Trust Computing Based on Argumentation Debates with Votes for Detecting Lying Agents

Jeferson José Baqueta[1], **Mariela Morveli-Espinoza**[1], **Cesar A. Tacla**[1]

[1]Programa de Pós-Graduação em Engenharia Elétrica e Informática Industrial
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 - Rebouças - CEP 80230-901 - Curitiba - PR

jefersonbaqueta@gmail.com, morveli.espinoza@gmail.com, tacla@utfpr.edu.br

Abstract. *In a multi-agent system (MAS), it is very usual that agents delegate tasks to each other. However, due to the subjectivity of the information used by agents during the decision-making process, an agent may end up delegating a task to an untrustworthy partner. In this work, we present a trust computing approach based on quantitative argumentation with votes (QuAD-V), where a trust measure is estimated according with the agents' opinions about the service provided by a partner. Besides, such an approach provides a mechanism to evaluate the credibility of agents that play as information sources. In our results, we demonstrate how our trust computing approach can be employed for detecting lying agents, which are able to slander or promote other agents.*

Resumo. *Em um sistema multi-agente (SMA), é muito comum que os agentes deleguem tarefas uns aos outros. Contudo, devido à subjetividade das informações utilizadas pelos agentes durante o processo de tomada de decisão, um agente pode acabar delegando uma tarefa a um parceiro não confiável. Neste trabalho, apresentamos uma abordagem de cálculo de confiança baseada em argumentação quantitativa com votação (QuAD-V), onde a confiança é estimada de acordo com a opinião dos agentes sobre o serviço prestado por um parceiro. Além disso, tal abordagem fornece um mecanismo para avaliar a credibilidade dos agentes que atuam como fontes de informação. Como resultado, demonstramos como nossa abordagem de cálculo de confiança pode ser empregada para detectar agentes mentirosos capazes de caluniar ou promover outros agentes.*

1. Introduction

The design of artificial agents capable of operating together is a great challenge [Singh 2018]. Predicting the behavior and intentions of other agents is a delicate task since the information collected by agents for this end may be incomplete, inconsistent, or even uncertain [Castelfranchi and Falcone 1998]. For instance, in a society where agents are able to lie (*e.g.*, slandering, self-promoting, or promoting other agents), there are no guarantees that a piece of information circulating in the society is, in fact, true [Buccafurri et al. 2015]. Thus, in this case, an agent can receive and share a rumor, created by an unknown source to harm someone.

In this context, social control mechanisms, such as the reputation and trust models, have been used as a way to provide security and efficiency in multi-agent

systems (MAS) [Conte and Paolucci 2002] [Sabater and Sierra 2001] [Griffiths 2005] [Sabater et al. 2006] [Castelfranchi and Falcone 2010] [Buccafurri et al. 2015]. These models allow the agents to be evaluated (*e.g.*, good or bad) as they interact with each other. In this sense, an agent can punish undesirable behavior of other agents, for example, by not selecting a certain partner for a given task [Pinyol and Sabater-Mir 2013].

On the other hand, several of these models consider that the partner's attributes combined with the task requirements are enough for ensuring the selection of good partners [Sabater and Sierra 2001] [Griffiths 2005] [Buccafurri et al. 2015]. However, in a scenario where an agent builds his opinion about a partner based on the opinions received from third parties, verifying the credibility of the agents that act as sources of information (*i.e.*, agents that share their opinions about a partner with other agents in the society), may prevent the propagation of lies and reduce the chances of an agent relies on someone untrustworthy.

Therefore, in this work, we propose a trust computing approach focused on task delegation scenarios, where an agent (*trustor*) needs to delegate a task to another agent (*trustee*) to achieve his goals. This approach employs some elements from reputation theory proposed by [Conte and Paolucci 2002], as social image, shared evaluations, and reputation for building and propagating the agents' opinions. Whereas, the trust in a given partner is calculated based on a quantitative argumentation framework for debates with votes (QuAD-V) [Rago and Toni 2017]. Moreover, the adoption of social image and shared evaluations allows the agents to compare their opinions in order to estimate the credibility of a source of information. As presented in our results, this credibility validation allows the agents to ignore the information shared by lying agents, which are able to slander and promoting other agents.

