# Towards a Probabilistic Model of Distributed Reputation Management *

Bin Yu and Munindar P. Singh

Department of Computer Science
North Carolina State University
Raleigh, NC 27695-7535, USA

{byu, mpsingh}@eos.ncsu.edu

**Abstract.** A probabilistic model of reputation management is proposed to help agents (users) avoid interaction with non-cooperative participants. Our approach adjusts the ratings of agents based on their observations as well as the testimony from others. Our former work used a scalar value to represent the reputation ratings and combine testimonies using combination schemes from the certainty factor model. One problem is that certainty factors do not represent measures of absolute belief. Rather, they are meant to represent changes in belief. In this paper the mathematical theory of evidence is used to represent and propagate the reputation information in an electronic community. Our specific approach to reputation management leads to a decentralized society in which agents help each other weed out undesirable players.

## 1 Introduction

The worldwide expansion of network access is driving an increase in interactions among people. We view an electronic community as a social network in which each user is assigned a software agent and software agents help automate the process of word-of-mouth by a series of "referral chains." For example, users pose queries to their agents in the form of *Where is the best Chinese restaurant in the Bay Area*? The queries by the user are first seen by his agent who decides the potential contacts to whom to send the query. After consultation with the user, the agent sends the query to the agents for other likely people. The agent who receives a query can decide if it suits its user and let the user see that query. In addition to or instead of just forwarding the query to the user, the agent may respond with referrals to other users. If the agent or user wish they can discard the query and never respond to it in any way.

Moreover, the agents assist their users in evaluating the services provided by others, and find the most helpful and reliable parties to deal with. In this manner, the agents build and manage the reputations of other agents. Reputation

* This research was supported by the National Science Foundation under grant IIS-9624425 (Career Award). We are indebted to the anonymous reviewers for helpful comments.

is different from trust. We view trust as a kind of "belief" of one agent about another, and reputation as the "cumulative beliefs" from a group of agents. Previously, we used a scalar value to represent an agent's trust about another and combine testimonies using combination schemes from the certainty factor model [15]. One problem with this work is that certainty factors do not represent measures of absolute belief. Rather, they are meant to represent changes in belief [7]. The drawbacks of the certainty factor models led us to consider alternate approaches. Particularly appealing is the mathematical theory of evidence developed by Arthur Dempster and Glenn Shafer [13].

Before introducing the Dempster-Shafer theory, we attempt to show some simple justification for the approach. In general, an agent $A_i$ does not know with full certainty whether another agent $A_j$ is trustworthy or not, but $A_i$ may be able to estimate the degree of trust about $A_j$. Dempster-Shafer theory handles this uncertainty explicitly, and with more ease than the Bayesian model [9]. Moreover, some evidence available to agent $A_i$ may neither support $A_j$'s being a trustworthy or nontrustworthy agent. Dempster-Shafer theory models this "ignorance" naturally, which is cited as a major motivation for the Dempster-Shafer theory [14].

The rest of this paper is organized as follows. Section 2 presents some necessary background on belief, trust and reputation. Section 3 introduces our approach, giving the key definitions and some propagation algorithms in a TrustNet. Our experimental results are given in section 4. Section 5 presents some related work in reputation management. Section 6 concludes our paper with a discussion of the main results and directions for future research.

## 2 The Dempster-Shafer Theory of Evidence

A *frame of discernment* is defined as the whole set of propositions in which each is known to be true. In our example, suppose the *frame of discernment* $\Theta$ contains only $T$ and $\neg T$, where $T$ stands for the trustworthy relationship between any two agents. The Dempster-Shafer theory assigns a number in the range $[0, 1]$ to every subset of $\Theta$ (excluding the empty set), called *basic probability assignment (bpa)*. And the sum of all the bpa's must equal 1.

**Definition 1.** If $\Theta$ is a frame of discernment, then a function $m : 2^{\Theta} \mapsto [0, 1]$ is called a *basic probability assignment* whenever (1) $m(\phi) = 0$, and (2) $\sum_{\hat{A} \subset \Theta} m(\hat{A}) = 1$, where $\hat{A}$ is a subset of $\Theta$.

