An Evidential Model of Distributed Reputation Management

Bin Yu
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-7535, USA
byu@eos.ncsu.edu

Munindar P. Singh
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-7535, USA
singh@ncsu.edu

ABSTRACT

For agents to function effectively in large and open networks, they must ensure that their *correspondents*, i.e., the agents they interact with, are trustworthy. Since no central authorities may exist, the only way agents can find trustworthy correspondents is by collaborating with others to identify those whose past behavior has been untrustworthy. In other words, finding trustworthy correspondents reduces to the problem of distributed reputation management.

Our approach adapts the mathematical theory of evidence to represent and propagate the ratings that agents give to their correspondents. When evaluating the trustworthiness of a correspondent, an agent combines its local evidence (based on direct prior interactions with the correspondent) with the testimonies of other agents regarding the same correspondent. We experimentally studied this approach to establish that some important properties of trust are captured by it.

Keywords

distributed reputation management, belief functions, trust networks

1. INTRODUCTION

With the expansion of the Internet, people and services are called upon to interact with independent parties. This is so in application areas such as e-commerce, knowledge sharing, and even game playing. We envision agents being used to assist in such interactions for each relevant party or *principal*. Because the parties are autonomous and potentially subject to different administrative and legal domains, it is important that each agent be able to identify trustworthy parties or *correspondents* with whom its principal should interact and untrustworthy correspondents with whom its principal should avoid interaction.

Trust has recently begun to attract attention in the multiagent systems community. Previously, multiagent systems were restricted to those in which the participants were assumed to be fully cooperative or those in which the rules of engagement were set in such a manner that the participants were competitive, e.g., in computational markets, but their trustworthiness never came into play. In many real-life settings, agents will not necessarily cooperate with one another. In such settings, agents must be able to deal with fraud and deception. The only way to do so is by developing a robust notion of trust. Ultimately, we imagine trust being a building-block concept for multiagent architectures with trust management services attracting as much importance as knowledge representation in today's multiagent systems.

Further, since we consider large distributed systems of autonomous and heterogeneous agents, it is generally inadvisable to assume that there are universally accepted trustworthy authorities, who can declare the trustworthiness of different agents. Even the authorities, such as they are, may not be considered trustworthy by all. This is a fundamental limitation of traditional methods, the so-called "hard security" techniques based on passwords, keys, and digital certificates. What hard security can ensure, at most, is that the given agent obtained credentials from another credentialed party, ultimately one that is trusted by fiat (i.e., PGP-style web of trust, or X.509-style certifying authority trees). This opens up two objections. First, no such trusted third party might exist. Second, even knowing the identity of a correspondent does not justify placing trust in it, because you have no reason to believe it will act in your interest.

Trust is more than creating, acquiring, and distributing certificates. A party might be authenticated and authorized, but this does not ensure that it exercises its authorizations in a way that is expected. Our approach, therefore, is to develop a social mechanism, a so-called "soft security" technique. We begin with a perfect peer-to-peer model of a multiagent system and study how to automatically and efficiently detect non-cooperative agents.

In our settings, all the agents are in principle equal, and agents will necessarily form ratings of others that they interact with. The trustworthiness of a correspondent is viewed as the expectation of cooperative behavior from that correspondent. But to evaluate the trustworthiness of a correspondent, especially prior to any frequent direct interactions, the agents will have to rely on social mechanisms for incorporating the knowledge of other agents. In other words, we model trustworthiness as a reputation for good behavior; an agent will place trust in a correspondent based on the latter's reputation combined with the outcomes of its direct interactions, if any, with the correspondent.

AAMAS'02, July 15-19, 2002, Bologna, Italy.

Let us now describe our conceptual model. We consider interactions in electronic communities, where agents assist and represent principal entities, such as people and businesses. Each agent attempts to determine the trustworthiness of a given correspondent based on its own prior interactions with that correspondent in conjunction with the testimonies given by other trustworthy agents or *witnesses* who have interacted with the same correspondent. The mechanism for finding the right witnesses relies upon referrals being generated by the agents who are directly acquainted with the given agent. We develop an evidential model of reputation management based on the Dempster-Shafer theory. In this scheme, if no information about the agent is available, it has no reputation at all. It should be noted that there is a difference between having a bad reputation and no reputation at all.

The proposed approach builds on our work on referral networks [24]. An agent-based referral network is a multiagent system whose member agents give referrals to one another (and are able to follow referrals received from other agents). To do so effectively presupposes certain representation and reasoning capabilities on the part of each agent. Each agent has a set of *acquaintances*, a subset of which are identified as its *neighbors*. The neighbors are the agents that the given agent would contact and the agents that it would point (refer) others to. An agent maintains a model of each acquaintance. This model includes the agent's abilities to act in a trustworthy manner and to refer to other trustworthy agents, respectively. The first ability we term *expertise* and the second ability we term *sociability*.

