Accurate Colluding Agents Detection by Reputation Measures

Attilio Marcianò

Department of Information Engineering, Infrastructure and Sustainable Energy (DIIES), Mediterranea University of Reggio Calabria, via Graziella snc, loc. Feo di Vito - 98123 Reggio Calabria

Abstract

Software agents can form multidimensional, relationship based networks potentially able to realize not trivial forms of interaction and cooperation among agents. In similar contexts, honest agents could be exposed to malicious behaviors acted by unqualified potential partners. *Trust and Reputation Systems* are effective tools able to mitigate such risks by providing the agent community with suitably information about the trustworthiness of the potential partners in order to allow a good partner choice. In such a framework, we propose: (*i*) a method to preliminarily identify the best promising candidates as malicious assigning them the role of pre-untrusted entities and (*ii*) a novel reputation model capable to accurately identify malicious agents without introducing collateral effects on the reputation scores of honest ones.

Keywords

Agent System, Colluding, Malicious, Pre-untrusted, Reputation System

1. Introduction

A major trait of humans is their natural attitude to having mutual social relationships. One of the major ambitions of Artificial Intelligence is reproducing relevant features of human behaviors in virtual societies also including their social skills [1]. The adoption of agent technologies can augment the spectrum of the interactions in virtual societies. In particular, the adoption of the concept of social interaction applied to Multi-Agent Systems (MAS) is crucial to realize more intelligent, human-like, autonomous artificial systems capable of being engaged in complex forms of social relationships [2, 3], cooperation [4, 5], auto-coordination [6, 7], group formation [8, 9], etc. To this aim, software agents need both to be provided with emotions, intentions and beliefs in order to model cognitive aspects and personality like benevolence, selfishness, honesty, meanness, and to consider the presence of other social agents in planning strategies for reaching own goals. Therefore, in the following we will assume agents relationships qualitatively analogous to those occurring in human societies. Consequently, in human-like agent-based simulations the emergent behavior of a MAS arises as a result of the aggregation of individual agents' behaviors.

In the context outlined above, the presence of malicious agents should be considered. They are aimed to gain undue benefits by cheating on a wide range of different (and frequently concurrent)

WOA 2022: 23rd Workshop From Objects to Agents, September 1–2, Genova, Italy

attilio.marciano@unirc.it (A. Marcianò)
https://www.unirc.it/scheda_persona.php?id=1130 (A. Marcianò)
0000-0002-7229-5464 (A. Marcianò)

malicious behaviors, from simple ones (e.g. selfishness, misjudgment) to more sophisticated ones (e.g. colluding). In addition, in presence of large agent environments, nomadic agents, or other events, the probability of interacting with partners whose trustworthiness is still unknown might increase, in such a way agents will be exposed to greater risks of deceptions. Obviously, minimizing such risks is a main aspect for enhancing social aspect in agent communities.

Trust (i.e. reputation) is a powerful means to achieve this goal and making social interactions more satisfying as possible. Therefore, to minimize risks each agent could be provided with appropriate trust measures about the other agents belonging to the same community in order to improve the probability of interacting with reliable partners or, conversely, deciding of not interacting with anyone.

However, the concept of trust cannot be uniquely defined and measured because it is influenced by measurable and not measurable properties involving multiple dimensions (like competence, honesty, security, reliability, etc.) and depending on the specific situational context under which the interactions between the two agents happen. Due to this multifaceted nature, several meanings have been associated with the term "trust". From our viewpoint, we have interest in addressing subjectivity and situational risks, two aspects playing a fundamental role in the agent societies. To this purpose, in the following we will define trust as:

- "Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which its welfare depends" [10];
- "Trust is the extent to which one party is willing to depend on something or somebody in a given situation with a feeling of relative security, even though negative consequences are possible" [11].

Trust and Reputation Systems (TRSs) are equipping a great number of applications belonging to a large variety of scenarios [12, 13, 14, 15]. From a practical viewpoint, TRSs are Decision Support Tool because they enable agents' choices providing them with information about the trustworthiness of their potential partners on the basis of (*i*) direct information derived by a direct knowledge of the trustor about a trustee and/or (*ii*) indirect information considering ratings and/or opinions provided by other members of the own community about that trustee. Generally, TRSs represent the agent's trustworthiness by means of a unique score which can be used to identify dishonest actors from honest ones.

However, a main problem of TRSs in detecting malicious agents is that of adopting computational processes that not penalize the reputation scores of honest agents. Unfortunately, this is not unusual as, for instance, in the well known reputation system EigenTrust [16]. To this aim, our contribution is focused on proposing a novel reputation model, designed for agent-based social communities, which careful preserves the reputation scores of honest agents in detecting malicious ones. Moreover, we developed a technique to preliminarily select the most promising agent candidates to be identified as colluding and using them as pre-untrusted agents.

The rest of the paper is organized as follows. Section 2 introduces some related work. The Section 3 describes the reference agent scenario adopted for our reputation model, presented in Section 4. In Section 5, a case study provides a practical example for our proposed reputation model and, finally, some conclusions and future works are drawn in Section 6.

2. Related Work

In social agent communities a relevant issue is represented by recognizing malicious actors [17, 18, 19, 20, 21, 22]). *Trust and Reputation Systems* (TRSs) provide a defense against cheaters that might perform various misleading behaviors [23, 24, 25, 26] and the greater the degree of TRSs robustness to malicious attacks, the more reliable their trustworthiness measures will be [27].