For example, we must have that $m(\{T\}) + m(\{\neg T\}) + m(\{T, \neg T\}) = 1$. This is similar to a probability assignment except that it is not necessary that the sum of the bpa's assigned to the members of $\Theta$ be equal to 1. For example, given the assignment of $m(\{T\}) = 0.8$, $m(\{\neg T\}) = 0$, $m(\{T, \neg T\}) = 0.2$, we have $m(\{T\}) + m(\{\neg T\}) = 0.8 < 1$

When agent $A_i$ is evaluating the trustworthiness of agent $A_j$, there are two sides of evidence. The first is the services offered by agent $A_j$. The second is testimonies from other agents. Suppose agent $A_i$ has the latest $h$ responses from

agent $A_j$, $S_j = \{s_{j1}, s_{j2}, \ldots, s_{jh}\}$. We use the distinct values of $\{0.0, 0.1, \ldots, 1.0\}$ to denote the quality of service (QoS) $s_{jk}, 1 \leq k \leq h$ (the quality of service $s_{jk}$ is equal to 0 if there is no response from agent $A_j$). Following Marsh [10], we define for each agent an upper and a lower threshold for trust.

**Definition 2.** For each agent $A_i$, there are two thresholds $\omega_i$ and $\Omega_i$, where $0 \leq \omega_i \leq 1$, $0 \leq \Omega_i \leq 1$, and $\omega_i \geq \Omega_i$.

We use $f(x_k)$ to denote the probability that a particular value $x_k$ of quality of services from agent $A_j$ happens, where $x_k \in \{0.0, 0.1, \ldots, 1.0\}$. $\sum_{x_k=\omega_i}^{1} f(x_k)$ indicates the possibility that agent $A_i$ trusts agent $A_j$ and will cooperate with $A_j$; $\sum_{0}^{x_k=\Omega_i} f(x_k)$ indicates the possibility that agent $A_i$ mistrusts $A_j$ and will defect against $A_j$.

**Definition 3.** Given a series of responses from agent $A_j$, $S_j = \{s_{j1}, s_{j2}, \ldots, s_{jh}\}$, and the two thresholds $\omega_i$ and $\Omega_i$ of agent $A_i$, we can get the *bpa* toward agent $A_j$: $m(\{T\}) = \sum_{x_k=\omega_i}^{1} f(x_k)$, $m(\{\neg T\}) = \sum_{0}^{x_k=\Omega_i} f(x_k)$, and $m(\{T, \neg T\}) = \sum_{x_k=\Omega_i}^{x_k=\omega_i} f(x_k)$.

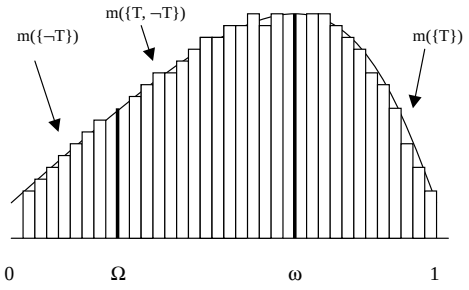


**Fig. 1.** Probability distribution of QoS of agent $A_j$

Returning to our original example, for a given subset $\hat{A}$ of $\Theta$, the *belief function* $Bel(\hat{A})$ is defined as the sum of all the belief committed to the possibilities in $\hat{A}$. For example,

$$Bel(\{T, \neg T\}) = m(\{T\}) + m(\{\neg T\}) + m(\{T, \neg T\}) = 1$$

For individual members of $\Theta$ (in this case, $T$ and $\neg T$), $Bel$ and $m$ are equal. Thus

$$Bel(\{T\}) = m(\{T\}) = 0.8, \text{ and } Bel(\{\neg T\}) = m(\{\neg T\}) = 0$$

A subset $\hat{A}$ of a frame $\Theta$ is called a *focal element* of a belief function $Bel$ over $\Theta$ if $m(\hat{A}) > 0$. Given two belief functions over the same frame of discernment

but based on distinct bodies of evidence, *Dempster's rules of combination* enables us to compute a new belief function based on the combined evidence. For every subset $\hat{A}$ of $\Theta$, Dempster's rule defines $m_1 \oplus m_2(\hat{A})$ to be the sum of all products of the form $m_1(X)m_2(Y)$, where $X$ and $Y$ run over all subsets whose intersection is $\hat{A}$. The commutativity of multiplication ensures that the rule yields the same value regardless of the order in which the functions are combined.