Each agent may modify its models of its acquaintances, potentially based on its direct interactions with the acquaintance, based on interactions with agents referred to by the acquaintance, and based on ratings of this acquaintances received from other agents. More importantly, in our approach, agents can adaptively choose their neighbors, which they do every so often from among their current acquaintances. An agent may estimate the trustworthiness of a given party based on its own past interactions or may consult other trusted agents who have directly interacted with the same party. These agents are termed *witnesses*. An agent can find the right witnesses by seeking and following referrals from its neighbors.

The rest of this paper is organized as follows. Section 2 introduces our technical approach, giving the key definitions for local trust rating and propagation through referrals. Section 3 presents our experimental results. We summarize some related work in reputation management in Section 4. Section 5 concludes our paper with a discuss of the main results and directions for future research.

2. REPUTATION MANAGEMENT

Traditionally, there are two main view of trust. The cognitive view postulates trust as made up of underlying beliefs. That is, trust is a function of the value of these beliefs [7]. The mathematical view ignores the role of underlying beliefs and uses a (scalar) metric to model a subjective probability with which an agent will perform a particular action [16]. Our approach represents an enhancement of the mathematical view. While we do not directly consider the specific cognitive notions that might apply in judging trustworthiness, our model is richer than traditional mathematical models in accommodating testimonies and in considering both trustworthiness and untrustworthiness.

The idea that the rating of a correspondent be based on direct observations as well the ratings assigned by other sources is well-known in the literature on reputation. However, some important challenges must be addressed.

- How does an agent rate a correspondent based on their direct interactions? Our approach does so by capturing the ratings of the last several interactions, which are recorded in the given agent's *history*.
- How does the agent find the right witnesses? Our approach applies a process of referrals through which agents help one another find witnesses.
- How does the agent systematically incorporate the testimonies of those witnesses? Our approach includes the TrustNet representation through which the ratings can be combined in a principled manner.

In particular, our approach has an advantage over other approaches in terms of the second and third challenges. These are discussed in Section 4.

Before we can describe the three elements of our approach, we must consider a representational framework over which they are layered. There are three main choices in this regard.

- Certainty factors are scalar values to represent an agent's belief ratings about another. The certainty factor model provides a mechanism for combining testimonies. However, certainty factors do not represent measures of absolute belief. Rather, they are meant to represent changes in belief [12].
- The agents' inherent uncertainty about their correspondent can be expressed within a Bayesian framework [17]. The Bayesian approach offers a mechanism for combining evidence. However, the Bayesian approach is limited in not being able to distinguish between lack of belief and disbelief. Lack of belief must be modeled through the artificial construct of equiprobable probability distributions.
- The Dempster-Shafer calculus handles the notion of evidence pro and con explicitly [15]. There is no causal relationship between a hypothesis and its negation, so lack of belief does not imply disbelief. Rather, lack of belief in any particular hypothesis implies belief in the set of all hypotheses, which is referred to as the state of uncertainty. This leads to the intuitive process of narrowing a hypothesis [11], in which initial uncertainty is replaced with belief or disbelief as evidence is accumulated.

For the above reasons, we uses the Dempster-Shafer theory of evidence as the underlying computational framework.

2.1 Dempster-Shafer Theory

We now introduce the key concepts of the Dempster-Shafer approach. Let T mean that the given agent considers a given correspondent to be trustworthy. A *frame of discernment* $\Theta = \{T, \neg T\}$ is the set of propositions under consideration.

DEFINITION 1. *Let Θ be a frame of discernment. A basic probability assignment (bpa) is a function $m : 2^{\Theta} \mapsto [0, 1]$ where (1) $m(\phi) = 0$, and (2) $\sum_{\hat{A} \subset \Theta} m(\hat{A}) = 1$.*

Thus $m(\{T\}) + m(\{\neg T\}) + m(\{T, \neg T\}) = 1$. A bpa is similar to a probability assignment except that its domain is the subsets and not the members of Θ. The sum of the bpa's of the singleton subsets of Θ may be less than 1. For example, given the assignment of $m(\{T\}) = 0.8$, $m(\{\neg T\}) = 0$, $m(\{T, \neg T\}) = 0.2$, we have $m(\{T\}) + m(\{\neg T\}) = 0.8$, which is less than 1.

