A Protocol Based on a Game-Theoretic Dilemma to Prevent Malicious Coalitions in Reputation Systems

Grégory Bonnet

ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-187

A Protocol Based on a Game-Theoretic Dilemma to Prevent Malicious Coalitions in Reputation Systems

Grégory Bonnet[1]

Abstract. In decentralized and open systems, a large number of agents interact and make collective decisions in order to share resources. As those systems are open, the presence of malicious agents needs to be considered. A way to deal with such agents in a decentralized fashion is to use reputation systems. But, as reputation systems are based on the aggregation of local trust between the agents, they are vulnerable to malicious coalitions, particularly to self-promotion based on false identities. In this paper, we propose a game-theoretic approach to prevent such manipulations. Its main feature is that honest agents use in turn a false-name manipulation to fool malicious agents and to drive them into a dilemma. We show that the best response to that dilemma in terms of mixed strategy equilibrium leads the malicious agents to give up most of their manipulations.

1 Introduction

In decentralized and open systems, a large number of agents interact and make collective decisions in order to share resources such as skills, knowledge, computational power, or mass memory. Those systems are designed to provide a decentralized service to their members such as composing Web services, grid computing, or providing an electronic market place. In order to insure the nominal use of such systems, it is assumed that the agents follow the rules defined by a protocol or by norms. Even if the agents might be altruistic, cooperative or competitive, it is also assumed that they are honest entities within the system. However, as those systems are open, some malicious agents can misuse the rules to their own profit or disrupt the service. Consequently, their presence needs to be considered. Such problematics lead to study the concept of trust and the use of reputation systems. Those systems allow the agents to modelize the interactions they observe or they make in order to decide if interacting with a given agent is *a priori* acceptable. This acceptance (or trust) notion means that the investigated agent behaves well and is reliable. Even if the reputation systems are designed to detect the behavior of a single agent, they are vulnerable to malicious coalitions [12]. Indeed, reputation systems are based on the aggregation of local subjective trust values between the agents. In addition of the aggregation of opinion issues, a set of malicious agents is able to report a high trust level for each other in order to artificially increase their reputation. Such manipulation is a self-promoting attack and can be used for instance to fool eBay's reputation system [10], Google's PageRank algorithm [5], or even to free-ride on peer-to-peer networks [18]. Conversely, a set of malicious agents is able to report a low trust level for an honest agent and artificially decrease its reputation. Such manipulation is a slandering attack and can be used jointly with self-promotion in order to enhance its effects. Moreover, in any system where the authentification mechanisms can be fooled, a single malicious agent can enter the system with multiple identities and create a virtual coalition to manipulate the reputation system. Such manipulation is called a Sybil attack [11]. As dealing with malicious coalitions is critical for reputation systems, much work have been done towards this end [14]. These proposals cover a broad area ranging from cryptographic puzzles to insure the unicity of the agents, to the detection of communities inside social networks, and the design of robust reputation functions. A recent way assumes that the malicious agents are rational agents and proposes to use game-theoretic techniques to provide incentives not to fool the system [8]. Such approaches are interesting because they can be extended to other applications such as combinatorial auctions and voting procedures. In this context, this paper proposes a game-theoretic approach to prevent some specific collusions in reputation systems. Its specificity is based on the fact that the honest agents use in turn a Sybil attack to fool the malicious agents. The paper is organized as follows. We first introduce related work in Section 2. In Section 3, we give details about the proposed protocol before analysing it under a game-theoretic perspective in Section 4. Before concluding, we show the advantages and the limits of our protocol with simulation results in Section 5, raising some questions to be answered by future work.

2 Related work

In the literature about trust, reputation functions can be symmetric, meaning that each agent in the system contributes to the reputation calculus, or asymmetric, meaning that the local trust is only propagated through *a priori* trusted agents such as in a social network. [4] shows that a symmetric reputation function cannot prevent manipulations whereas asymmetric reputation functions can, if and only if they satisfy strong properties that make them weakly informative and difficult to design. Moreover, if reputation functions reduce the trust values of the witness agents when the agents they recommend act in a bad way, such approach is vulnerable to whitewashing where malicious witnesses can change their identities to reset to a default trust value. Hence, other solutions based on detecting the malicious witnesses before interacting with the recommended agent were proposed [14]. A first class of solutions consists in preventing the Sybil attacks that facilitate the manipulations. Some approaches such as [3] propose to use a central trusted authority that certifiates each agent, but it reduces the decentralization and openness properties of the system. Moreover, a central authority is alway a failure point in the system. Another approach introduces a recurring cost to join the system, such as solving a cryptographic puzzle [2], paying a financial

