation process, and the outcomes obtained by applying our framework.

4.1 | Dataset description

In our study, we employed a diverse dataset that is the foundation for evaluating our framework's effectiveness in detecting fake users within social networks. This dataset comprises real and fake user networks that mimic specific real-world characteristics. For the real user network, we utilized the Karate club friends network dataset and the Facebook ego network dataset, integrating genuine friendship relationships and Facebook social network attributes. Trust indicators represented by impression and confidence values and edge creation timestamps were randomly assigned to mirror authentic user behavior. In contrast, to simulate the deceptive behavior of fake users, we employed the Erdos-Renyi model to generate social network graphs, emphasizing a higher probability of edge creation to capture the densely interconnected nature of fake user communities. Like real users, fake user nodes were endowed with impression and confidence values, with edge creation timestamps randomly assigned. This comprehensive dataset allows us to rigorously assess our approach to detecting fake users in a realistic social network context. For our experiment, we have considered only the users within a 100-mile radius. This is done by calculating the distance using latitude and longitude coordinates.

4.2 | Experimentation and Results

We have employed four distinct community detection algorithms, Clauset-Newman-Moore, Label Propagation, Girvan-Newman, and Louvain, to identify communities within social networks. Specifically, we restricted our analysis to real social network graph data, including the karate networks and Facebook ego networks. The objective was to conduct a comparative assessment of these diverse community detection algorithms.

To evaluate the quality of the communities identified by these algorithms, we utilized three key quality metrics: Modularity, coverage, and performance. Figures 2 and 3 concisely summarize the metric values obtained for each community detection algorithm applied to both datasets. It is important to note that different algorithms may exhibit varying degrees of performance across these metrics, and we leveraged an artificial conscience model to make informed selections based on the specific characteristics of each network and attack scenario.

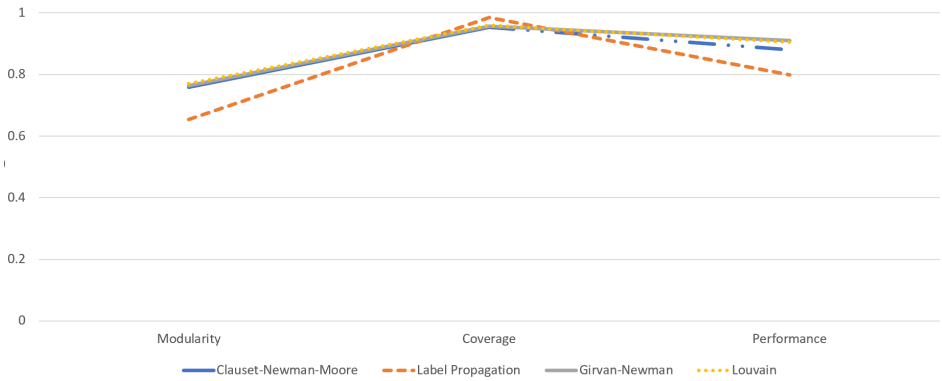

FIGURE 2 Modularity, Coverage, and Performance value for all the community detection algorithms on the Facebook Dataset.

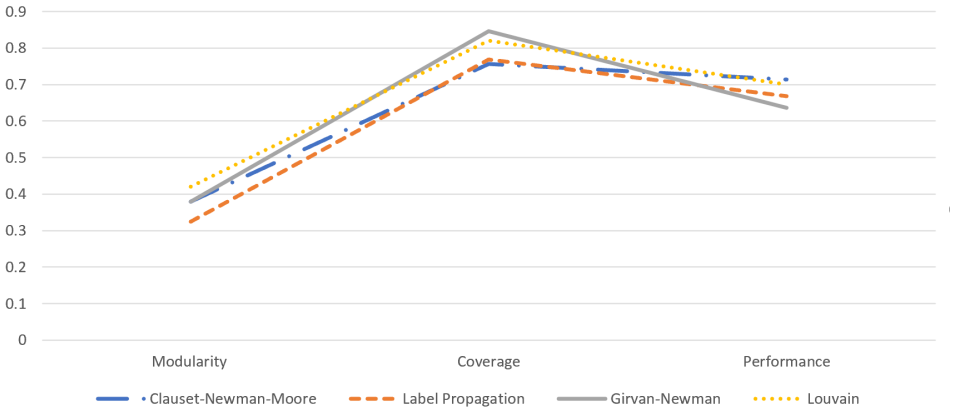

FIGURE 3 Modularity, Coverage, and Performance value for all the community detection algorithms on the Friend Dataset.