**Definition 4.** Suppose $Bel_1$ and $Bel_2$ are *belief functions* over the same frame $\Theta$, with basic probability assignments $m_1$ and $m_2$, and focal elements $\hat{A}_1, \ldots, \hat{A}_k$, and $\hat{B}_1, \ldots, \hat{B}_l$, respectively (here $\phi$ is the empty set). Suppose

$$\sum_{i,j,\hat{A}_i \cap \hat{B}_j = \phi} m_1(\hat{A}_i)m_2(\hat{B}_j) < 1$$

Then the function $m : 2^\Theta \mapsto [0,1]$ defined by $m(\phi) = 0$, and

$$m(\hat{A}) = \frac{\sum_{i,j,\hat{A}_i \cap \hat{B}_j = \hat{A}} m_1(\hat{A}_i)m_2(\hat{B}_j)}{1 - \sum_{i,j,\hat{A}_i \cap \hat{B}_j = \phi} m_1(\hat{A}_i)m_2(\hat{B}_j)}$$

for all non-empty $\hat{A} \subset \Theta$ is a basic probability assignment. [13]

The belief function given by $m$ is called the *orthogonal sum* of $Bel_1$ and $Bel_2$ and is denoted $Bel_1 \oplus Bel_2$. Let us now look at how beliefs obtained from two separate agents are combined. Suppose

$$m_1(\{T\}) = 0.8,\ m_1(\{\neg T\}) = 0,\ m_1(\{T, \neg T\}) = 0.2$$
$$m_2(\{T\}) = 0.9,\ m_2(\{\neg T\}) = 0,\ m_2(\{T, \neg T\}) = 0.1$$

Then $m_{12}$ is obtained as follows:

| | $m_2(\{T\})$ 0.9 | $m_2(\{T, \neg T\})$ 0.1 |
|---|---|---|
| $m_1(\{T\})$ 0.8 | $\{T\}$ 0.72 | $\{T\}$ 0.08 |
| $m_1(\{T, \neg T\})$ 0.2 | $\{T\}$ 0.18 | $\{T, \neg T\}$ 0.02 |

The new belief committed to $T$ is obtained by summing all the components committed to $T$:

$$m_{12}(\{T\}) = 0.72 + 0.18 + 0.08 = 0.98$$
$$m_{12}(\{\neg T\}) = 0$$
$$m_{12}(\{T, \neg T\}) = 0.02$$

Next suppose that one piece of the evidence confirms $T$, while the other disconfirms $T$. That is, we have the following situation:

$$m_1(\{T\}) = 0.8,\ m_1(\{\neg T\}) = 0,\ m_1(\{T, \neg T\}) = 0.2$$
$$m_2(\{T\}) = 0,\ m_2(\{\neg T\}) = 0.9,\ m_2(\{T, \neg T\}) = 0.1$$

Then $m_{12}$ is obtained as follows:

| | $m_2(\{\neg T\})$ 0.9 | $m_2(\{T, \neg T\})$ 0.1 |
|---|---|---|
| $m_1(\{T\})$ 0.8 | $\phi$ 0.72 | $\{T\}$ 0.08 |
| $m_1(\{T, \neg T\})$ 0.2 | $\{\neg T\}$ 0.18 | $\{T, \neg T\}$ 0.02 |

In this case, 0.72 of our belief is committed to the empty set. Since there are no possibilities in this set, the belief in our other sets must be normalized to 1. This yields

| | $m_2(\{\neg T\})$ 0.9 | $m_2(\{T, \neg T\})$ 0.1 |
|---|---|---|
| $m_1(\{T\})$ 0.8 | $\phi$ 0 | $\{T\}$ 0.29 |
| $m_1(\{T, \neg T\})$ 0.2 | $\{\neg T\}$ 0.64 | $\{T, \neg T\}$ 0.07 |

The new belief committed to $T$ is obtained as follows:

$$m_{12}(\{T\}) = 0.29$$
$$m_{12}(\{\neg T\}) = 0.64$$
$$m_{12}(\{T, \neg T\}) = 0.07$$

## 3 Our Approach

To better understand the notion of trust in electronic communities, let's discuss the famous prisoners' dilemma [1]. The prisoner's dilemma arises in a non-cooperative game with two agents. The agents have to decide whether to *cooperate* or *defect* from a deal. The payoffs in the game are such that both agents would benefit if both cooperate. However, if one agent were to try to cooperate when the other defects, the cooperator would suffer considerably. This makes the locally rational choice for each agent to be to defect, thereby leading to a worse payoff for both agents than if both were to cooperate.