In more detail, *Trust Systems* (TSs) combine information both direct raising by own past experiences (i.e. ”reliability”) and indirect given by opinions provided by other members of the community (i.e. ”reputation”), as in [28]. Differently, *Reputation Systems* (RSs) rely only on indirect information [29]. TRSs can also exploit single or multiple information sources, adopt centralized or distributed architectures or consider global or local approaches [30, 31, 32, 33].

Even though it is difficult comparing systems developed for specific scenarios, a number of studies addressed this issue with respect to a more or less wide range of malicious behaviors. To this aim, in [25, 34, 35] the defense mechanisms implemented by those TRSs against some common malicious attacks were compared. However, these studies are lack in providing well-defined quantitative approaches to assess the TRSs robustness.

To test TRSs, malicious attacks can be simulated on different scenarios, also in the form of competition among TRSs [36, 37]. In particular, some testbeds have been proposed, among which ART (Agent Reputation and Trust testbed) [38] is well known, other examples of testbeds can be found in [39, 40, 41]. Alternative mechanisms to testbeds exploit mathematical/analytical approaches [42, 43] that, from one hand, allow a more comprehensive verification of a TRS but, on the other hand, each TRS requires to develop specific test modalities.

eBay [44] is a popular and simple, but not robust RS [45, 44, 46], particularly with respect to collusive activities. Its reputation model consists of summing the single feedback provided by counterpart to increase or decrease a reputation score, leaving to each user its evaluation on the basis of their risk attitude. All newcomers receive a null reputation score, i.e. the minimum rating in eBay.

Both PeerTrust [47] and Hypertrust [48] are robust, distributed RS adopting a peer-to-peer overlay network, To identify the most suitable peers to interact, the former exploits several information referred to the specific context, direct feedback, credibility of the indirect feedback sources, number and nature of the transactions performed by each peer. The other one was conceived for large, competitive federations of utility computing infrastructures. In Hypertrust the nodes, linked via overlay, form clusters and a distributed algorithm discovers and allocates resources associated with trusted nodes to limit deceptive activities. The search of potentially interesting resource is reduced to an eligible region based on reputation information.

The well known EigenTrust [16] computes the global reputation of each peer, assuming the reputation transitivity, on the basis of the local trust matrix, storing the (normalized) trust scores that each peer has about the trustworthiness of the other peers in its community, weighted by the trustworthiness of each trustor peer. Based on their reputation scores, the peers will be differentiated in colluding and not colluding, but its computational process flatten the reputation scores of all the agents including honest ones.

A number of TRSs in distributed, semi-distributed, centralized and/or blockchain-based IoT environment has been presented in [12, 13, 49, 50]. In particular, open IoT environments are particularly risky for the increased possibilities to realize malicious behaviors. In [19], a distributed RS is proposed. It implements some countermeasures to detect malicious or cheating actors. A simulation dealing with vehicular mobility was proposed to test the effectiveness of the reputation model in quickly detecting malicious devices. RESIOT [20] is a framework to form groups of reliable IoT devices based on the reputation scores of their associated agents. In this proposal, a novel reputation model is presented and the results of a set of experiments simulating malicious attacks acting different, concomitant cheating strategies, have been compared with those of other RSs. Finally, in [51] a TS for a SIOT scenario is suggested; it adopts a machine learning technique to realize a resilient system towards a significant number of attacks and capable to detect the most part of cheaters at the number of transactions increases.

The TRSs presented in this section implement different and effective approaches to identify malicious actors. However, to the best of our knowledge none of them was explicitly designed to preserve the trust scores of honest actors in looking for malicious, unlike the TRS presented in Section 4.

3. The Agent-based Reference Scenario

In this section, we describe the agent-based reference scenario which we will refer to in the following. This scenario involves a potentially large number of agents, which can mutually interact among them, on behalf of their associated devices. Within this agent community, we assume that the interactions carried out by agents satisfy the following desirable properties [52]:

- agents are long living entities so that past behaviors will provide information about expected, future behaviors;
- new agent interactions are driven only by the counterpart's past behaviors;
- agents' reputation scores are spread into the community.

In particular, we consider rational agents, i.e. artificial intelligence software developed to make autonomous and rational choices on the basis of a system of rules, knowledge and data available. The actions to be taken are chosen on the basis of the information collected, previous knowledge and basic knowledge available to the agent.

The rational agent is composed of interacting elements and is equipped with special devices, such as sensors or actuators, able to (*a*) capture information from the surrounding environment and (*b*) intervene to modify it.

In our reference scenario, the information coming from the external environment perceived by the agents can be the reputations of the individual agents belonging to the whole community. As regards the methods of intervention, we refer to the possibility of attributing a high or low trust value based on one's experience with the individual agent.

With the development of Multi-Agent Systems, several organizational paradigms have been developed. These organizations establish a framework for relationships and interactions between agents. The community we consider is a collection of various agents who interact and communicate. They have different goals, they do not have the same level of rationality, nor the same skills, but they are all subject to common laws.

Among the agents of the community there are also the malicious agents (colluding): they gang up and collaborate as their individual interests are shared. Their goal is to maximize the interests of the whole coalition at the expense of other agents.

4. The Reputation Model

This section describes in the detail the reputation model we designed to detect colluding agents by preserving the reputation scores of honest agents, wherever other well known approaches decrease them as a collateral effect of their computational processes.