[1] University of Caen Lower-Normandy, UMR CNRS 6072 GREYC, France, email: gregory.bonnet@unicaen.fr

fee or using *captchas*. Therefore, creating a large number of false identities to manipulate the system is difficult. However, malicious agents can have a huge amount of computational resources thanks to botnets, and these methods are still very constraining for honest agents. A second class of solutions [19] consists in detecting malicious communities inside a social network. Such approaches assume the Sybil agents present either a high clustering coefficient with a few links outside [7, 21], or a common identifier range [6]. They use clustering techniques from link mining or graph analysis litterature to cut the graph between honest and Sybil agents. However, these approaches address the Sybil attack problem and cannot consider malicious coalitions between true distinct agents. In order to propose a general framework that can address both Sybil attacks and distinct agents collusions, recent works focus on game-theoretic approaches [8, 15, 16, 17] by assuming that the malicious agents are rational. Those approaches make the link with the false-name manipulation problem in weighted voting game [1] and combinatorial auctions [20] where false-name-proofness means that an agent never benefits from participating more than once. For instance, the Informant protocol [16] and its application to Tor [17] is based on a game where the honest agents use a Dutch auction to reward the malicious agents that reveal themselves. However, it assumes the reward is built on a recurring fee, and the protocol may incite new malicious agent to join the system. In this context, we propose to prevent self-promoting manipulation on reputation functions without recurring fee or huge computational cost for the honest agents. For this, we propose a protocol based on a game that causes a dilemma only to malicious agents if they are rational. We use the answer to this dilemma to partition agents into honest and malicious agents.

3 A Sybil-based protocol

We first introduce the reputation system we consider, the manipulations we address and highlight the major features of our protocol.

3.1 Reputation system

Let us reuse the definition of a reputation system given by [4]. For convenience, we denote by A the truster agent, by B the trustee agent and by W_i a witness agent:

Definition 1 *Let $\mathcal{G} = (V, E)$ be an oriented graph where V is a set of agents and $E \subseteq V \times V$ an interaction relation labeled by a trust value $c : E \longmapsto [0, 1]$. The reputation of an agent B according to an agent A with respect to a trust network $\mathcal{G}$ is given by a function $f_{\mathcal{G}} : V \times V \longmapsto [0, 1]$ where:*

$$f_{\mathcal{G}}(A, B) = \max_{\mathcal{P}_{AB} \in \mathcal{G}} \oplus_{P \in \mathcal{P}_{AB}} \odot (P)$$

$\mathcal{P}_{AB}$ is the maximum set of disjoint paths between A and B in $\mathcal{G}$ in the sense of inclusion, $\odot$ is an aggregation operator on c along a single path P between A and B, and $\oplus$ is an aggregation operator on g along all disjoint paths between A and B.

Example 1 *The FlowTrust reputation system proposed by [4] is defined by $\odot = \prod$, $\oplus = \max$. The reputation of an agent B according to an agent A is given by the maximum value over the products of the trust values among all disjoint paths between A and B.*

Each agent A makes a decision about trusting or not the agent B with respect to its reputation.

Definition 2 *Let $d : V \times V \longmapsto \{0, 1\}$ be a decision function where 0 means that A distrusts B and 1 means that A trusts B. The value of $d(A, B)$ is given by a threshold function over $f_{\mathcal{G}}(A, B)$.*

Example 2 *Let us assume that $f_{\mathcal{G}}(A, B) \in [0; 1]$ where 0.5 means indifference, then $d(A, B)$ can be defined in $\{0, 1\}$ by $d(A, B) = (f_{\mathcal{G}}(A, B) > 0.5)$.*

In our model, as $f_{\mathcal{G}}$ represents the collective agregation mechanism, g and $\oplus$ are common to all agents, whereas d is a private and subjective decision function. Therefore, d can be different for all agents. Whatever d is, computing the reputation of a given agent involves searching through the interaction graph and asking the involved agents their trust value.