Clearly, if the agents trusted each other, they could both cooperate and avert the situation where both suffer. Such trust can only build up in a setting where the agents have to repeatedly interact with each other. In our present domain, cooperation can be cast as delivering the desired quality of service. When agents have to engage in multiple interactions with others, it is rational for them to try to cooperate. A reputation mechanism sustains such cooperation, because the good agents are rewarded by society whereas the bad agents are penalized. Both the rewards and penalties from a society are greater than from an individual.

### 3.1 TrustNet

In our approach, agent $A_i$ evaluates the trustworthiness of agent $A_j$ based on (1) its direct observations of $A_j$ *as well as* (2) the belief ratings of $A_j$ as given by $A_j$'s neighbors (We also call the neighbors here as the *witnesses* of agent $A_j$). The second aspect makes our approach a social one and enables information about reputations to propagate through the network. A *TrustNet* encodes how agents estimate the reputation of other agents that they have not met before.

**Definition 5.** A referral to agent $A_j$ returned from agent $A_i$ is defined as $r_{\langle A_i, A_j \rangle}$, where $A_i$ is the source and $A_j$ the destination of the referral.

Given a set of referrals $R$, we define a numbering as a bijection that assigns each referral $r_{\langle A_i, A_j \rangle}$ a unique number in $\{1, 2, \ldots, n\}$ according to the sequence of returning from other agents, denoted as $r_i$, where $1 \leq i \leq n$.

**Definition 6.** Let $Q$ be a query from agent $A_i$. Assume that after $l$ referrals, agent $A_j$ returns a service. The entire referral chain in this case would be $\langle A_i, A_{i+1}, \ldots, A_{j-1}, A_j \rangle$, where $l$ is the length of the referral chain.

When searching for a potential witness in a social network, usually the further removed a witness is from the requester, the less likely the witness will respond. Similarly, the more steps away from the requester, the less accurate of the referrals provided. In our experiment we set a bound of 6 for the length of any referral chain.

**Definition 7.** In order to evaluate the trustworthiness of agent $A_g$, the requester agent $A_r$ may construct a TrustNet $TN$ which is defined as a directed graph $TN(\Lambda, R, Ar, Ag)$, where $\Lambda$ is a finite set of agents $\{A_1, \ldots, A_N\}$, and R is a set of referrals $\{r_1, \ldots, r_n\}$.

So given a series of referrals $\{r_1, r_2, \ldots, r_n\}$, the requester agent $A_r$ will append each referral $r_i$ to the TrustNet $TN$. In our experiment we use an *adjacency list* to represent the TrustNet. The adjacency list consists of an array *adj* of $N$ lists, one for each agent in $TN$. For each $A_i \in \Lambda$, the adjacency list $adj(A_i)$ contains the name and its belief functions towards each agents $A_j$ such that there is a referral $r_k$ from $A_i$ to $A_j$, where $1 \leq k \leq n$.

### 3.2 Incorporating Testimonies from Different Witnesses

Traditional approaches either ignore the social aspects altogether or employ a simplistic approach that directly combines the ratings assigned by different sources. However, such approaches do not consider the trustworthiness of the witnesses themselves. Clearly, the weight assigned to a rating should depend on the reputation of the rater.

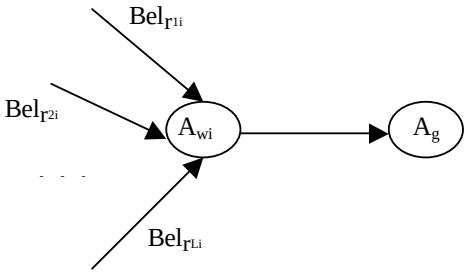


**Fig. 2.** The reputation of the witness $A_{w_i}$

Suppose agent $A_r$ wants to evaluate the trustworthiness of agent $A_g$. $\{w_1, \ldots, w_k\}$ are a group of witnesses towards agent $A_g$. For any witness $w_i$, $\{r_{1i}, \ldots, r_{Li}\}$ are a series of referrals to the witness $w_i$. $Bel_{r_{li}}(\{T_{w_i}\})$, $Bel_{r_{li}}(\{\neg T_{w_i}\})$ and $Bel_{r_{li}}(\{T_{w_i}, \neg T_{w_i}\})$ are belief ratings to $w_i$ by the referral $r_{li}$.