To this end, let $\mathcal{A}$ an agent community of n agents, where each pair of agents is uniquely identified as a_i and a_j, with $i \neq j \in [1, n]$, and let τ_{ij} be a real number ranging in $[0; 1]$ that represents the *trust* perceived by a_j about a_i, and let $\tau_{ii} = 0$ be the *trust* of an agent about itself for all the agents in $\mathcal{A}$.

In $\mathcal{A}$ the *reputation* ρ_i of the generic agent a_i (i.e. the trustee) is computed as the ratio between the sum of the trust values ρ_{ij} perceived by the other agents of $\mathcal{A}$ (i.e. the trustors) about a_i, with $j = 1, \ldots, n$ and $a_j \neq a_i$, weighted by their own reputation scores, and the sum of the reputation scores of all the trustor agents. More formally, ρ_i is computed as:

$$\rho_i = \frac{\displaystyle\sum_{j=1}^{n} \tau_{ij}\ \rho_j}{\displaystyle\sum_{j=1}^{n} \rho_j}, \quad \text{with} \quad i = 1, \ldots, n \tag{1}$$

To represent the agents' reputation in $\mathcal{A}$, we define the trust matrix $\mathbf{T} = [\tau_{ij}]$ as:

$$\mathbf{T} = \begin{pmatrix} \tau_{11} & \tau_{12} & \cdots & \tau_{1n} \\ \tau_{21} & \tau_{22} & \cdots & \tau_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{n1} & \tau_{n2} & \cdots & \tau_{nn} \end{pmatrix}.$$

By assuming $\mathbf{T}$ as the transpose of the weighted adjacency matrix $\mathbf{A} = [A_{ij}]$ which, in turn, corresponds to the directed graph $\mathcal{G}$, where each node is associated with an agent and each link (i, j) is associated with a non-negative value representing the trust value perceived by a_i about a_j, then we can reformulate the equations (1) as:

$$\mathbf{T}\rho = \mathbf{R}, \quad \|\mathbf{R}\|_1 = 1, \tag{2}$$

where the i-th element of the reputation vector $\mathbf{R} = (\rho_1, \ldots, \rho_n)^T$ is the reputation of the agent a_i. Note that $\|\mathbf{R}\|_1$ is the sum of the absolute values of the elements of $\mathbf{R}$ and its closure to 1 warranties to (2) the uniqueness of the solution. Besides, requiring that the sum of the trust value τ_{ij}, given by each agent a_j to the other n agents of $\mathcal{A}$, be 1, then the matrix $\mathbf{T}$ will be column-stochastic. The solution of the eigensystem problem (2) can be reformulated as the computation of the stationary distribution for a Markov chain represented by the matrix $\mathbf{T}$.

For the Perron Frobenius Theorem, $\lambda = 1 = \varphi(\mathbf{T})$ is an eigenvalue of $\mathbf{T}$ (the other eigenvalues will be < 1 in modulus) and if $T_{ij} > 0$, then a unique vector $\mathbf{R} \in \mathbb{R}$, $\|\mathbf{R}\|_1 = 1$ there exists, such that $\mathbf{T}\ \mathbf{R} = \varphi(\mathbf{A})\ \mathbf{R} = \mathbf{R}$, where $\varphi(\mathbf{A})$ is the spectral radius that is a unique positive reputation.

A modified version of the eigensystem (2) is the PageRank model that can be formulated as:

$$\big(\omega\, \mathbf{T} + (1-\omega)\, \mathbf{V}\, \mathbf{U}^T\big)\, \mathbf{R} = \mathbf{R} \tag{3}$$

where the parameter $\omega \in \mathbb{R}$ ranges in $0 \leq \omega \leq 1$, $\mathbf{U}$ is a unitary vector and $\mathbf{V}$ (generally named *teleportation vector*) is a non-negative vector with unitary 1-norm, i.e. $\mathbf{U}^T\ \mathbf{V} = 1$. If $\omega \neq 0, 1$, the solution (3) exists, and it is unique.

In the PageRank algorithm all the elements of $\mathbf{V}$ are set to $v = \frac{1}{n}u$. In [53], to decrease the reputation of malicious agents it is proposed to introduce some agents, whose opinions are considered highly reliable. Let $\mathcal{M}$ be the set of such mentor agents so that $v_i = 1/|\mathcal{M}|$ if the agent belongs to $\mathcal{M}$ and 0 otherwise.

To detect colluding agents, the vector $\mathbf{R}$ is computed from the matrix $\mathbf{T}$.
To this aim, the agents a_i and a_j are considered as malicious when the following three conditions on the trust scores are verified at the same time:

(*i*) τ_{ij} and τ_{ji} are high;

(*ii*) τ_{ij} and τ_{ji} are similar;

(*iii*) the sum of the remaining trust scores from rows i and j, respectively associated with a_i and a_j, is low.

In other words, we consider all those cases where the trust values of a_i and a_j are high, while the majority of the other agents of $\mathcal{A}$ consider them untrustworthy.