3.2 Manipulation properties

In this paper, we only consider self-promoting manipulations where the malicious agents support each other and want to fool every other agent in the system. It corresponds to many real-world attacks, such as PageRank manipulation, Tor poisoning and free-riding on peer-to-peer networks. Indeed, the aim of the malicious agents is that all honest agents interact with at least one member of the malicious coalition. Consequently, a self-promoting manipulation is defined as follows.

Definition 3 *$\forall\ A \in V$ that asks an agent B its value $c(B, W_i)$, if W_i is in collusion with B then $c(B, W_i) > 0.5$.*

This definition means that every agent within a malicious coalition reports a high trust towards the others. They form a high mutual trust cluster. We can deduce two things from this definition. Firstly, the malicious coalition is characterized by a high mutual trust among its members. Therefore, we assume that the honest agents know a suspicion function which computes the probability that two agents are in collusion with respect to a given trustee agents. Such a function is a heuristic representing a knowledge about the assumption we made on the malicious coalition. Such function may be defined as follows:

Definition 4 *Let $M : V \times V \longmapsto [0, 1]$ be a* suspicion function *such that $M(W_1, W_2) = c(W_1, W_2) \times c(W_2, W_1)$. The higher is $M(W_1, W_2)$, the more likely W_1 and W_2 are in collusion.*

Considering two distinct witnesses that individually trusts the trustee agent B, the more the witnesses trust each other, the more likely they are in collusion. Secondly, as the malicious agents want to fool every honest agent within the system, their behavior is the same in the front of all the other agents. Consequently, they can be fooled in return. Indeed, an honest agent A can ask a witness W_1 its trust value for another witness W_2, then ask W_2 its trust value for W_1, both under the pretence of computing a trust network whereas A uses these trust values to determine if there is a malicious collusion.

3.3 Building the trust network

The protocol we propose needs to detect the colluding agents within the set of witnesses according to Definition 4. To this end, the protocol needs to build the trust network from the witnesses to the trustee and from the trustee to the witnesses in order to compute the suspicion function. However, the malicious agents can hide their relationship if a single honest agent asks for mutual trust. In order to incite the malicious agents to reveal themselves according to Definition 3,

the honest agent uses a Sybil attack to conceal its investigation: both malicious agents are asked their trust value by two apparently distinct agents (the honest agent and its Sybil). As both malicious agents believe that those truster agents are honest, they are incited to reveal a high mutual trust value in order to fool them. Consequently, we can define a protocol based on a Sybil attack to force the malicious agents to reveal their collusions. Algorithm 1 presents the main steps of this protocol. First, A uses a given reputation system to compute the trustee's reputation (line 1). Thereby, A gets a set W of witnesses for B and can compute $d(A, B)$. As noticed by [8], B may be an honest agent under a slandering attack if $d(A, B) = 0$ or B may be a malicious agent doing self-promotion if $d(A, B) = 1$. Indeed, only successful manipulations need to be considered; and a manipulation is successful if and only if it leads A to make the *right* decision: trusting if there is self-promotion, or distrusting if there is slandering. As we only consider self-promoting manipulations, the protocol only considers the case where $d(A, B) = 1$ (line 2).

1: A computes $f_{\mathcal{G}}(A, B)$ with a given reputation system
2: **if** $d(A, B) = 1$ **then**
3: A selects the subset W' of W that trusts B
4: **for** each $W_i, W_j \in W'$ $(W_i \neq W_j)$ **do**
5: A generates two Sybil agents A' and A''
6: A' asks W_i its trust value $c(W_i, W_j)$
7: A'' asks W_j its trust value $c(W_j, W_i)$
8: A computes $M(W_i, W_j)$
9: **end for**
10: A uses all the $M(W_i, W_j)$ values to revise $f_{\mathcal{G}}(A, B)$
11: **end if**

Algorithm 1: A Sybil-based protocol

Next, A selects the subset of witnesses that trust B (line 3). For each couple of witnesses, A generates two Sybil agents (line 4 and 5) that will be used to fool the possible malicious agents. Those Sybil ask W_i if it trusts the witness W_j (line 6) and ask W_j if it trusts W_i (line 7). If W_i (respectively W_j) is an honest agent, it will answer honestly and, if W_i is a malicious agent, it will answer honestly too because it is fooled by the Sybil according to Definition 3. Consequently, the agent A can build the real mutual trust network and decide if there is a malicious collusion with respect to Definition 4.