**Definition 8.** Suppose in a TrustNet $TN(A, R, A_r, A_g)$, agent $A_{w_i}$ is one of the witnesses of agent $A_g$ and $\{r_1, r_2, \ldots, r_L\}$ are a series of referrals to agent $A_{w_i}$ (Figure 2). Then the cumulative belief for agent $A_{w_i}$ is computed as

$$Bel_{r_i} = Bel_{r_{1i}} \oplus Bel_{r_{2i}} \oplus \ldots \oplus Bel_{r_{Li}}$$

and the reputation of the witness $A_{wi}$ is defined as

$$\Gamma(A_{w_i}) = Bel_{r_i}(\{T_{w_i}\})$$

**Input:** Given a series of referrals $\{r_1, r_2, \ldots, r_n\}$, and for each referral $r_{\langle Ai, Aj\rangle}$, there is a *bpa* assigned to agent $A_j$ by agent $A_i$.
**Output:** The *bpa* of agent $A_g$ given by each witness $A_{wi}$, and the reputation of witness $A_{wi}$.

1. (**Forward**) For each referral $r_{\langle Ai, Aj\rangle}$, append it to the end of adj($A_i$) if exists, otherwise initialize adjacency list $A_i$ and append it to adj($A_i$).
2. (**Backward**) Reverse the TrustNet TN, and find the node $A_g$ in the reversed TrustNet, for each node in adj($A_g$), we name it as one of the witness $A_{wi}$.
3. For each witness $A_{wi}$, find a list of agents in adj($A_{wi}$) in the reversed TrustNet and compute the reputation of $A_{wi}$ using *Dempster's rule of combination*.
4. Return the *bpa* of agent $A_g$ given by each witness $A_{wi}$, and the reputation of $A_{wi}$.

**Fig. 3.** Testimony propagation algorithm

We now show how testimonies from different agents can be incorporated into the belief rating by a given agent.

**Definition 9.** For agent $A_r$, the reliability of a testimony $e_i$ from agent $A_{w_i}$ about agent $A_g$ is computed as

$$Bel_{e_i}(\{T_{A_g}\}) = \Gamma(A_{w_i}) Bel_{w_i}(\{T_{A_g}\});$$
$$Bel_{e_i}(\{\neg T_{A_g}\}) = \Gamma(A_{w_i}) Bel_{w_i}(\{\neg T_{A_g}\});$$
$$Bel_{e_i}(\{T_{A_g}, \neg T_{A_g}\}) = 1 - Bel_{e_i}(\{T_{A_g}\}) - Bel_{e_i}(\{\neg T_{A_g}\})$$

where $Bel_{w_i}(\{T_{A_g}\})$ and $Bel_{w_i}(\{\neg T_{A_g}\})$ are the belief ratings to agent $A_g$ given by witness $A_{w_i}$.

Therefore, agent $A_r$ will update its belief rating of agent $A_g$ as follows:

**Definition 10.** Given a set of testimonies $\Delta = \{e_1, e_2, \ldots, e_H\}$, agent $A_r$ will update its trust value of agent $A_g$ as follows

$$Bel_{A_r} = Bel_{A_r} \oplus Bel_{e_1} \oplus \ldots \oplus Bel_{e_H}$$

and the updated reputation of agent $A_g$ is

$$\Gamma(A_g) = Bel_{A_r}(\{T_{A_g}\})$$

Figure 3 summarizes the process of testimony propagation. The requester agent $A_r$ will update its trust upon the testimony from each witness.

## 4 Experimental Results

In our simulation, we treat the agent and user simply as just the agent. Each agent has an *interest* vector, an *expertise* vector, and several *neighbor* models. In general, the neighbor models depend on how many agents know the given agent, how many agents it knows, which community it belongs to, and so on. In our case, the neighbor models kept by an agent are the given agent's representation of the other agents' expertise and belief rating. We introduce a probability between 0 and 1 to model the responsiveness of each agent $A_i$, called *responsiveness factor*, and denoted as $F(A_i)$. Agent $A_i$ will generate an answer from his *expertise* vector upon a query with the probability $F(A_i)$ even when there is a good match between the query and his expertise vector.

In each simulation cycle, we randomly designate an agent to be the "requester agent." The queries are generated as vectors by perturbing the interest vector of the requester agent. When an agent receives a query, it will try to answer it based on its expertise vector, or refer to other agents it knows. The originating agent collects all possible referrals, and continues the process by contacting some of the suggested referrals. At the same time, it changes its models for other agents.