In particular, as discussed in [Dung 1995], an argumentation framework can be defined in terms of arguments and the relations established between them. For instance, in the case of the bipolar argumentation frameworks (BAFs) [Cayrol and Lagasquie-Schiex 2005], the dialectical exchanges established between arguments are represented through the attack and support relations. In turn, the QuAD-V is a type of bipolar framework based on IBIS methodology [Kunz and Rittel 1970], where the arguments are divided into answer, pro, and con arguments. Moreover, in a QuAD-V framework, a strength is assigned to arguments based on a voting process [Rago and Toni 2017]. In the QuAD-V framework adopted in this work, the agents can vote for or against the pro and con arguments that better express their satisfaction with certain aspects of a partner's behavior.

The rest of this paper is organized as follows. Section 2 presents some basic concepts and definitions adopted in this work, as the mechanism of social evaluation employed and the details of the QuAD-V framework. Section 3 presents the proposed trust computing approach, highlighting the details of the trust assessment performed by agents. Section 4 describes the case study and some considerations of our implementation. Section 5 exhibits the experimental results. The conclusions, discussions about the obtained results, and future works are summarized in Section 6.

2. Basic Concepts

This section reviews some of the main definitions and concepts regarding the social mechanisms used by agents to delegate a task. Herein, we also discuss and present in detail the quantitative argumentation with votes (QuAD-V) framework used to compute a trust measure during a task delegation situation.

2.1. Trust

As discussed in [Cho et al. 2015], trust is a multidisciplinary concept, which has been used in different disciplines to model different types of relationships. In this work, we adopt the trust definition suggested by [Castelfranchi and Guerini 2007], where trust is defined through five components, which are represented by the *5-tuple* TRUST(X, Y, C, τ, G_x). This tuple can be read as, *X* (*trustor*) trusts *Y* (*trustee*) in a context *C* for performing an action α (through the task τ) and obtaining as result *p* (the outcome expected by *X*, which corresponds to *X*'s goal ($G_x = g_x$)).

Trust is fundamental for environments where agents must work together to achieve a goal. In this kind of scenario, agents tend to delegate tasks to each other, because many times an agent may not have the capabilities or resources to achieve his goals alone [Griffiths 2005]. Therefore, as discussed in [Castelfranchi and Falcone 2010] and [Solhaug et al. 2007], the act of trusting someone is not a simple activity since it requires the fulfillment of some conditions. Firstly, the *trustor* must perform a preventive evaluation about the characteristics and virtues of all possible partners (*i.e.*, validating the minimum requirements for trusting in a *trustee*). Following, the *trustor* compares the potential partners, considering the risks and costs of delegating the task. In the last stage, the *trustor* must select a partner and delegate the task to him (*i.e.*, establishing a trust relationship with the selected partner, where the *trustor* creates expectations about the fulfillment of the task and starts relying on the *trustee*).

2.2. Social Image and Reputation

As pointed out by [Conte and Paolucci 2003], reputation is a multi-purpose social and cognitive artifact that can be used as a partner selection mechanism. In this sense, the agents themselves are capable of punishing non-desirable behaviors (*e.g.*, not selecting a given partner to a certain task) [Pinyol and Sabater-Mir 2013]. Based on these characteristics, a reputation theory was introduced in [Conte and Paolucci 2002]. In this theory, the authors discuss the differences and the interrelationships between social image and reputation, both mechanisms of social evaluation employed to assess the attitudes of an agent (target) based on his behavior.

To clarify the reputation and social image concepts, in Figure 1 the basic components of the reputation theory are shown. In particular, the social image is defined as a belief produced from the direct experiences of an agent, expressing a personal opinion about a target [Conte and Paolucci 2002]. For instance, in Figure 1, agent *A*, after interacting with agent *B* (target), produces his own social image about *B*'s behavior (*i.e.*, defining whether *B* is *good* or *bad* with respect to a norm, a standard, or a skill [Miceli and Castelfranchi 2000]).

On the other hand, reputation is defined as a meta-belief, since it is produced based on third-party opinions [Pinyol and Sabater-Mir 2013]. Thus, reputation can be