Therefore, let $\alpha, \beta, \gamma \in \mathbb{R} \in [0, 1]$, with $i = 1, 2, 3$, be three thresholds suitably set to detect colluding agents and let $\mathbf{K}$ be a vector representing the output degree of the nodes of the graph $\mathcal{G}$ or, analogously, the vector of the sums of the rows of $\mathbf{T}$, i.e. $\mathbf{K} = \mathbf{T}\ \mathbf{U}$. Besides, let $\overline{\mathbf{Z}}$ be an auxiliary matrix built from $\mathbf{T}$ by identifying the quasi-symmetric high-valued elements as follows:

$$\tilde{z}_{ij} = \begin{cases} \tau_{ij} & \text{if } |\tau_{ij} - \tau_{ji}| \leq \beta \text{ and } \tau_{ij} \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

The matrix $\overline{\mathbf{Z}}$ corresponds to a weighted adjacency matrix of the (undirected) sub-graph of $\mathcal{G}$ whose arcs connect potential colluding agents.
The real colluding agents are present among these potential collusive agents and to identify them correctly, we use the third threshold γ. In order to do this, we first build the vector $\overline{\mathbf{K}} = \mathbf{K} - \overline{\mathbf{Z}}\ \mathbf{U}$. Finally, based on the vector $\overline{\mathbf{K}}$, we classify as malicious an agent a_i belonging to $\mathcal{A}$ if $\overline{k}_i \leq \gamma$, where $\overline{k}_i$ is the sum of the remaining trust scores from row i associated with a_i, without affecting reputation scores of honest agents.
If we don't consider the third threshold, we could obtain as malicious colluding agents some honest agents, in particular the good honest agents.

5. Colluding Agent Detection

The knowledge of supposed malicious agents can be exploited to provide appropriate reputation vectors.

For instance, let $\mathcal{B}$ be the index set of the malicious agents previously identified and let $\mathbf{w}$ be the definitive weighted adjacency matrix of the subgraph of $\mathcal{G}$ consisting of all the both

colluded agents and edges connecting them.
We can proceed with two different approaches.

For the approach A, consider (3) and a vector $\mathbf{L}$, where its elements l_i are set to 0 if $i \in \mathcal{B}$ or $1/(n - |\mathcal{B}|)$ otherwise. By adopting this setting all the honest agents will receive the same trust score regardless from their starting trust score. This is exactly the same algorithm as EigenTrust, however we exploit here the additional information about the pre-trusted mentors.

The last approach B, consists of building a new matrix $\overline{\mathbf{T}}$ as follows:

- Set a threshold $\eta > 0$;
- make the $\overline{\mathbf{T}}$ columns stochastic by normalizing each of them to 1.
- For $a_i, a_j \in \mathcal{B}$ then let $\overline{\tau_{ij}} = \eta$, otherwise $\overline{\tau_{ij}} = \tau_{ij}$;

In the following, we present a simple example by considering an agent community formed by twelve agents that is represented in Fig. 1. Moreover, we assume that suspected malicious agents correspond to 30 percent of the total agents and they are identified as $4, 5, 6$ and 7.

For simplicity, we choose null the trust that an agent assigns to himself, as it is irrelevant.

The corresponding trust matrix is:

$$\mathbf{T} = \begin{pmatrix} 0 & 0.44 & 0.39 & 0.001 & 0.002 & 0.003 & 0.001 & 0.3 & 0.29 & 0.32 & 0.3 & 0.29 \\ 0.382 & 0 & 0.43 & 0.001 & 0.001 & 0.001 & 0.002 & 0.28 & 0.31 & 0.29 & 0.31 & 0.29 \\ 0.46 & 0.44 & 0 & 0.001 & 0.002 & 0.001 & 0.002 & 0.29 & 0.3 & 0.3 & 0.28 & 0.32 \\ 0.001 & 0.002 & 0.004 & 0 & 0.91 & 0.005 & 0.001 & 0.001 & 0.002 & 0.004 & 0.002 & 0.001 \\ 0.002 & 0.003 & 0.002 & 0.89 & 0 & 0.002 & 0.001 & 0.003 & 0.003 & 0.001 & 0.002 & 0.004 \\ 0.003 & 0.001 & 0.001 & 0.001 & 0.005 & 0 & 0.93 & 0.004 & 0.001 & 0.003 & 0.003 & 0.003 \\ 0.002 & 0.001 & 0.003 & 0.002 & 0.001 & 0.92 & 0 & 0.002 & 0.004 & 0.002 & 0.003 & 0.002 \\ 0.01 & 0.06 & 0.03 & 0.003 & 0.009 & 0.01 & 0.02 & 0 & 0.02 & 0.03 & 0.04 & 0.02 \\ 0.02 & 0.01 & 0.04 & 0.011 & 0.01 & 0.018 & 0.01 & 0.02 & 0 & 0.01 & 0.03 & 0.02 \\ 0.03 & 0.013 & 0.03 & 0.03 & 0.02 & 0.01 & 0.01 & 0.03 & 0.03 & 0 & 0.02 & 0.02 \\ 0.04 & 0.02 & 0.03 & 0.03 & 0.03 & 0.02 & 0.02 & 0.04 & 0.02 & 0.01 & 0 & 0.03 \\ 0.05 & 0.01 & 0.04 & 0.03 & 0.01 & 0.01 & 0.003 & 0.02 & 0.02 & 0.03 & 0.01 & 0 \end{pmatrix}.$$

Determined the threshold α and β, we can build the following matrix

$$\overline{\mathbf{Z}} = \begin{pmatrix} 0 & 0.44 & 0.39 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.382 & 0 & 0.43 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.46 & 0.44 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.91 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.89 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.93 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.92 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