3.4 Using the suspicion value

Once an honest agent A has computed a set of suspicious values $M(W_i, W_j)$ over the selected witnesses, it needs to use them to compute $f_{\mathcal{G}}(A, B)$ (line 10) once again. The higher the suspicion values for W_i, the less trustable the testimony $c(W_i, B)$ by W_i. Consequently, we propose a mechanism, given in Algorithm 2, that *a posteriori* removes the testimony from the $f_{\mathcal{G}}(A, B)$ calculus.

1: Let $\mathcal{P}_{AB}$ be the set of paths between A and B
2: **for** $\forall P \in \mathcal{P}_{AB}$ **do**
3: **for** $\forall W_i \in P$ such that $c(W_i, B) > 0$ **do**
4: $\mathcal{P}_{AB} \leftarrow \mathcal{P}_{AB} \setminus \{P\}$ with probability $\max_{W_j \in W'} M(W_i, W_j)$
5: **end for**
6: **end for**
7: A computes $f_{\mathcal{G}}(A, B)$

Algorithm 2: Using the suspicion value

The honest agent proceeds as follows: for each path in the trust network that relies on a testimony given by an agent W_i in favor of B, the honest agent removes this path with a probability equal to the highest suspicion value computed for W_i. If there is several suspect testimonies on a single path, each of them can independently removes the path.

3.5 Overall cost of the protocol

As our protocol is a layer implemented over a reputation system given by $f_{\mathcal{G}}(A, B)$, it increases the global communication cost of this latter. Such communication cost is the number of messages exchanged in one round of communication.

Proposition 1 *Let $|W|$ be the number of selected witnesses in Algorithm 1 line 3. The communication cost of the protocol given by Algorithm 1 is in $\mathcal{O}(|W|^2)$*

Proof 1 *Each time the reputation for an agent B is computed by an agent A, a Sybil agent A' asks the $|W|$ witnesses their trust value for each of the other witnesses. Consequently, our protocol adds $|W|^2$ new messages to the original protocol that computes $f_{\mathcal{G}}(A, B)$.*

4 Analyzing the dilemma

As this protocol is known by all the agents in the system, a malicious agent needs to decide if it will answer honestly when any other agent asks for its trust values. Making such a decision is a dilemma that we now analyze from a game-theoretic perspective. To analyze the dilemma, we first present a strategic form game to represent it. Then, we show some of its properties.

4.1 Strategic form game

In the sequel, we consider the following notations as we assume the agents are rational.

Definition 5 *Let $g \in \mathbb{R}$ and $c \in \mathbb{R}$ such that $g > c$ be the reward for the malicious agent that fool the system and the cost of being identified as a malicious agent respectively. Let $\delta \in [0, 1]$ be the probability that a given agent asking a malicious one its trust value is a Sybil agent.*

Table 1 gives the strategic form of the game from the malicious agents' point of view as this game is a zero-sum game.

Agent	Reveal	Conceal
Honest	$(1-\delta)g$	0
Sybil	$-\delta c$	δg

Table 1. Strategic form game for the malicious agent

Informally, the game is the following. If the malicious agent reveals its trust in a witness to an honest agent, it fools the honest agent. If the malicious agent conceals its trust to a Sybil agent, it fools the protocol, and then fools the honest agent that generated the Sybil. If the malicious agent reveals its trust to a Sybil agent, it receives a penalty due to the cost of a collusion between it and the witness. In the last case, the malicious agent has no reward, nor penalty if it conceals its trust to an honest agent. The structure of the payoff matrix

corresponds to a matching pennies game [9] where δ is the mixed strategy parameter of the protocol (the malicious agent's opponent). As this dilemma is a variant of the matching pennies game, we know that there is no pure strategy profile that allows a player to maximize its reward. Consequently, the malicious agent needs to play a mixed strategy.

4.2 Mixed strategy Nash equilibrium

Let us denote for a malicious agent R the action of revealing its trust and C concealing it. Let us also denote for the protocol S the fact that the truster agent is a Sybil and H the fact it is an honest agent. Consequently, we can denote $\pi_B = \langle \sigma(R) = (1-m), \sigma(C) = m \rangle$ the malicious agent's mixed strategy profile, and $\pi_p = \langle \sigma(H) = (1-\delta), \sigma(S) = \delta \rangle$ the protocol's mixed strategy profile.