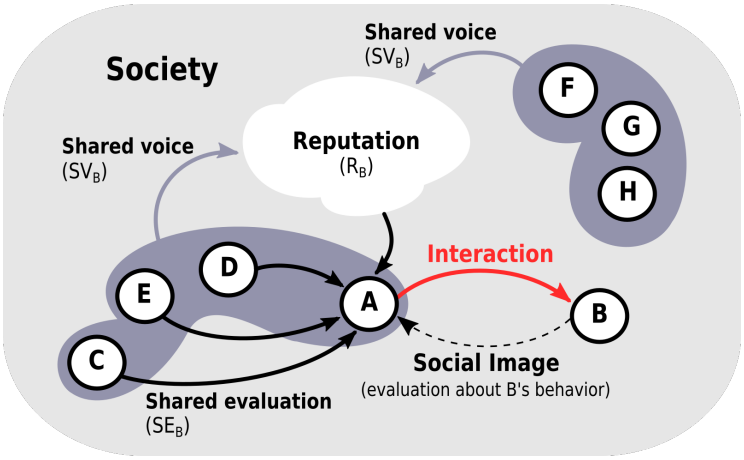

Figure 1. Components of the reputation theory. The interaction between the *A* and *B* rely on the evaluations shared by *C*, *D*, and *E*, concerning *B*'s behavior, and the reputation of *B* that circulates in the society. After his interaction with *B*, *A* can produce his own evaluation (social image) about *B*'s behavior based on the experiences obtained from such an interaction.

seen as a general opinion about something or someone that is shared by the majority of the members of the society. Notice in Figure 1 that the *B*'s reputation is produced through shared voices, which are opinions shared by sets of agents about the behavior of a target, in this particular case, opinions concerning the behavior of the agent *B*.

Additionally, before interacting with *B*, the agent *A* may use the reputation of *B*, which is circulating in the society, and the evaluations shared by other agents (*C*, *D*, and *E*) about the *B*'s behavior to decide whether it is worth interacting with *B*. Note that a shared evaluation is essentially a special case of social image [Sabater and Sierra 2001], which is shared with the agents of a given set.

2.3. The QuAD-V Framework

The QuAD-V framework has been proposed by [Rago and Toni 2017] as an extension of the quantitative argumentation debate framework (QuAD)[Baroni et al. 2015]. Its main advantage is the possibility to solve a debate using a voting system, where a set of users vote for or against arguments. As presented in [Rago and Toni 2017], a QuAD-V is a *6-tuple* (A, C, P, R, U, V), in which *A* is a finite set of answer arguments, *C* is a finite set of con arguments, *P* is a finite set of pro arguments, $R \subseteq (C \cup P) \times (A \cup C \cup P)$ is an acyclic binary relation, *U* is a finite set of users, and $V : U \times (A \cup C \cup P) \rightarrow \{-, ?, +\}$ is a total function, such as $V(u, a)$ is the vote of user $u \in U$ on argument $a \in (A \cup C \cup P)$.

In the QuAD-V framework, the arguments can attack or support one another. The attackers and supporters of an argument are defined based on the con and pro arguments, respectively. Thus, for any argument $a \in (A \cup C \cup P)$, the set of *attackers* of *a* is $R^-(a) = \{b \in C | (b, a) \in R\}$ and the set of *supporters* of *a* is $R^+(a) = \{b \in P | (b, a) \in R\}$. Besides, each argument *a* has a vote base score $\tau_v : A \cup C \cup P \rightarrow \mathbb{I}$ (for scale $\mathbb{I} = [0, 1]$), which is computed according with the users

voting, such as following:

$$\tau_v(a) = \begin{cases} 0.5 & \textit{if } |U| = 0 \\ 0.5 + (0.5 * \frac{N^+(a) - N^-(a)}{|U|}) & \textit{if } |U| \neq 0 \end{cases} \quad (1)$$

where, $N^+(a)$ is a counter that sums the positive votes for *a*, such that, for any argument $a \in (A \cup C \cup P), N^+(a) = |V^+(a)|$, and $V^+(a) = \{u \in U : V(u, a) = +\}$ is the set of users voting for *a*. Whereas, $N^-(a)$ is a counter that sums the negative votes for *a*, such that, for any argument $a \in (A \cup C \cup P), N^-(a) = |V^-(a)|$, and $V^-(a) = \{u \in U : V(u, a) = -\}$ is the set of users voting against *a*.

The Discontinuity-Free QuAD (DF-QuAD) algorithm [Rago et al. 2016] is adopted to aggregate the base score of each argument with the strength of its attackers and supporters. The DF-QuAD is recursive algorithm based on the following strength aggregation function $F : \mathbb{I}^* \rightarrow \mathbb{I}$:

$$F(S) = \begin{cases} 0 & \textit{if } n = 0 \\ v_1 & \textit{if } n = 1 \\ f(v_1, v_2) & \textit{if } n = 2 \\ f(F(v_1, ..., v_{n-1}), v_n) & \textit{if } n > 2 \end{cases} \quad (2)$$

where, $S = (v_1, ..., v_n) \in \mathbb{I}^*$ is an arbitrary permutation of the attackers or supporters. For instance, for any argument $a \in (A \cup C \cup P)$, let $(a_1, ..., a_n)$ be an arbitrary permutation of the attackers in $R^-(a)$, while $(a_1, ..., a_m)$ is an arbitrary permutation of the supporters in $R^+(a)$. On the other hand, the base function $f : \mathbb{I} \times \mathbb{I} \rightarrow \mathbb{I}$ for $v_1, v_2 \in \mathbb{I}$ can be defined as:

$$f(v_1, v_2) = v_1 + v_2 - v_1 * v_2 \quad (3)$$

Finally, after aggregating separately the strengths of the attackers (v^-) and supporters (v^+) of an argument $a \in (A \cup C \cup P)$ through the function F, the final score of a can be obtained using the combination function, defined as $c : \mathbb{I} \times \mathbb{I} \times \mathbb{I} \rightarrow \mathbb{I}$:

$$c(a, v^+, v^-) = \begin{cases} \tau_v(a) - \tau_v(a) * |v^+ - v^-| & \textit{if } v^- \geq v^+ \\ \tau_v(a) + (1 - \tau_v(a)) * |v^+ - v^-| & \textit{if } v^- < v^+ \end{cases} \quad (4)$$

Note that the final score of the arguments is recursively computed based on their attack and support relationships. This process is performed from the leaf arguments up to the answer arguments. In particular, the final score of an argument defines its acceptability degree (strength) (*e.g.*, 1 to be accepted, 0.5 to be neutral, and 0 to be rejected).

3. The Trust Computing Approach

As pointed out by [Cho et al. 2015], the trust in someone can be estimated based on several different elements. In general, a simple way of computing trust is by considering the internal factors, such as the partner's capabilities or the requirements of the task (*e.g.*, the velocity, strength, or expertise of a partner, and the cost, time, or quality of a task) [Griffiths 2005] [Braga et al. 2018] [Buccafurri et al. 2015]. However, as discussed in

[Castelfranchi and Falcone 2010], external factors, as the environmental conditions and the risks associated with the decision to delegate a task to another agent, also should be considered on trust computing.

In this sense, to compute the trust of the *trustor* in a partner, we use a QuAD-V framework where the pro and con arguments represent claims about the competencies and availability of a partner (internal factors), as well as the conditions and risks associated with the task execution (external factors). In particular, these arguments are predefined according with the service provided by the partner. For instance, the choice of a musician (partner) for playing at a party can be justified by a set of pro and con arguments that confirm his abilities as a good musician. Furthermore, the adopted QuAD-V framework has only one answer argument, which is associated with the trustworthiness of a partner. Therefore, the trust measure can be computed by a voting process that determines the final score of the answer argument and consequently whether a partner is or not trustworthy.

Basically, the QuAD-V framework is employed in two distinct situations. In the first of them, after interacting with a partner, the *trustor* performs a personal voting considering the experiences obtained during the interaction. (*i.e.*, the *trustor* produces an own opinion (social image) about the partner's behavior, which is defined as a set of personal votes that expresses his agreement or disagreement with the pro and con arguments in the QuAD-V framework). In the second situation, through collective voting, the *trustor* aggregates the evaluations shared with him, concerning a partner, into a trust measure. In this case, the QuAD-V framework is used by the *trustor* to select a partner based on the personal opinions of other agents. In case of no one directly shares evaluations with the *trustor*, he can use the partner's reputation that circulates in the society to decide whether such a partner is or not trustworthy. In particular, the reputation of a partner is computed by a global voting process where the personal opinions of all agents are aggregated into a single trust measure.

At last, as the agents can share their evaluations about a partner with the *trustor*, which may or not be true evaluations, we implement a simple mechanism to check the credibility of these agents (sources of information). In this case, the *trustor* can validate the credibility of a source of information by comparing the evaluation shared by this source with his social image. As the *trustor*'s social image is produced through an interaction between the *trustor* and a partner, the veracity of the shared evaluation can be confirmed based on the behavior presented by such a partner during his interaction with the *trustor*.

4. Case Study

In our study case, a given agent (*trustor*) needs to have surgery, but he does not have the abilities and resources to perform it alone. Thus, the *trustor* must delegate the surgery (task) to another agent who can carry out it (a doctor). To select a doctor, the *trustor* may use either the evaluation shared with him or the doctor's reputation. After interacting with a doctor, the *trustor* has conditions to evaluate the doctor's behavior based on the provided service and updates his social image about the doctor. In the case where the doctor is selected based on the evaluations of other agents, the credibility of each agent, playing as a source of information, can be validated by comparing the *trustor*'s image to the evaluation shared by the agent.