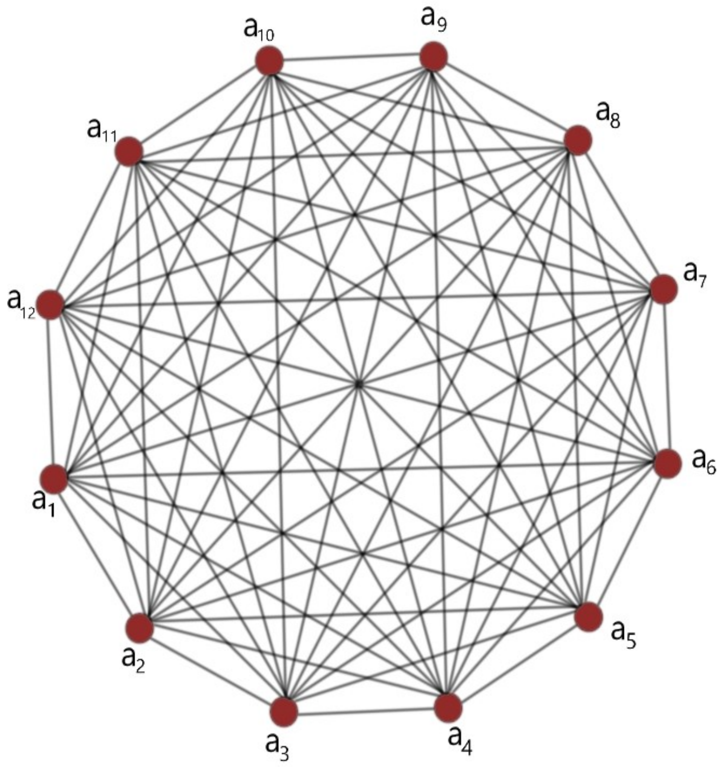

Figure 1: An example of IoT community formed by 12 IoT devices (i.e. agents).

Remember that the matrix $\overline{\mathbf{Z}}$ corresponds to a weighted adjacency matrix of the (undirected) subgraph of $\mathcal{G}$ whose arcs connect potential collusive agents.

With an appropriate choice of γ , $\mathcal{B} = \{a_4, a_5, a_6, a_7\}$ is obtained.

Then, applying the approach A, set the parameter ω to 0.2 and considering $\mathbf{L} = (1/8, 1/8, 1/8, 0, 0, 0, 0, 1/8, 1/8, 1/8, 1/8, 1/8)^T$, the updated trust matrix is computed as:

$$\tilde{\mathbf{T}}_{\mathbf{A}} = \begin{pmatrix} 0.1000 & 0.1880 & 0.1780 & 0.1002 & 0.1004 & 0.1006 & 0.1002 & 0.1600 & 0.1580 & 0.1640 & 0.1600 & 0.1580 \\ 0.1764 & 0.1000 & 0.1860 & 0.1002 & 0.1002 & 0.1002 & 0.1004 & 0.1560 & 0.1620 & 0.1580 & 0.1620 & 0.1580 \\ 0.1920 & 0.1880 & 0.1000 & 0.1002 & 0.1004 & 0.1002 & 0.1004 & 0.1580 & 0.1600 & 0.1600 & 0.1560 & 0.1640 \\ 0.0002 & 0.0004 & 0.0008 & 0 & 0.1820 & 0.0010 & 0.0002 & 0.0002 & 0.0004 & 0.0008 & 0.0004 & 0.0002 \\ 0.0004 & 0.0006 & 0.0004 & 0.1780 & 0 & 0.0004 & 0.0002 & 0.0006 & 0.0006 & 0.0002 & 0.0004 & 0.0008 \\ 0.0006 & 0.0002 & 0.0002 & 0.0002 & 0.0010 & 0 & 0.1860 & 0.0008 & 0.0002 & 0.0006 & 0.0006 & 0.0006 \\ 0.0004 & 0.0002 & 0.0006 & 0.0004 & 0.0002 & 0.1840 & 0 & 0.0004 & 0.0008 & 0.0004 & 0.0006 & 0.0004 \\ 0.1020 & 0.1120 & 0.1060 & 0.1006 & 0.1018 & 0.1020 & 0.1040 & 0.1000 & 0.1040 & 0.1060 & 0.1080 & 0.1040 \\ 0.1040 & 0.1020 & 0.1080 & 0.1022 & 0.1020 & 0.1036 & 0.1020 & 0.1040 & 0.1000 & 0.1020 & 0.1060 & 0.1040 \\ 0.1060 & 0.1026 & 0.1060 & 0.1060 & 0.1040 & 0.1020 & 0.1020 & 0.1060 & 0.1060 & 0.1000 & 0.1040 & 0.1040 \\ 0.1080 & 0.1040 & 0.1060 & 0.1060 & 0.1060 & 0.1040 & 0.1040 & 0.1080 & 0.1040 & 0.1020 & 0.1000 & 0.1060 \\ 0.1100 & 0.1020 & 0.1080 & 0.1060 & 0.1020 & 0.1020 & 0.1006 & 0.1040 & 0.1040 & 0.1060 & 0.1020 & 0.1000 \end{pmatrix}.$$

Finally, by adopting the threshold strategy (approach B) the trust matrix, in correspondence