For a subset $\hat{A}$ of Θ, the *belief function* $\mathsf{Bel}(\hat{A})$ is defined as the sum of the beliefs committed to the possibilities in $\hat{A}$. For example,

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

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

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

2.2 Local Trust Ratings

When agent A_i is evaluating the trustworthiness of agent A_j, there are two components to the evidence. The first component is the services offered by agent A_j. The second component is the testimonies from other agents in case A_i has no transactions with A_j before. 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) $s_{jk}, where 1 \leq k \leq H$ (note that the quality of service was rated by users. s_{jk} is equal to 0 if there is no response from agent A_j).

Following Marsh [16], we define for each agent an upper and a lower threshold for trust. For each agent A_i, there are two thresholds ω_i and Ω_i, where $0 \leq \omega_i \leq \Omega \leq 1$. 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 $x_k \in \{0.0, 0.1, \ldots, 1.0\}$. For example, given a specific value x_k, there are three services with that quality in the latest H responses, then $f(x_k) = 3/H$. But if there are less than H responses available, say h, then $f(x_k) = 3/h$.

Definition 2. *Given a series of responses from agent A_j, $S_j = \{s_{j1}, s_{j2}, \ldots, s_{jH}\}$, and the two thresholds ω_i and Ω_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 $m(\{T, \neg T\}) = \sum_{x_k=\omega_i}^{x_k=\Omega_i} f(x_k)$.*

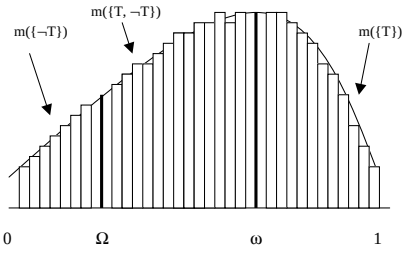

Figure 1: Distribution of trust ratings

2.3 Combining Belief Functions

When an agent has not interacted often enough with a correspondent, it must seek the testimonies of other witnesses. Next we discuss how to combine such evidence.

A subset $\hat{A}$ of a frame Θ is called a *focal element* of a belief function Bel over Θ if $m(\hat{A}) > 0$. Given two belief functions over the same frame of discernment but based on distinct bodies of evidence, Dempster's rule of combination enables us to compute a new belief function based on the combined evidence. For every subset $\hat{A}$ of Θ, 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 3. *Let Bel_1 and Bel_2 be belief functions over Θ, 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. Suppose*

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

Then the function $m : 2^{\Theta} \mapsto [0, 1]$ that is defined by

$$m(\phi) = 0, \text{ and}$$
$$m(\hat{A}) = \frac{\sum_{i,j,\hat{A}_i \cap \hat{B}_j = \hat{A}} m_1(\hat{A}_i) m_2(\hat{B}_j)}{1 - \sum_{i,j,\hat{A}_i \cap \hat{B}_j = \phi} m_1(\hat{A}_i) m_2(\hat{B}_j)}$$

for all non-empty $\hat{A} \subset \Theta$ is a basic probability assignment [22].

Bel, the belief function given by m, is called the *orthogonal sum* of Bel_1 and Bel_2. It is written $\mathsf{Bel} = \mathsf{Bel}_1 \oplus \mathsf{Bel}_2$. 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:

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

2.4 Deciding Whether to Trust

It helps to distinguish between two kinds of beliefs: *local belief* and *total belief*. An agent's local belief about a correspondent is from direct interactions with it and can be propagated to others upon request. An agent's total belief about a correspondent combines the local belief (if any) with testimonies received from any witnesses. Total belief can be used for deciding whether the correspondent is trustworthy. To prevent non-well-founded cycles, we restrict agents from propagating their total beliefs. However, in principle, the necessary information underlying a total belief can be obtained by the requesting agent from the original witnesses.

Agent A_r models all of the information he gets about A_g using belief functions, and then decides whether to cooperate with A_g. Figure 2 shows the whole process. We argue that the total belief is needed only if they have no transaction before. One intuition is the trustworthiness is people or agent specific. Agent A may trust agent B, but not C.

To evaluate the trustworthiness of agent A_g, agent A_r will check if A_g is one of its acquaintances. If so, A_r will use its existing local belief to evaluate the trustworthiness of A_g. Otherwise, A_r will query its neighbors about A_g. When an agent receives a query about A_g's trustworthiness, it will check if A_g is one of its acquaintances. If yes, it will return

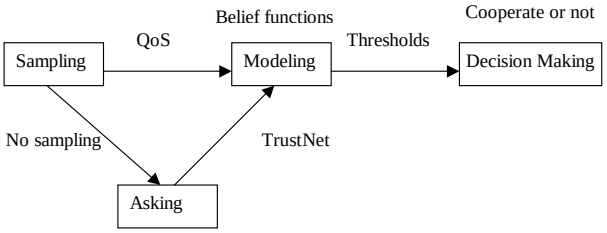

Figure 2: The process of deciding whether to cooperate with another agent

the information about A_g; otherwise, it will return referrals to A_r. A_r, if it chooses, can then query the referred agents.