The QuAD-V framework adopted in our case study is presented in Figure 2. This framework has been modeled in order to represent the internal and external factors of the task delegation scenario. Note that the internal and external factors are represented by con and pro arguments. The set of arguments $\{S_2, S_3, S_5, S_6, S_7, S_8\}$ refers to the capabilities, experiences, and availability of a doctor (internal factors), and the set of arguments $\{S_4, S_9, S_{10}\}$ refers to the risk of having the surgery and the surgery's consequences (external factors).

In the QuAD-V framework, an agent can express his opinion about a doctor by voting for or against the pro and con arguments (social image). For instance, if the agent believes that the doctor is an expert, he must vote for the argument S_2, another way, he must vote against this argument. Once the voting is closed, the DF-QuAD algorithm is executed to compute the final score of arguments. The final score associated with the answer argument S_1 expresses the doctor's trustworthiness. In this case study, a doctor is considered trustworthy case the final score of S_1 is greater than or equal to 0.5.

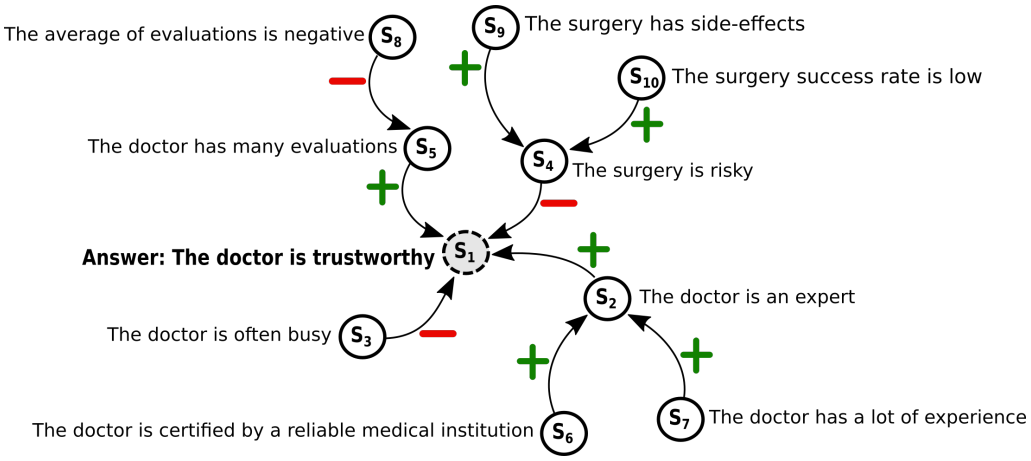

Figure 2. The QuAD-V framework adopted in our case study. The trustworthiness of a doctor is estimated based on a voting process, where the agents express their opinions by voting for or against the pro and con arguments.

4.1. Sharing Evaluations

Every time an agent shares an evaluation about a doctor with a *trustor*, he must decide whether to tell the truth or lie to the *trustor*. When an agent lies, he shares an opposite opinion to his social image (*e.g.*, for any argument $a \in (A \cup C \cup P)$, a vote for a is changed to a vote against a). Nevertheless, when the agent tells the truth, he shares his social image without alteration. Therefore, by lying, the agent may be slandering a doctor by good behavior or promoting a doctor by bad behavior, and at the same time, this lying may lead the *trustor* to delegate a task to an untrustworthy doctor.

4.2. Voting Principles

According to [Rago and Toni 2017], in a QuAD-V framework, the users must vote rationally to ensure the coherence of the voting process. For instance, a user is voting irrationally when agrees with some argument a, agrees with one of its attackers b but does not agree with any of its supporters. Therefore, to ensure the rationality of the voting process employed by the agents, we are assuming that agents are totally for or totally against

a point of view. In the first case, when an agent agrees with an argument *a* (voting for it), he also agrees with all its supporters and disagrees with all its attackers (voting against them). On the other hand, when an agent disagrees with an argument *a*, he also agrees with all its attackers and disagrees with all its supporters.

5. Experiments

In this section, we present the results obtains from the implementation of our case study. In particular, two different scenarios are considered in our experiments. In the first scenario, the *trustor* uses the information about the doctors' reputation to select a partner. In the second scenario, the partner selection is based on the mechanism of shared evaluations. In this particular case, besides computing the trust in a doctor, the *trustor* is able to evaluate the credibility of agents that shared evaluations with him (*i.e.*, the agents that are playing as sources of information).