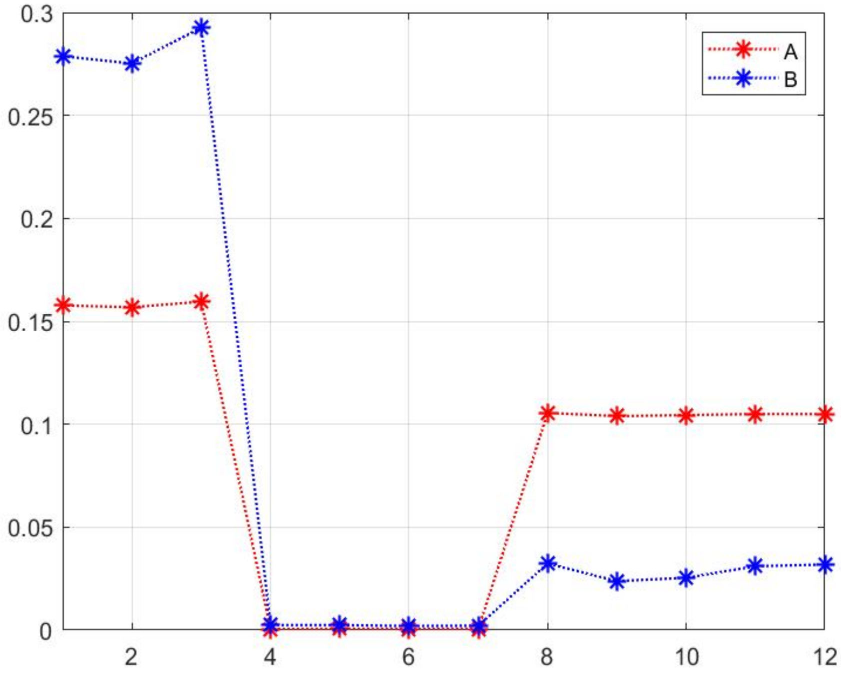

Figure 2: The reputation vectors obtained with the approaches A and B

of $\eta = 0.00016$, will be:

$$\tilde{\mathbf{T}}_{\mathbf{B}} = \begin{pmatrix} 0 & 0.4400 & 0.3900 & 0.0091 & 0.0222 & 0.0374 & 0.0143 & 0.3030 & 0.2900 & 0.3200 & 0.3000 & 0.2900 \\ 0.3820 & 0 & 0.4300 & 0.0091 & 0.0111 & 0.0125 & 0.0285 & 0.2828 & 0.3100 & 0.2900 & 0.3100 & 0.2900 \\ 0.4600 & 0.4400 & 0 & 0.0091 & 0.0222 & 0.0125 & 0.0285 & 0.2929 & 0.3000 & 0.3000 & 0.2800 & 0.3200 \\ 0.0010 & 0.0020 & 0.0040 & 0 & 0.0018 & 0.0624 & 0.0143 & 0.0010 & 0.0020 & 0.0040 & 0.0020 & 0.0010 \\ 0.0020 & 0.0030 & 0.0020 & 0.0015 & 0 & 0.0250 & 0.0143 & 0.0030 & 0.0030 & 0.0010 & 0.0020 & 0.0040 \\ 0.0030 & 0.0010 & 0.0010 & 0.0091 & 0.0555 & 0 & 0.0023 & 0.0040 & 0.0010 & 0.0030 & 0.0030 & 0.0030 \\ 0.0020 & 0.0010 & 0.0030 & 0.0182 & 0.0111 & 0.0020 & 0 & 0.0020 & 0.0040 & 0.0020 & 0.0030 & 0.0020 \\ 0.0100 & 0.0600 & 0.0300 & 0.0272 & 0.0998 & 0.1248 & 0.2851 & 0 & 0.0200 & 0.0300 & 0.0400 & 0.0200 \\ 0.0200 & 0.0100 & 0.0400 & 0.0999 & 0.1109 & 0.2246 & 0.1425 & 0.0202 & 0 & 0.0100 & 0.0300 & 0.0200 \\ 0.0300 & 0.0130 & 0.0300 & 0.2723 & 0.2218 & 0.1248 & 0.1425 & 0.0303 & 0.0300 & 0 & 0.0200 & 0.0200 \\ 0.0400 & 0.0200 & 0.0300 & 0.2723 & 0.3327 & 0.2495 & 0.2851 & 0.0404 & 0.0200 & 0.0100 & 0 & 0.0300 \\ 0.0500 & 0.0100 & 0.0400 & 0.2723 & 0.1109 & 0.1248 & 0.0428 & 0.0202 & 0.0200 & 0.0300 & 0.0100 & 0 \end{pmatrix}.$$

Note that with approach B only the columns of the matrix T corresponding to the colluding agents are modified, leaving the columns of the honest agents unchanged. While, with approach A, all columns of the matrix T are modified and this implies that the final reputation obtained of the agents will be different based on the approach used. In Fig. 2 the reputation vectors corresponding to cases A and B are plotted.

Indeed, as concern honest agents, the approach A provides very similar reputation, while our approach B maintains some difference between the reputations of honest good agents and low honest agents, i.e. it does not penalize the reputation scores of good honest agents.

6. Conclusions

By interacting with the world around them, agents can form multidimensional, relationship based networks potentially rich of social interactions. In a such agent scenario, the presence of malicious actors should be considered in order to minimize the risks of their deception. Such risks can be increased, for instance in presence of large communities and/or nomadic agents, and their minimization is of primary importance to promote satisfactory social agent interactions. To this purpose, an effective solution is to provide each player with appropriate trust measures about potential its partners and, in this respect, we presented: *i*) a method to preliminarily identify the best candidates as malicious (colluding) in order to use such agents as pre-trusted mentors; (*ii*) a novel reputation model to detect colluding malicious agents that does not penalize reputation scores of honest agents in detecting malicious ones.