Proposition 2 *The* mixed strategy Nash equilibrium *of the game depicted in Table 1 is:*

$$m = \frac{g+c}{2g+c} \qquad \textit{and} \qquad \delta = \frac{g}{2g+c}$$

Proof 2 *The expected utility of π_B with respect to m and δ is:*

$$\begin{aligned} u_{\pi_B}(m,\delta) &= (1-m)((1-\delta)g - \delta c) + m\delta g \\ &= g - \delta g - \delta c + m(2\delta g + \delta c - g) \end{aligned}$$

As the malicious agent wants to maximize $u(\pi_B)$, we find the roots of the partial derivative of $u_{\pi_B}(m,\delta)$ with respect to δ.

$$\begin{aligned} -g - c + m(2g+c) &= 0 \\ m &= \frac{g+c}{2g+c} \end{aligned}$$

Likewise, as the protocol wants to minimize $u(\pi_B)$, we find the roots of the partial derivative of $-u_{\pi_B}(m,\delta)$ with respect to m.

$$\begin{aligned} -2\delta g - \delta c + g &= 0 \\ \delta &= \frac{g}{2g+c} \end{aligned}$$

□

Moreover, we can notice that $m = 1 - \delta$ which is a feature of matching pennies games. Even if the penalty is zero, a rational malicious agent that maximize its reward needs to play a mixed strategy such that $m = \delta = \frac{1}{2}$.

4.3 Successful attack probabilities

However, even if the malicious agent has a mixed strategy that maximize its reward, manipulating a reputation system needs to play the game twice. Indeed, a malicious agent needs to fool both the honest agent and its Sybil. In the other cases, the manipulation is a failure. Concealing first in front of an honest agent leads to give up the attack whatever is the next game, and revealing in front of a Sybil agent in the second leads to being sanctionned for collusion. Consequently, a successful manipulation is defined as follows:

Definition 6 *A manipulation is* successful *if and only if the malicious agent plays reveal against an honest agent on a first game then plays conceal against a Sybil agent on another game.*

Agent	Reveal	Conceal
Honest	$m(1-m)$	m^2
Sybil	$(1-m)^2$	$m(1-m)$

Table 2. Occurence probabilities of joint strategies

According to the mixed strategy Nash equilibrium we determined in the previous section (and $m = 1 - \delta$), the Table 2 gives the occurence probabilities of the joint strategies. We assume that due to the simultaneous requests in the system, and due to the fact that the Sybil agents wait before asking the trustee its trust valuation, a malicious agent cannot determine if two dilemmas are correlated. Hence,

Proposition 3 *A manipulation is successful if the joint strategy (honest,reveal) then (Sybil,conceal) occurs. Consequently, the probability that a given manipulation is successful is $m^2 - 2m^3 + m^4$.*

As we know the best value for m thanks to Proposition 2, we can express the probability of success in terms of g and c. Moreover, we can express c as a fraction of g. In this case, even if $c = 0$, the probability that a manipulation is successful is only 0.0625. However, it is important to notice that this success probability only concerns the trust relationship between two malicious agents. To conclude this analyzis, the protocol we defined forces the malicious agents to reveal their high mutual trust and, hence, enables some honest agents to detect them, or to give up some manipulations to avoid being suspected.

5 Simulation results

In order to evaluate our approach, we implemented it over a reputation system proposed by [4] and compared its performances with and without the dilemma. Although, all the results are strongly dependent on a huge number of parameters from the topology, to the distribution of the trust values within, through the kind of reputation system, they give us insights about the efficiency of our protocol.

5.1 Common setting

We consider the trust network as a binomial Erdös-Rényi graph with $p = 0.15$, and where the trust values are fixed according to a uniform distribution. For each experiment, we fix the number of agents and the proportion of malicious agents without making any assumption about their position within the graph. We also consider that there is no cost for colluding ($c = 0$). We launched $10,000$ simulations where a random honest agent evaluates a subset of random trustee agents. Indeed, a decentralized reputation system cannot evaluate all the agents in the system without scaling problems. Then the honest agent computes the reputation of each trustee without our protocol, then with our protocol under a pure and a mixed malicious strategy. In each case, if a malicious agent maximizes its reputation with respect to the other trustee agents, we consider that there is a successful manipulation. Our performance criterion is the proportion of successful manipulations over the $10,000$ simulations. Obviously, we need to use a given reputation system in order to compare and implement our protocol. We chose the FlowTrust reputation system given in Example 1. This reputation system is known to be robust against manipulation when the agents have a global view of the system although they are not very informative. All the results are given