5.1. Experiments Setup

In our experiments, we are considering a society composed of 100 common agents, one *trustor* agent, and two doctors agents. The common agents play as sources of information, generating and sharing evaluations about the doctors' behaviors with the *trustor*. In turn, the *trustor* agent needs to decide what is the best doctor to perform a surgery. The *trustor* must make his decision based on either the evaluations shared by common agents or the doctor's reputations. Finally, the doctors are service providers, performing surgeries. Moreover, one of the doctors plays as a good doctor, for which the quality of the provided service is high [0.8, 1], whereas another is a bad doctor, for which the quality of provided service is low [0, 0.3]. Values in the range of [0.31, 0.79] are never reached by the framework. It is important to remark that the quality of service provided by a doctor represents just his competencies to perform a surgery, we do not use such value to compute the trust or the reputation of a doctor, even though the quality of the service is directly associated with the social behavior of a doctor. In this work, the reputation and trust measures are estimated based only on the agents' votes.

In each experiment, the agents interact with each other by 11 iterations (i) (*i.e.*, from the iteration 0 up to 10). The number of common agents that interact with the doctors per iteration ($N_{agents}(i)$), consuming the service provided by them, is defined as follows:

$$N_{agents}(i) = 100 * (0.1 * i) \tag{5}$$

On the other hand, the number of common agents that tell lies per iteration ($N_{liars}(i)$) is defined as follows:

$$N_{liars}(i) = N_{agents}(i) * (0.1 * i) \tag{6}$$

In conclusion, in each iteration, the *trustor* selects a doctor based on the trust measure estimated by him. Therefore, if the trust in the good doctor is higher or equal to the trust in the bad doctor, the good doctor is selected as the *trustor*'s partner, another way, the *trustor* selects the bad doctor as his partner.

5.2. Experiment A: Trust Based on Reputation

As discussed in [Sabater et al. 2006], differently from the other social evaluations (*e.g.*, social image, shared evaluations, and shared voices), reputation does not take a stand on what is true but just on what is told, since there is no personal commitment of the speaker concerning the main content of the information delivered. Thus, in a situation where *trustor* decides to trust in the *trustee* considering only information about the *trustee*'s reputation, it is not possible to identify the sources that produced such information, as well as the sources' credibility.

Therefore, as presented in Figure 3 (a), when the trust is computed based on the doctors' reputation, the *trustor* makes no difference between the evaluations produced by honest and lying agents. Moreover, as the number of agents that tell lies increases along the time, when the number of lying agents becomes higher than the number of honest agents, the *trustor* ends up delegating the surgery to the bad doctor (Figure 3 (c), iteration 6). Note that this trust inversion happens because the majority of the members of the society are slandering the good doctor and promoting the bad doctor.

5.3. Experiment B: Trust Based on Shared Evaluations

Every time the *trustor* makes trust decisions considering the shared evaluations, his social image can be used as a validation mechanism to identify reliable sources of information and expose lying agents. Thus, for identifying the liars, the *trustor* compares his social image about a doctor to the evaluations of other agents shared with him. As the social image is produced from the *trustor*'s direct experiences, there is a high chance that it expresses the doctors' real behavior. Consequently, if shared evaluation is contrary to the *trustor*'s social image, possibly such an evaluation is a lie. However, due to the uncertainty involved in this process, it is not prudent to presume that an agent is a liar by just interacting with him once. In general, the *trustor*'s uncertainty about a partner decreases as the number of interactions between this partner and the *trustor* increases [Ashtiani and Azgomi 2014]. Therefore, we assume that the credibility of a source of information (σ_s) rely on the perceived uncertainty by the *trustor*, which can be estimated as follows:

$$\sigma_s = \begin{cases} 1 & \textit{if } \frac{1}{|I^*_{t\to s}|} \leq \phi \vee |I^*_{t\to s}| = 0 \\ 0.5 + (0.5 * \frac{|I^+_{t\to s}| - |I^-_{t\to s}|}{|I^*_{t\to s}|}) & \textit{otherwise} \end{cases} \tag{7}$$

where, ϕ is the uncertainty threshold that determines when the number of interactions between the *trustor* and the source of information is enough to make a correct judgment about the source's credibility, $|I^-_{t\to s}|$ is the number times that the source of information *s* told a lie to the *trustor*, $|I^+_{t\to s}|$ is the number times that the source of information *s* told a truth to the *trustor*, and $|I^*_{t\to s}|$ is the number of interactions between the source of information *s* and the *trustor*.