Our experiments are based on the simulation testbed we have developed so far in the expertise location setting (The only difference is in the referring process: a referral is given only if the referral agent places some trust in the agent being referred.), which involves between 20 and 60 agents with interest and expertise vectors of dimension 5. Each agent keeps the latest 10 responses from another agent if there are more than 10 responses. The agents are limited in the number of neighbors they may have - in our case the limit is 4. However, each agent may keep track of more peers than his neighbors (others will be put in his *cache*). Periodically he decides which peers to be kept as neighbors, i.e., which are worth remembering.

### 4.1 Metrics

We now define some useful metrics in which to intuitively capture the results of our experiments.

**Definition 11.** Suppose there are $L$ agents who know agent $A_i$ (we say that agent $A_j$ knows agent $A_i$ if and only if agent $A_i$ is a neighbor of agent $A_j$.), and $\{r_1, r_2, \ldots, r_L\}$ are a series of referrals to agent $A_i$. Then the cumulative belief of agent $A_i$ is computed as

$$Bel_r = Bel_{r_1} \oplus Bel_{r_2}, \ldots, \oplus Bel_{r_L}$$

and the reputation of agent $A_i$ is defined as $\Gamma(A_i) = Bel_r(\{T_{A_i}\})$. If $L = 0$ then $\Gamma(A_i) = 0$.

**Definition 12.** The average reputation of a group of agents is defined as:

$$\Pi = 1/N \sum_{i=1}^{N} \Gamma(A_i),$$

where N is the total number of agents in the group.

### 4.2 Bootstrapping

In the first simulation we evaluate the convergence speed of the algorithm. One has 60 agents with uniformly distributed responsiveness factors. Each agent starts with some random interest vectors, expertise vectors and 4 neighbors. For any two agents $A_i$ and $A_j$, $Bel_{A_i}(\{T_{A_j}\}) = Bel_{A_i}(\{\neg T_{A_j}\}) = 0$, $Bel_{A_i}(\{T_{A_j}, \neg T_{A_j}\}) = 1$ in the beginning. Then agents send queries, referrals, and responses to one another, all the while learning about each others' interest and expertise vectors. We assume that one has reached equilibrium when the average reputation of one agent converges to its real responsiveness factor.

Consider the following example. Assume agent $A_1$ is a cooperative agent with a responsiveness factor 0.75, and agent $A_{60}$ is a non-cooperative agent with a responsiveness factor 0.25. Their initial average reputations are zero. After 2,000 simulation cycles, we found that the average reputation of agent $A_1$ increased to a high level, and the average reputation of agent $A_{60}$ increased to a low level. The average reputation of the whole group agents increased rapidly in the beginning, but slowed down later. Figure 4 confirms our hypothesis.

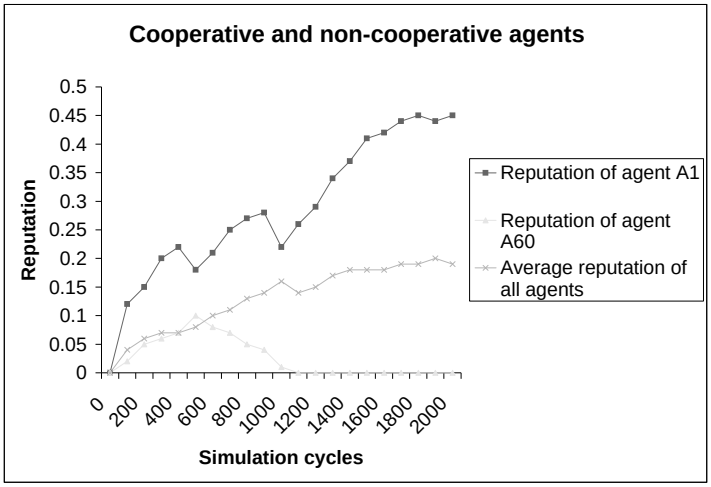


**Fig. 4.** Reputations of cooperative and non-cooperative agents in the bootstrapping stage

### 4.3 Reputation Buildup

Clearly, a social network will not remain stable for long, because agents will continually introduce and remove themselves from the network. In the second simulation, we show that a new agent $A_{61}$ who joins the electronic community at the simulation cycle 2000, behaves cooperatively with a responsiveness factor 1 until he/she reaches a high reputation value, and then starts abusing his/her reputation by decreasing his/her responsiveness factor to 0.25. Thus, his average reputation starts dropping because of his/her non-cooperative behavior (Figure 5).