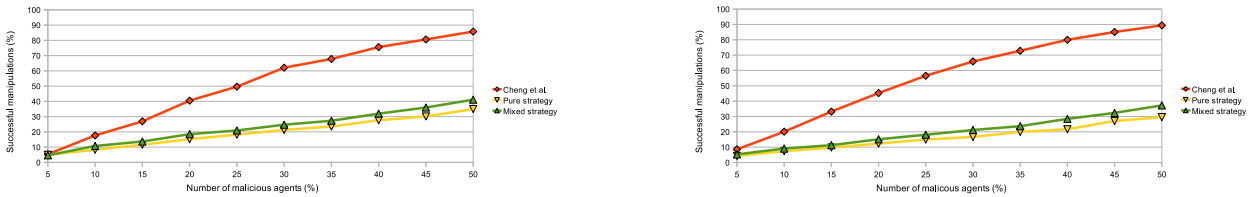

Figure 1. Successful attacks w.r.t. the proportion of malicious agents over a network of 50 and 100 agents

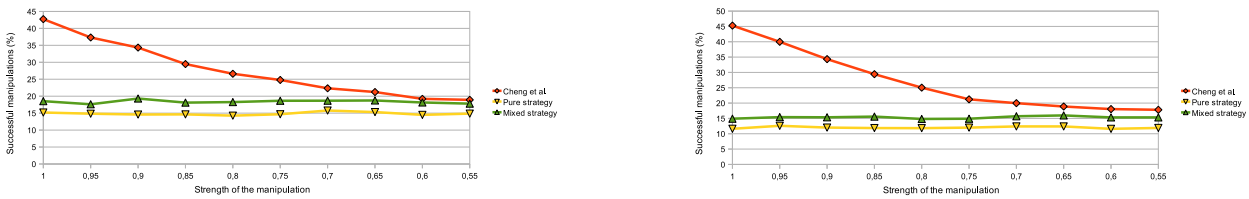

Figure 2. Successful attacks w.r.t. the manipulation strength over a network of 50 and 100 agents

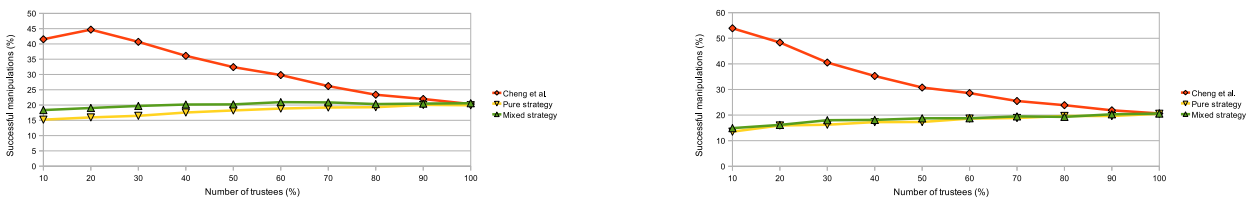

Figure 3. Successful attacks w.r.t. the number of evaluated agents over a network of 50 and 100 agents

in Figure 1, 2 and 3. For each of those figures, the red diamond-shaped curves represent the successful manipulations with the original reputation system. The green head-up-triangle-shaped curves represent the successful manipulations with our approach under a mixed malicious strategy whereas the yellow upside-down-triangle-shaped curves represent the successful manipulations under pure malicious strategy (always revealing).

5.2 About the number of malicious agents

In order to highlight the influence of the network in terms of size and malicious agents, we ran two experiments with a network of 50 agents and 100 agents respectively. The proportion of malicious agents varies from 10% to 50%, and in each simulation an honest agent evaluates 5 trustee agents chosen at random before deciding the one it can trust. The results are shown on Figure 1. Obviously, the number of successful manipulations increases in all cases as the number of malicious agents grows, with a slightly higher number of manipulations for the network of size 100. However, in average, the mixed strategy reduces by 55% the number of successful manipulations, whereas a pure strategy sees its manipulations reduced by 62%. Finally, the results have the same shape for both size of networks with a difference of 11% in average in favor of large networks. Consequently, our protocol seems to be sensitive to the size of the network. The more agents in the network, the more efficient the protocol. Moreover, the theoretical results are confirmed as the malicious agents need to play a mixed strategy if they want to maximize their number of successful manipulations. Even though the malicious agents play an optimal mixed strategy, their manipulations are reduced by more than a half.