As evident from the visual representations, the label propagation algorithm achieves the highest coverage in the Facebook dataset, whereas the Girvan-Newman algorithm demonstrates superior coverage in the Friends dataset. This observation underscores the importance of context and specific requirements in algorithm selection.

The artificial conscience model is crucial in tailoring the algorithm selection process to meet specific requirements. When a user emphasizes a single metric, we opt for the algorithm that performs best in that regard. However, in cases where the user assigns greater importance to more than one metric, the algorithm selection is adapted accordingly to align with those priorities. This highlights why the artificial conscience model must tailor the algorithm selection based on the network properties and user requirements. For our experimentation, we utilized the Louvain algorithm, as it outperformed other algorithms across all metrics and for both datasets. Following identifying communities within the dataset, these communities are assessed using the evaluation metrics described in section 3.4.

Once communities have been identified, they are categorized as either genuine or fraudulent user communities, depending on criteria such as community density, community trust value, and the time taken for their formation. Within our dataset, we have identified 11 communities, and Figure 3 illustrates the trust and density values associated with these communities.

Based on the trust and density value, we have observed the following patterns across the communities:

In Figure 4, we can observe the following patterns:

- Communities with low trust and low-density values (Community 4 and 5) indicate users who primarily observe others' interactions, engaging minimally with others. They exert little to no influence on other users.

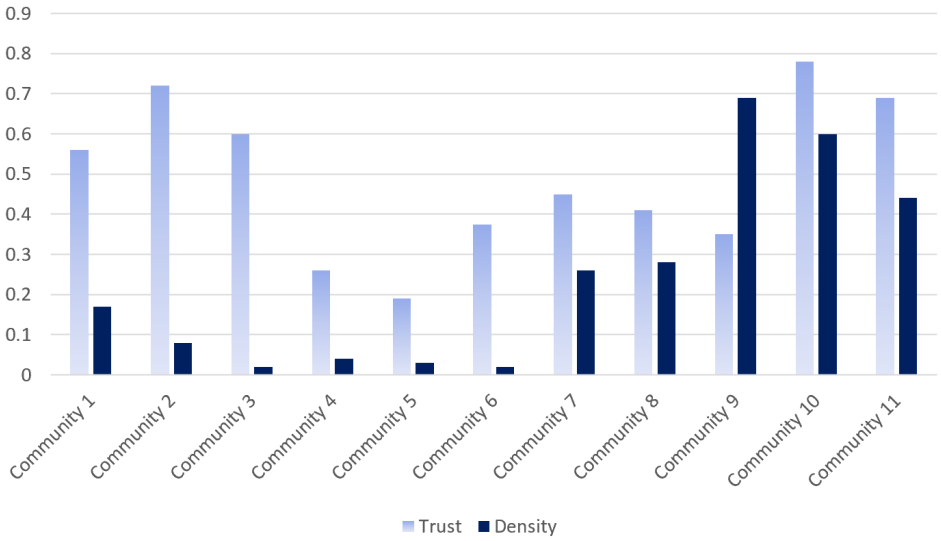

FIGURE 4 Trust and Density values for all the communities detected.

- Communities displaying low trust and high density (Communities 7, 8, and 9) suggest that users within them have received fewer positive interactions from other tightly connected users, making them less trustworthy and easily detectable.
- Communities characterized by high trust, low density, and a high number of interactions (Communities 1, 2, 3, and 6) consist of well-trusted users who engage in numerous positive interactions within the community.
- Communities exhibiting high trust and high-density values (Community 10 and 11) are potentially the communities of fake users seeking to boost their trust levels to intentionally spread malicious information artificially. Users in these communities tend to have exceptionally high trust connections within a small, dense community while having significantly fewer connections with the outside world, raising concerns about their potential for causing harm.

To conduct a more detailed examination of these communities, we have chosen four communities, each representing one of the social networks (Community 11, 9, 8, and 4). For each of these selected communities, we've undertaken two key analyses: We first determine the time it takes for each community to form based on the timestamp of edge creation. Second, we analyze the trust dynamics over time. In Figure 5, we present a six-month trust variation graph for these communities, offering valuable insights:

- Communities with potential fake users inclined to cause harm tend to maintain consistently high trust values with minimal fluctuations over time.
- Genuine user communities, whether established over a year ago or formed within six months, generally exhibit an average trust value with significant variations over time. This variation results from their diverse interactions with various users within the network.
- Communities characterized by low trust values and minimal variations typically consist of less trustworthy users or passive observers who exert minimal influence on the overall network dynamics.