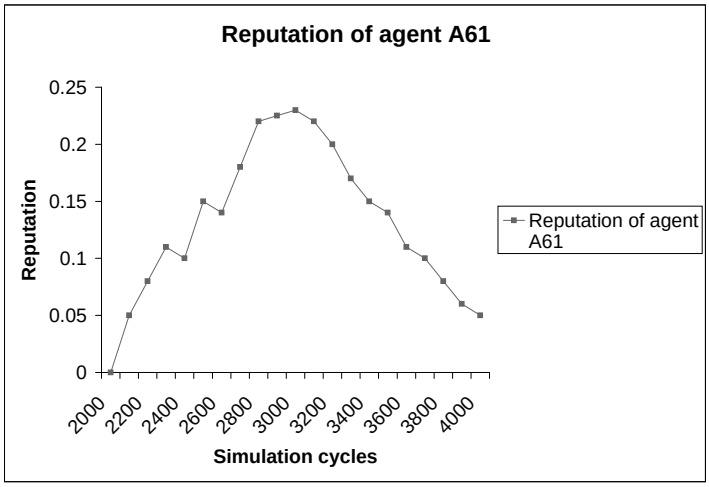


**Fig. 5.** Reputation buildup and crash of a new agent

### 4.4 Community Size

Usually there is a better chance to select a partner in a large (virtual) city of 300,000 people than in a small town of 3,000 people. On the other side, it is much easier to collect "bad" testimonies in a small town. We conjecture that the average reputation of an agent in a smaller group should change faster than that in a larger community.

Given two groups of agents, group1 and group2, with the number of agents 20 and 60, respectively. Suppose agent $A_{group1-1}$ and agent $A_{group2-1}$ are two cooperative agents in the beginning with the responsiveness factors 1. After a series of simulation cycles, i.e., 2,000, both of them decrease their responsiveness factor to 0.25. Thus, their average reputation starts dropping because of their non-cooperative behaviors. Figure 6 shows that the average reputation of agent $A_{group1-1}$ drops faster since it is in a smaller community.

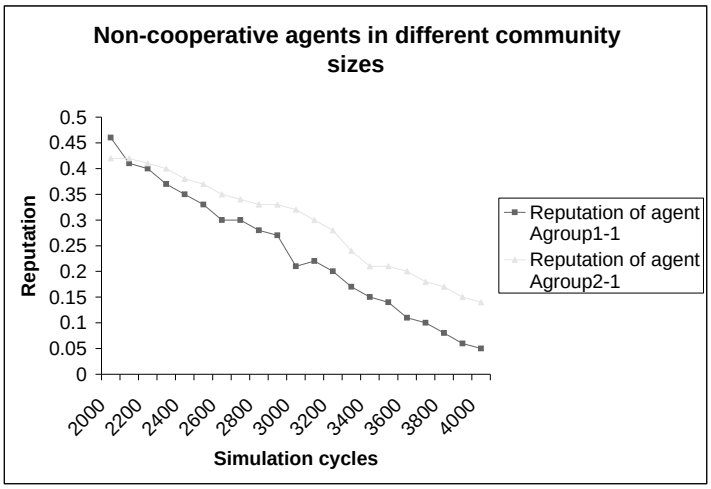


**Fig. 6.** Non-cooperative agents in different community sizes

## 5 Related Work

OnSale Exchange and eBay are important practical examples of reputation management. OnSale allows its users to rate and submit textual comments about sellers. The overall reputation of a seller is the average of the ratings obtained from his customers. In eBay, sellers receive feedback $(+1, 0, -1)$ for their reliability in each auction and their reputation is calculated as the sum of those ratings over the last six months. In OnSale, the newcomers have no reputation until someone rates them, while on eBay they start with zero feedback points. Both approaches require users to explicitly make and reveal their ratings of others. As a result, the users lose control to the central authority.

Some prototype approaches are relevant. Weaving a web of trust [8], and Kasbah [16] require that users give a rating for themselves and either have a central agency (direct ratings) or other trusted users (collaborative ratings). A central system keeps track of the users' explicit ratings of each other, and uses these ratings to compute a person's overall reputation or reputation with respect to a specific user. These systems require preexisting social relationships among the users of their electronic community. It is not clear how to establish such relationships and how the ratings propagate through this community.