5.3 About the strength of the manipulation

A rational malicious agent reduces the strength of its manipulation in order to avoid being suspected by the honest agents. In this case, the malicious coalition reduces the value of their mutual trust. In order to highlight the influence of such malicious behavior, we ran two experiments with a network of 50 agents and 100 agents respectively. In each of them, we considered 20% of malicious agents, and the trust value they reported varies in the range of $[0.5, 1]$. The results are shown on Figure 2. Obviously, the number of successful manipulations decreases with respect to the original reputation system as the malicious agents reduce their manipulation. It converges to around 20%, the proportion of malicious agents within the network. We can notice that our protocol is very efficient as, in all cases, the successful manipulations remain around 20% for the mixed strategy and 15% for the pure strategy. When the network grows in size, the successful manipulations are reduced to 15% and 10% respectively. Thus, our protocol can prevent manipulations and cannot be manipulated by the malicious agents. Moreover, as Algorithm 2 removes the suspected witnesses stochastically, the performance is not reduced when malicious agents reduce their mutual trust value to mimic honest agents.

5.4 About the amount of information

In the previous experiments, we considered a decentralized case where the honest agent only evaluates a subset of agents in the network: this is the trade-off between exploration and exploitation. In order to highlight the influence of the amount of information owned by an agent, we ran two experiments with a network of 50 agents and 100 agents respectively. In each of them, we considered 20% of malicious agents and the number of random trustee agents the honest agent evaluates varied from 10% to 100% of the network. The results are shown on Figure 3. Once again, the results present the same structure with 50 or 100 agents. We can notice that the number of successful manipulations decreases under the original system as the number of considered trustees increases. The efficiency of our protocol under a mixed strategy decreases from around a reduction of manipulations by 64% to a null gain. We can also notice that the mixed strategy is as efficient as the pure strategy: whatever the strategy the malicious agents play, they cannot increase the number of successful manipulations. Finally, in both cases, the performance of our protocol converges towards the performance of the original system. However, the original reputation system is robust to manipulation in a centralized system. In real systems, we cannot assume that an agent can evaluate all the other agents in the system. Consequently, our protocol is still very efficient in general.

6 Conclusion

In order to insure the nominal use of decentralized and open systems, the presence of malicious agents needs to be considered. Such problematics are addressed by reputation systems, but even if those systems are designed to detect the malicious behavior of a single agent, they are vulnerable to malicious coalitions and Sybil attacks. As dealing with malicious coalitions is critical for reputation systems, much work has been done towards this end. These proposals cover a broad area ranging from cryptographic puzzles to insure the unicity of the agents, to the detection of communities inside social networks, and the design of robust reputation functions. However, those approaches cause a partial centralization of the system, or are costly for the honest agents. A recent way proposes to use game-theoretic techniques to provide incentives not to fool the system. In this context, we propose a protocol based on a game-theoretic dilemma to detect, and therefore prevent, self-promoting coalitions in reputation systems. Its specificity is based on the fact that the honest agents use in turn a Sybil attack to fool the malicious agents. Our protocol leads the malicious agents to reveal their mutual trust relationships, that are then used as a heuristic to detect collusions by the honest agents. Our theoretical analysis shows that the malicious agents need to play a mixed strategy and give up some manipulations in order to maximize their efficiency. Our simulations show that our protocol reduces in average the manipulations by more than a half. Moreover, its efficiency remains high even if the malicious agents hide themselves. However, our work raises several perspectives. Firstly, we need to improve our experiments to highlight the limits of our protocol. How the protocol behaves when the trust values of the honest agents are correlated? Moreover, how the protocol behaves when compared to other reputation systems such as EigenTrust [13]? Secondly, we need to address the problem of the suspicion function to enhace the overall performance of our protocol. Indeed, this function represents a heuristic about what is a malicious behavior and many definitions can be applied. For instance, we can suspect a single agent that provides too many testimonies. As we did not make any assumption about the topological relationships between the malicious agents, we can also combine our heuristic with others based on the topology of the network, such as SybilLimit [21]. This information might enhace the efficiency of our approach. Another way is to consider the dynamics of the system. Indeed, our protocol only considers a snapshot of the system at a given time. However, if a malicious agent plays a mixed strategy, an honest agent does not. Consequently, considering several answers about the dilemma may allow to detect if some agents are playing a mixed strategy, and therefore to deduce that they are malicious agents. Reasoning about the strategy, and not the answer in itself, is a way for overcoming the current limits of our protocol.