Our forthcoming researches will be focused on realizing a more complete campaign of experiments on real and simulated data including also different malicious behaviors. In particular, it will be of our interest to use a new metric that identifies the colluding agents of the network, analyzing the starting trust matrix T. An interesting feature that real networks present is the property of clustering or community structure, according to which the graph is called community or cluster. The peculiarity is that the nodes of the same community (honest agents or colluding agents) are very similar while, on the contrary, the nodes between the communities have a low similarity. Therefore, we will start to analyze the fundamental concepts and the methodological bases on which the graph clustering algorithms are based. Later, we will try to identify potential colluding groups, through new metrics and strategies. Finally, we will evaluate the property of a good cluster in a direct graph.

References

[1] C. Castelfranchi, F. D. Rosis, R. Falcone, S. Pizzutilo, Personality traits and social attitudes in multiagent cooperation, Applied Artificial Intelligence 12 (1998) 649–675.

[2] R. Hortensius, F. Hekele, E. S. Cross, The perception of emotion in artificial agents, IEEE Transactions on Cognitive and Developmental Systems 10 (2018) 852–864.

[3] M. Rheu, J. Y. Shin, W. Peng, J. Huh-Yoo, Systematic review: Trust-building factors and implications for conversational agent design, International Journal of Human–Computer Interaction 37 (2021) 81–96.

[4] L. Fotia, F. Messina, D. Rosaci, G. M. L. Sarné, Using local trust for forming cohesive social structures in virtual communities, The Computer Journal 60 (2017) 1717–1727.

[5] C. Misselhorn, Collective agency and cooperation in natural and artificial systems, in: Collective agency and cooperation in natural and artificial systems, Springer, 2015, pp. 3–24.

[6] O. Perrin, C. Godart, A model to support collaborative work in virtual enterprises, Data & Knowledge Engineering 50 (2004) 63–86.

[7] M. Uhl-Bien, R. Marion, B. McKelvey, Complexity leadership theory: Shifting leadership from the industrial age to the knowledge era, The leadership quarterly 18 (2007) 298–318.

[8] F. Amin, A. Ahmad, G. Sang Choi, Towards trust and friendliness approaches in the social internet of things, Applied Sciences 9 (2019) 166.

[9] G. Fortino, F. Messina, D. Rosaci, G. M. L. Sarné, Using blockchain in a reputation-based model for grouping agents in the internet of things, IEEE Transactions on Engineering Management 67 (2019) 1231–1243.

[10] D. Gambetta, et al., Can we trust trust, Trust: Making and breaking cooperative relations 13 (2000) 213–237.

[11] D. H. McKnight, N. L. Chervany, The meanings of trust (1996).

[12] A. Altaf, H. Abbas, F. Iqbal, A. Derhab, Trust models of internet of smart things: A survey, open issues, and future directions, Journal of Network and Computer Applications 137 (2019) 93–111.

[13] G. Fortino, L. Fotia, F. Messina, D. Rosaci, G. M. L. Sarné, Trust and reputation in the internet of things: state-of-the-art and research challenges, IEEE Access 8 (2020) 60117–60125.

[14] Z. Yan, P. Zhang, A. V. Vasilakos, A survey on trust management for internet of things, Journal of network and computer applications 42 (2014) 120–134.

[15] M. N. Postorino, G. M. L. Sarné, An agent-based sensor grid to monitor urban traffic, in: Proceedings of the 15th Workshop dagli Oggetti agli Agenti, WOA 2014, volume 1260 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014.

[16] S. D. Kamvar, M. T. Schlosser, H. Garcia-Molina, The eigentrust algorithm for reputation management in p2p networks, in: Proc. of the 12th international conference on World Wide Web, ACM, 2003, pp. 640–651.

[17] F. Messina, G. Pappalardo, D. Rosaci, C. Santoro, G. M. L. Sarné, A trust model for competitive cloud federations, in: 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems, IEEE, 2014, pp. 469–474.

[18] A. Ahmed, K. A. Bakar, M. I. Channa, K. Haseeb, A. W. Khan, A survey on trust based detection and isolation of malicious nodes in ad-hoc and sensor networks, Frontiers of Computer Science 9 (2015) 280–296.

[19] P. De Meo, F. Messina, M. N. Postorino, D. Rosaci, G. M. L. Sarné, A reputation framework to share resources into iot-based environments, in: IEEE 14th Int. Conf. on Networking, Sensing and Control, IEEE, 2017, pp. 513–518.

[20] G. Fortino, F. Messina, D. Rosaci, G. M. L. Sarné, Resiot: An iot social framework resilient to malicious activities, IEEE/CAA Journal of Automatica Sinica 7 (2020) 1263–1278.

[21] H. Jnanamurthy, S. Singh, Detection and filtering of collaborative malicious users in reputation system using quality repository approach, in: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2013, pp. 466–471.

[22] S. M. Sajjad, S. H. Bouk, M. Yousaf, Neighbor node trust based intrusion detection system for wsn, Procedia Computer Science 63 (2015) 183–188.

[23] A. J. Bidgoly, B. T. Ladani, Benchmarking reputation systems: A quantitative verification approach, Computers in Human Behavior 57 (2016) 274–291.