Much theoretical work has been done in how to learn strategies of agents and how to react to a variety of behaviors [2, 12]. These approaches build on the success of Axelrod's experiments with round-robin tournaments of agent strategies in the Iterated Prisoner's Dilemma, where each strategy had to play against all other strategies [1]. Work in social psychology, however, has shown that selecting the right partners to play yields a better game performance compared to spending the same effort on *how* to play the game itself [6]. Especially in the context of open systems, there is at least some good chances to find an alternative partner.

Marsh presents a formalization of the concept trust [10]. His formalization considers only an agent's own experiences and doesn't involve any social mechanisms. Hence, a group of agents cannot collectively build up a reputation for others. A more relevant computational method is from *Social Interaction FrameWork* (SIF) [11]. In SIF, an agent evaluates the reputation of another agent based on direct observations as well through other *witnesses*. However, SIF does not describe how to find such witnesses, whereas in the electronic communities, deals are brokered among people who probably have never met before.

There has been much work on social abstractions for agents, e.g., [3, 5]. The initial work on this theme studied various of relationships among agents. There have been some studies of the aggregate behavior of social systems that is relevant to some of our tasks. More recent work on these themes has begun to look at the problems of deception and fraud [4]. However, the proposed approach goes significantly beyond their approach in the kinds of representations of trust, propagation algorithms, and formal analysis.

## 6 Conclusion

Trust and reputation management are becoming hot topics in agents and multiagent systems. Although we present our results in the context of electronic communities, our approach applies to multiagent systems in general. Most current multiagent systems assume benevolence, meaning that the agents implicitly assume that other agents are trustworthy and reliable. With the growth of network services, agents may find themselves confronted with deception and fraud. Approaches for explicit reputation management can help the agents finesse their interactions depending on the reputations of the other agents. The ability to deal with selfish, antisocial, or unreliable agents can lead to more robust multiagent systems.

Our present approach adjusts the ratings of agents based on their interactions with others. However, it does not fully protect against spurious ratings generated by malicious agents. It relies only on there being a large number of agents who offer honest ratings to override the effect of the ratings provided by the malicious agents. In future work, we plan to study the special problems of lying and rumors as well as of community formation. We also want to study the evolutionary situations where groups of agents consider rating schemes for other agents. The purpose is not only to study alternative approaches for achieving more efficient communities, but also to test if our mechanism is robust against invasion and, hence, more stable.

## References


1. Robert Axelrod. *The Evolution of Cooperation.* Basic Books, New York, 1984.
2. David Carmel and Shaul Markovitch. Exploration and adaptation in multiagent system: A model-based approach. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, pages 606–611, 1997.

3. Cristiano Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.
4. Cristiano Castelfranchi and Rino Falcone. Principle of trust for MAS: cognitive anatomy, social importance, and quantification. In *Proceedings of Third International Conference on MultiAgen t Systems*, pages 72–79, 1998.
5. Les Gasser. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence*, 47:107–138, 1991.
6. Nahoko Hayashi and Toshio Yamagishi. Selective play: choosing partners in an uncertain world. *Personality and Social Psychology Review*, 2:276–289, 1998.
7. David Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In *Uncertainty in Artificial Intelligence*, pages 167–196, 1986.
8. Rohit Khare and Adam Rifkin. Weaving a web of trust. *World Wide Web*, 2(3):77–112, 1997.
9. Henry E. Kyburg. Bayesian and non-bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.
10. P. Steven Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, Department of Computing Science and Mathematics, University of Stirling, April 1994.
11. Michael Schillo, Petra Funk, and Michael Rovatsos. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence*, 14:825–848, 2000.
12. Sendip Sen. Reciprocity: a foundational principle for promoting cooperative behavior among self-interested agents. In *Proceedings of the second International Conference on Multiagent Systems*, pages 315–321, 1996.
13. Glenn Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.
14. Glenn Shafer. Probability judgement in artificial intelligence and expert systems. *Statistical Science*, 2:3–44, 1987.
15. Bin Yu and Munindar P. Singh. A social mechanism of reputation management in electronic communities. In *Proceedings of Fourth International Workshop on Cooperative Information Agents*, pages 154–165, 2000.
16. Giorgos Zacharia and Pattie Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14:881–908, 2000.