In particular, a source of information is seen as reliable when his credibility (σ_s) is greater than or equal to 0.8. Sources of information with credibility lower than 0.8 are considered liars, and they are ignored by the *trustor*, such as shown in Figure 3 (b). Notice that, due to the uncertainty principle discussed previously [Ashtiani and Azgomi 2014], an agent playing as a source of information becomes unreliable over time as the number of lies told by him increases. Also, as it is possible to see in Figure 3 (b), at the last iteration (iteration 11), due to uncertainty of the information, even when all members of

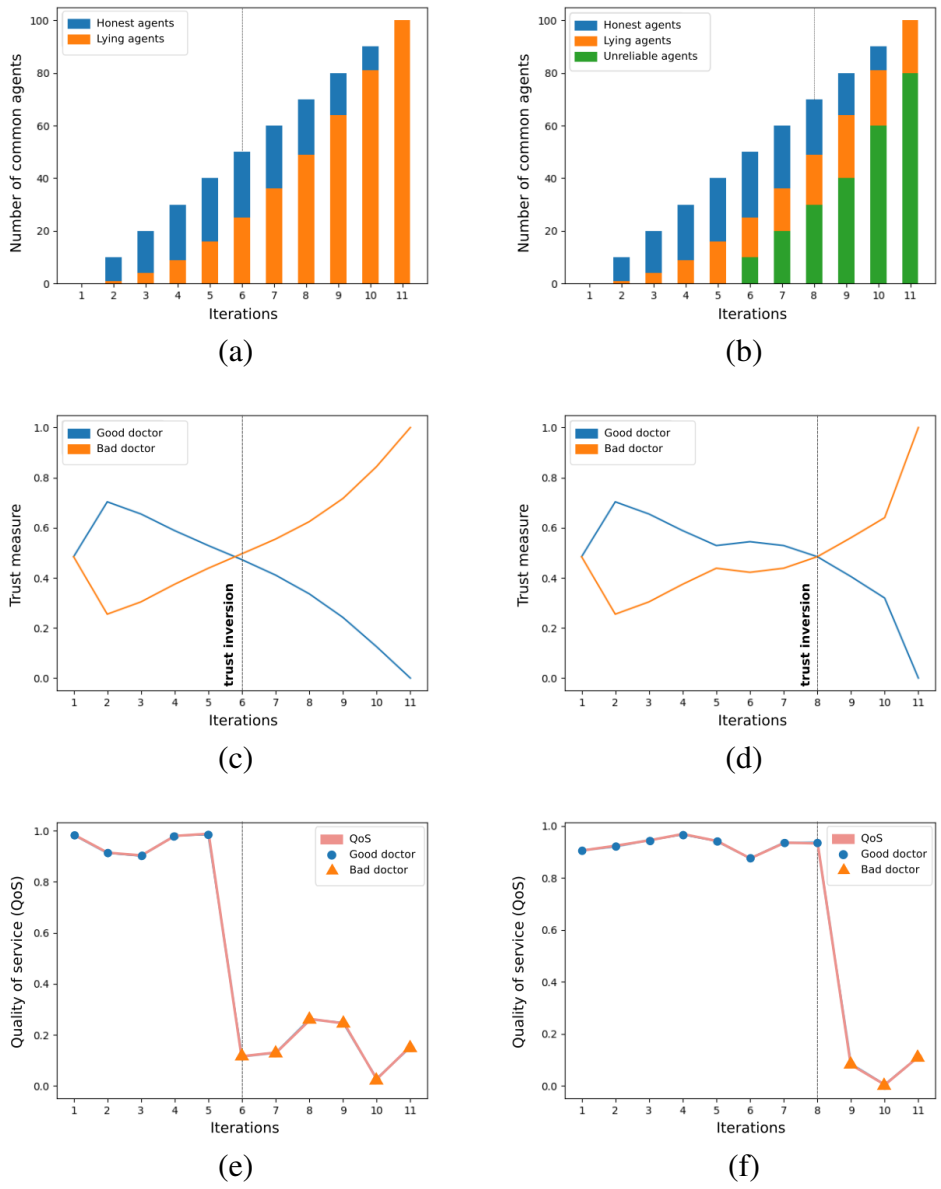

Figure 3. Results of the experiments A (trust computing based on reputation) and B (trust computing based on shared evaluations): (a) growth of the number of lying agents per iteration for experiment A, (b) growth of the number of lying agents per iteration and identifying unreliable sources of information for experiment B, (c) trust inversion for experiment A, (d) trust inversion for experiment B, (e) task assignment for experiment A, and (f) task assignment for experiment B.

the society are telling lies, the *trustor* still considers some evaluations shared by lying agents as truths. Moreover, remark that in this experiment, due to the exclusion of liars performed by the *trustor*, the number of lying agents exceeds the number of honest agents only at iteration 8 (Figure 3 (d)), which ensures that the *trustor* has more interactions with the good doctor, even receiving recommendations from lying agents.