5 | CONCLUSION

This paper introduces a trust-based framework empowered by an artificial conscience control model for identifying fraudulent users within social networks. This innovative framework combines elements such as graphical network attributes, community detection algorithms, geo-location information, and community trust knowledge to categorize social network users effectively. It distinguishes itself from other detection techniques by its capacity to dynamically adjust according to user and attack requirements while leveraging users' inherent trustworthiness. This quality is challenging to counterfeit.

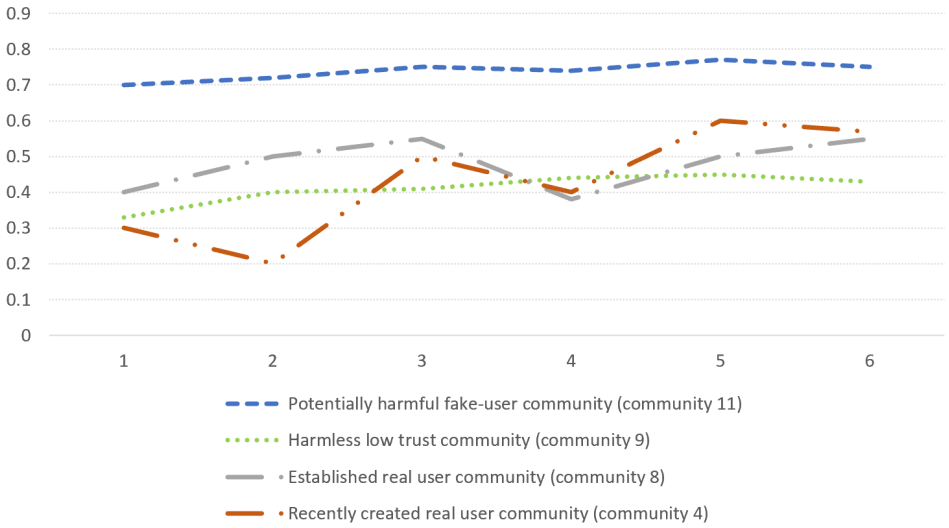

FIGURE 5 Trust variation across communities over time.

References

1. Gupta A, Jha RK. A survey of 5G network: Architecture and emerging technologies. *IEEE access* 2015; 3: 1206–1232.
2. Chella A, Manzotti R. *Artificial consciousness*. Andrews UK Limited . 2013.
3. Gamez D. *Human and machine consciousness*. Open Book Publishers . 2018.
4. Baars BJ. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA . 1997.
5. Blum L, Blum M. A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences* 2022; 119(21): e2115934119.
6. Russell S. *Human compatible: Artificial intelligence and the problem of control*. Penguin . 2019.
7. Solms M. *The Hidden Spring: A Journey to the Source of Consciousness*. Profile books . 2021.
8. Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* 2022; 55(2): 1–38.
9. Tang J, Gao H, Hu X, Liu H. Exploiting homophily effect for trust prediction. In: ; 2013: 53–62.
10. Kaur D, Uslu S, Durresi A. Trust-based security mechanism for detecting clusters of fake users in social networks. In: Springer. ; 2019: 641–650.
11. Ruan Y, Zhang P, Alfantoukh L, Durresi A. Measurement theory-based trust management framework for online social communities. *ACM Transactions on Internet Technology (TOIT)* 2017; 17(2): 1–24.
12. Kaur D, Uslu S, Durresi A, Mohler G, Carter JG. Trust-based human-machine collaboration mechanism for predicting crimes. In: Springer. ; 2020: 603–616.
13. Rittichier KJ, Kaur D, Uslu S, Durresi A. A Trust-Based Tool for Detecting Potentially Damaging Users in Social Networks. In: Springer. ; 2021: 94–104.
14. Kaur D, Uslu S, Durresi M, Durresi A. A geo-location and trust-based framework with community detection algorithms to filter attackers in 5G social networks. *Wireless Networks* 2022: 1–9.
15. Uslu S, Kaur D, Rivera SJ, Durresi A, Durresi M, Babbar-Sebens M. Trustworthy Fairness Metric Applied to AI-Based Decisions in Food-Energy-Water. In: Springer. ; 2022: 433–445.

16. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Trust-based decision making for food-energy-water actors. In: Springer. ; 2020: 591–602.

17. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M, Tilt JH. Control theoretical modeling of trust-based decision making in food-energy-water management. In: Springer. ; 2020: 97–107.

18. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M, Tilt JH. A trustworthy human–machine framework for collective decision making in food–energy–water management: The role of trust sensitivity. *Knowledge-Based Systems* 2021; 213: 106683.

19. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Trust-based game-theoretical decision making for food-energy-water management. In: Springer. ; 2019: 125–136.

20. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Decision support system using trust planning among food-energy-water actors. In: Springer. ; 2019: 1169–1180.

21. Kaur D, Uslu S, Durresi A, Badve S, Dundar M. Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable ai medical diagnostic systems. In: Springer. ; 2021: 35–46.

22. Kaur D, Uslu S, Durresi A. Trustworthy AI explanations as an interface in medical diagnostic systems. In: Springer. ; 2022: 119–130.

23. Uslu S, Kaur D, Durresi M, Durresi A. Trustability for resilient internet of things services on 5G multiple access edge cloud computing. *Sensors* 2022; 22(24): 9905.

24. Kaur D, Uslu S, Durresi A. Quantum Algorithms for Trust-Based AI Applications. In: Springer. ; 2023: 1–12.

25. Uslu S, Kaur D, Rivera SJ, Durresi A, Durresi M, Babbar-Sebens M. Trustworthy Acceptance: A New Metric for Trustworthy Artificial Intelligence Used in Decision Making in Food–Energy–Water Sectors. In: Springer. ; 2021: 208–219.

26. Kaur D, Uslu S, Durresi A. Requirements for trustworthy artificial intelligence–a review. In: Springer. ; 2020: 105–115.

27. Ruan Y, Durresi A. A survey of trust management systems for online social communities–trust modeling, trust inference and attacks. *Knowledge-Based Systems* 2016; 106: 150–163.

28. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. *Communications of the ACM* 2016; 59(7): 96–104.

29. Boshmaf Y, Muslukhov I, Beznosov K, Ripeanu M. The socialbot network: when bots socialize for fame and money. In: ; 2011: 93–102.

30. Breuer A, Eilat R, Weinsberg U. Friend or faux: Graph-based early detection of fake accounts on social networks. In: ; 2020: 1287–1297.

31. Roy PK, Chahar S. Fake profile detection on social networking websites: a comprehensive review. *IEEE Transactions on Artificial Intelligence* 2020; 1(3): 271–285.

32. Di Taranto R, Muppirisetty S, Raulefs R, Slock D, Svensson T, Wymeersch H. Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G. *IEEE Signal Processing Magazine* 2014; 31(6): 102–112.

33. Goyal M, Prakash S, Xie W, Bashir Y, Hosseini H, Durresi A. Evaluating the Impact of Signal to Noise Ratio on IEEE 802.15.4 PHY-Level Packet Loss Rate. In: ; 2010: 279-284

34. Yang T, Ikeda M, Mino G, Barolli L, Durresi A, Xhafa F. Performance Evaluation of Wireless Sensor Networks for Mobile Sink Considering Consumed Energy Metric. In: ; 2010: 245-250

35. Xie W, Goyal M, Hosseini H, et al. A Performance Analysis of Point-to-Point Routing along a Directed Acyclic Graph in Low Power and Lossy Networks. In: ; 2010: 111-116

36. Durresi A, Paruchuri V, Jain R. Geometric broadcast protocol for heterogeneous sensor networks. *Journal of Interconnection Networks* 2005; 6(03): 193–207.

37. Ikeda M, Barolli L, Hiyama M, Yang T, De Marco G, Durresi A. Performance evaluation of a manet tested for different topologies. In: IEEE. ; 2009: 327–334.

38. Barolli L, Honma Y, Koyama A, Durresi A, Arai J. A selective border-casting zone routing protocol for ad-hoc networks. In: IEEE. ; 2004: 326–330.

39. Ruan Y, Durresi A, Alfantoukh L. Using Twitter trust network for stock market analysis. *Knowledge-Based Systems* 2018; 145: 207–218.

40. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. *Physical review E* 2004; 70(6): 066111.

41. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 2007; 76(3): 036106.

42. De Meo P, Ferrara E, Fiumara G, Provetti A. Generalized Louvain method for community detection in large networks. In: IEEE. ; 2011: 88–93.

43. Newman ME, Girvan M. Finding and evaluating community structure in networks. *Physical review E* 2004; 69(2): 026113.

44. Kaur D, Uslu S, Durresi A. A Model for Artificial Conscience to Control Artificial Intelligence. In: Springer. ; 2023: 159–170.

45. Aggarwal CC, Wang H. A survey of clustering algorithms for graph data. *Managing and mining graph data* 2010: 275–301.

46. Girvan M, Newman ME. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 2002; 99(12): 7821–7826.

How to cite this article: D. Kaur, S. Uslu, M. Durresi, and A. Durresi, and (2023), AI Control for Trust-based Detection of Attackers in 5G Social Networks, , *2023;00:1–6*.