[24] W. Fang, W. Zhang, W. Chen, T. Pan, Y. Ni, Y. Yang, Trust-based attack and defense in wireless sensor networks: a survey, Wireless Communications and Mobile Computing

2020 (2020).

[25] K. Hoffman, D. Zage, C. Nita-Rotaru, A survey of attack and defense techniques for reputation systems, ACM Computing Surveys (CSUR) 42 (2009) 1–31.

[26] F. G. Mármol, G. M. Pérez, Security threats scenarios in trust and reputation models for distributed systems, computers & security 28 (2009) 545–556.

[27] A. Jøsang, Robustness of trust and reputation systems: Does it matter?, in: IFIP International Conference on Trust Management, Springer, 2012, pp. 253–262.

[28] D. Rosaci, G. M. L. Sarnè, S. Garruzzo, Integrating trust measures in multiagent systems, International Journal of Intelligent Systems 27 (2012) 1–15.

[29] F. Hendrikx, K. Bubendorfer, R. Chard, Reputation systems: A survey and taxonomy, Journal of Parallel and Distributed Computing 75 (2015) 184–197.

[30] P. De Meo, F. Messina, D. Rosaci, G. M. L. Sarné, An agent-oriented, trust-aware approach to improve the qos in dynamic grid federations, Concurrency and Computation: Practice and Experience 27 (2015) 5411–5435.

[31] P. De Meo, L. Fotia, F. Messina, D. Rosaci, G. M. L. Sarné, Providing recommendations in social networks by integrating local and global reputation, Information Systems 78 (2018) 58–67.

[32] A. Jøsang, R. Ismail, C. Boyd, A survey of trust and reputation systems for online service provision, Decision support systems 43 (2007) 618–644.

[33] A. Sharma, E. S. Pilli, A. P. Mazumdar, P. Gera, Towards trustworthy internet of things: A survey on trust management applications and schemes, Computer Communications (2020).

[34] A. Jøsang, J. Golbeck, Challenges for robust trust and reputation systems, in: Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France, volume 5, Citeseer, 2009.

[35] S. Vavilis, M. Petković, N. Zannone, A reference model for reputation systems, Decision Support Systems 61 (2014) 147–154.

[36] G. Lax, G. M. L. Sarné, CellTrust: a reputation model for C2C commerce, Electronic Commerce Research 8 (2006) 193–216.

[37] Y.-F. Wang, Y. Hori, K. Sakurai, Characterizing economic and social properties of trust and reputation systems in p2p environment, Journal of Computer Science and Technology 23 (2008) 129–140.

[38] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, M. Voss, A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies, in: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, 2005, pp. 512–518.

[39] A. A. Adamopoulou, A. L. Symeonidis, A simulation testbed for analyzing trust and reputation mechanisms in unreliable online markets, Electronic Commerce Research and Applications 13 (2014) 368–386.

[40] R. Kerr, R. Cohen, Treet: The trust and reputation experimentation and evaluation testbed, Electronic Commerce Research 10 (2010) 271–290.

[41] F. G. Mármol, G. M. Pérez, Trmsim-wsn, trust and reputation models simulator for wireless sensor networks, in: 2009 IEEE International Conference on Communications, IEEE, 2009,

pp. 1–5.

[42] A. J. Bidgoly, B. T. Ladani, Modelling and quantitative verification of reputation systems against malicious attackers, The Computer Journal 58 (2015) 2567–2582.

[43] S. A. Ghasempouri, B. T. Ladani, Modeling trust and reputation systems in hostile environments, Future Generation Computer Systems 99 (2019) 571–592.

[44] S. C. Hayne, H. Wang, L. Wang, Modeling reputation as a time-series: Evaluating the risk of purchase decisions on ebay, Decision Sciences 46 (2015) 1077–1107.

[45] L. Cabral, A. Hortacsu, The dynamics of seller reputation: Evidence from ebay, The Journal of Industrial Economics 58 (2010) 54–78.

[46] P. Resnick, R. Zeckhauser, Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system, in: The Economics of the Internet and E-commerce, Emerald Group Publishing Limited, 2002.

[47] L. Xiong, L. Liu, Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities, IEEE transactions on Knowledge and Data Engineering 16 (2004) 843–857.

[48] F. Messina, G. Pappalardo, D. Rosaci, C. Santoro, G. M. L. Sarné, A trust-aware, self-organizing system for large-scale federations of utility computing infrastructures, Future Generation Computer Systems (2015).

[49] W. Abdelghani, C. A. Zayani, I. Amous, F. Sèdes, Trust management in social internet of things: a survey, in: Conference on e-Business, e-Services and e-Society, Springer, 2016, pp. 430–441.

[50] I. U. Din, M. Guizani, B.-S. Kim, S. Hassan, M. K. Khan, Trust management techniques for the internet of things: A survey, IEEE Access 7 (2018) 29763–29787.

[51] C. Marche, M. Nitti, Trust-related attacks and their detection: a trust management model for the social iot, IEEE Transactions on Network and Service Management (2020).

[52] P. Resnick, R. Zeckhauser, F. E., K. Kuwabara, Reputation systems, Communication of ACM 43 (2000) 45–48.

[53] S. Kamvar, M. Schlosser, H. Garcia-Molina, The eigentrust algorithm for reputation management in P2P networks, in: Proc. of World Wide Web, 12th International Conference on, ACM, 2003, pp. 640–651.