6. Conclusion and Discussion

In this work, we present a way to compute trust using a QuAD-V framework, where agents are able to vote for or against arguments that better express their satisfaction degree with a service provided by a partner. Moreover, due to the use of social evaluations as social image, shared evaluation, and reputation, our trust computing approach presents two main advantages: (*i*) the trust in a partner can be estimated even the *trustor* has not ever interacted with such a partner yet, since in this situation the *trustor* can use the evaluations shared with him by other agents or the partner's reputation to estimate a trust measure; and (*ii*) making the use of social image, our approach provides a simple mechanism for validating the credibility of an agent that plays as a source of information. In particular, this mechanism allows the *trustor* to verify the veracity of an evaluation shared with him by comparing his social image to such an evaluation.

Moreover, as presented in our results, due to the validation mechanism of credibility based on shared evaluations and social image, the *trustor* can expose lying agents, ignoring the future information shared by them. As we demonstrated in our experiments, slandering and promotion attacks are more effective when the trust is computed based on just reputation since, in such an approach, it is not possible to identify the agents that produce and share fake information.

As future work, we intend to extend the task delegation scenario presented herein. In this extension, the agents will be able to have a partial point of view about a partner. This modification requires the implementation of an algorithm to ensure that agents present a rational behavior to vote on the arguments of the QuAD-V. Another issue that could be explored in future works is a study about the reasons that lead an agent to lie. In our experiments, the agents start lying from a given iteration, sharing fake information with the *trustor*, but they do not have a good reason to do that. For example, in the case study presented herein, in order to motivate the spreading of fake information, the bad doctor could offer a reward to agents that share good evaluations about the service provided by him.

References

Ashtiani, M. and Azgomi, M. A. (2014). Contextuality, incompatibility and biased inference in a quantum-like formulation of computational trust. *Advances in Complex Systems*, 17(05):1450020.

Baroni, P., Romano, M., Toni, F., Aurisicchio, M., and Bertanza, G. (2015). Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation*, 6(1):24–49.

Braga, D. D. S., Niemann, M., Hellingrath, B., and Neto, F. B. D. L. (2018). Survey on computational trust and reputation models. *ACM Computing Surveys (CSUR)*, 51(5):1–40.

Buccafurri, F., Comi, A., Lax, G., and Rosaci, D. (2015). Experimenting with certified reputation in a competitive multi-agent scenario. *IEEE Intelligent Systems*, 31(1):48–55.

Castelfranchi, C. and Falcone, R. (1998). Towards a theory of delegation for agent-based systems. *Robotics and Autonomous systems*, 24(3-4):141–157.

Castelfranchi, C. and Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*, volume 18. John Wiley & Sons.

Castelfranchi, C. and Guerini, M. (2007). Is it a promise or a threat? *Pragmatics & Cognition*, 15(2):277–311.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005). On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 378–389. Springer.

Cho, J.-H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):1–40.

Conte, R. and Paolucci, M. (2002). *Reputation in artificial societies: Social beliefs for social order*, volume 6. Springer Science & Business Media.

Conte, R. and Paolucci, M. (2003). Social cognitive factors of unfair ratings in reputation reporting systems. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 316–322. IEEE.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Griffiths, N. (2005). Task delegation using experience-based multi-dimensional trust. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 489–496.

Kunz, W. and Rittel, H. W. (1970). *Issues as elements of information systems*, volume 131. Citeseer.

Miceli, M. and Castelfranchi, C. (2000). The role of evaluation in cognition and social interaction. *Human cognition and agent technology*, pages 225–262.

Pinyol, I. and Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25.

Rago, A. and Toni, F. (2017). Quantitative argumentation debates with votes for opinion polling. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 369–385. Springer.

Rago, A., Toni, F., Aurisicchio, M., and Baroni, P. (2016). Discontinuity-free decision support with quantitative argumentation debates. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Sabater, J., Paolucci, M., and Conte, R. (2006). Repage: Reputation and image among limited autonomous partners. *Journal of artificial societies and social simulation*, 9(2).

Sabater, J. and Sierra, C. (2001). Regret: reputation in gregarious societies. In *Proceedings of the fifth international conference on Autonomous agents*, pages 194–195.

Singh, R. R. (2018). *Designing for multi-agent collaboration: a shared mental model perspective*. PhD thesis.

Solhaug, B., Elgesem, D., and Stolen, K. (2007). Why trust is not proportional to risk. In *The Second International Conference on Availability, Reliability and Security (ARES'07)*, pages 11–18. IEEE.