REFERENCES

[1] Y. Bachrach and E. Elkind, 'Divide and conquer: false-name manipulations in weighted voting games', in *Proceedings of the 7th AAMAS*, pp. 975–982, (2008).

[2] N. Borisov, 'Computational puzzles as Sybil defenses', in *Proceedings of the 6th P2P*, pp. 171–176, (2006).

[3] M. Castro, P. Drusche, A. Ganesh, A. Rowstron, and D.-S. Wallach, 'Secure routing for structured peer-to-peer overlay networks', in *Proceedings of the 5th OSDI Symposium*, (2002).

[4] A. Cheng and E. Friedman, 'Sybilproof reputation mechanisms', in *Proceedings of the 3rd P2PEcon*, pp. 128–132, (2005).

[5] A. Cheng and E. Friedman, 'Manipulability of PageRank under Sybil strategies', in *Proceedings of the 1st NetEco Workshop*, (2006).

[6] T. Cholez, I. Chrisment, and O. Festor, 'Efficient DHT attack mitigation through peers' ID distribution', in *Proceedings of the 24th IPDPS*, pp. 1–8, (2010).

[7] V. Conitzer, N. Immorlica, J. Letchford, K. Munagala, and L. Wagman, 'False-name-proofness in social networks', in *Proceedings of the 6th WINE*, pp. 1–17, (2010).

[8] V. Conitzer and M. Yokoo, 'Using mechanism design to prevent false-name manipulations', *AI Magazine*, **Vol. 31(4)**, 65–77, (2010).

[9] T. Dang, 'Gaming or guessing: mixing and best-responding in matching pennies', Technical report, University of Arizona, (2009).

[10] F. Dini and G. Spagnolo, 'Buying reputation on eBay: do recent changes help?', *IJEB*, **Vol. 7(6)**, 581–598, (2009).

[11] J.-R. Douceur, 'The Sybil attack', in *Proceedings of the 1st IPTPS*, (2002).

[12] K. Hoffman, D. Zage, and C. Nita-Rotaru, 'A survey of attack and defense techniques for reputation systems', *ACM Computing Survey*, **Vol. 42(1)**, 1–31, (2009).

[13] S.-D. Kamvar, M.-T. Schlosser, and H. Garcia-Molina, 'The EigenTrust algorithm for reputation management in P2P networks', in *Proceedings of the 12th WWW*, pp. 640–651, (2003).

[14] B.-N. Levine, C. Shields, and N.-B. Margolin, 'A survey of solutions to the sybil attack', Technical report, University of Massachussets Amherst, (2006).

[15] X. Liao, D. Hao, and K. Sakurai, 'A taxonomy of game theoretic approaches agains attacks in wireless ad hoc networks', in *Proceedings of the 28th SCIS*, pp. 1–8, (2011).

[16] N.-B. Margolin and B.-N. Levine, 'Informant: detecting Sybils using incentives', in *Proceedings of the 11th FC*, pp. 192–207, (2007).

[17] A.-K. Pal, D. Nath, and S. Chakreborty, 'A discriminatory rewarding mechanism for Sybil detection with applications to Tor', in *Proceedings of the 8th ICCCIS*, pp. 84–91, (2010).

[18] M. Sirivianos, J.-H. Park, R. Cheng, and X. Yang, 'Free-riding in BitTorrent networks with the large view exploit', Technical report, California Irvine, (2001).

[19] B. Viswanath, A. Post, K.-P. Gummadi, and A. Mislove, 'An analysis of social network-based Sybil defenses', in *Proceedings of SIGCOMM'10*, (2010).

[20] M. Yokoo, Y. Sakurai, and S. Matsubara, 'The effect of false-name bids in combinatorial auctions: new fraud in Internet auctions', *Game and Economic Behavior*, **Vol. 46**, 174–188, (2004).

[21] H. Yu, P.-B. Gibbons, M. Kaminsky, and X. Feng, 'SybilLimit: a near-optimal social network defense against Sybil attacks', *IEEE/ACM Transactions on Networking*, **Vol. 18(3)**, 885–898, (2010).