A referral r to agent A_j returned from agent A_i is written as $\langle A_i, A_j \rangle$. A series of referrals makes a referral chain. Observing that shorter referral chains are more likely to be fruitful and accurate [13] and to limit the effort expended in pursuing referrals, we define *depthLimit* as the bound on the length of any referral chain.

The referral process begins with A_r initially contacting a neighbor A_i, who then gives a referral, and so on. The process terminates in success when a rating is received and in failure when the depthLimit is reached or when it arrives at an agent neither gives an answer rating nor a referral.

To simplify the notation, we refer to the initial contact $\langle A_r, A_i \rangle$ as a referral as well. For simplicity, a chain is written as $\langle A_0, A_1, \ldots, A_k \rangle$, where A_0 is the querying agent and every agent A_i for $i < k$ gives a referral to agent A_{i+1}.

2.5 TrustNet

Now suppose A_r wants to evaluate the trustworthiness of A_g, after a series of l referrals, a testimony about agent A_g is returned from agent A_j. Let the entire referral chain in this case be $\langle A_r, \ldots, A_j \rangle$, with length l. A TrustNet is a representation built from the referral chains produced from A_r's query. It is used to systematically incorporate the testimonies of the various witnesses regarding a particular correspondent.

DEFINITION 4. *A TrustNet $TN(A_r, A_g, \mathbf{A}, R)$ is a directed graph, where $\mathbf{A}$ is a finite set of agents $\{A_1, \ldots, A_N\}$, and R is a set of referrals $\{r_1, \ldots, r_n\}$.*

Given a series of referrals $\{r_1, r_2, \ldots, r_n\}$, the requester A_r constructs a TrustNet TN by incorporating the each referral $r_i = \langle A_i, A_j \rangle$ into TN. A_r adds r_i to R if and only if $A_j \notin \mathbf{A}$ and $depth(A_i) \leq depthLimit$. Figure 3 shows how the testimonies propagate through the TrustNet.

Suppose agent A_r wants to evaluate the trustworthiness of agent A_g, and $\{w_1, \ldots, w_L\}$ are a group of witnesses towards agent A_g. We now show how testimonies from witnesses can be incorporated into the trust rating of a given agent. Let τ_{A_i} and π_{A_i} be the belief functions corresponding to agent A_i's local and total beliefs, respectively.

DEFINITION 5. *Given a set of witnesses $\Delta = \{w_1, w_2, \ldots, w_L\}$, agent A_r will update its total belief value of agent A_g as follows*

$$\pi_{A_r} = \tau_{w_1} \oplus \ldots \oplus \tau_{w_L}$$

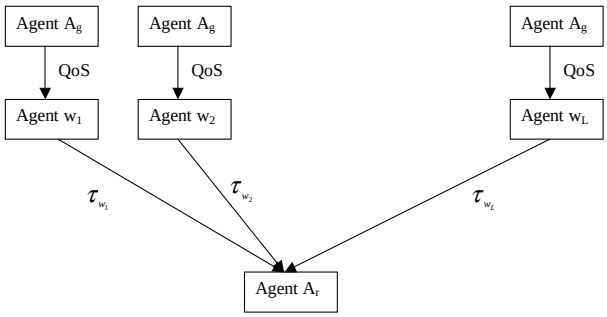

Figure 3: Testimony propagation through witnesses

Input: Suppose A_r is the request agent, set Λ is the agents being visited. Given a series of referrals $\{r_1, r_2, \ldots, r_n\}$, and for each referral $r_k = \langle Ai, Aj \rangle$, there is a *bpa* assigned to agent A_j by agent A_i.
Output: The *bpa* of agent A_g from a series of testimony given witnesses.

1. If $A_j \notin \Lambda$ then
   - If $A_j = A_g$, then append r_k to the TrustNet, and store the testimony from A_i (one of the witness to A_g), otherwise
   - If depth(A_j) < six, append r_k to the TrustNet, and send a query to A_j, otherwise
   - ignore the referral r_k.
2. If $A_j \in \Lambda$, ignore the referral r_k.
3. Compute the *bpa* of agent A_g using *Dempster's rule of combination*, and return the *bpa* of agent A_g.

Figure 4: Testimony propagation algorithm

Figure 4 summarizes the testimony propagation algorithm, where the depthLimit is six.

Next we consider the situation where A_r needs to compute its total belief regarding A_g.

- **Case 1:** A_r has interacted with A_g. A_r will trust A_g if $\tau_{A_r}(\{T_{A_g}\}) - \tau_{A_r}(\{\neg T_{A_g}\}) \geq \rho$, where ρ is a threshold for trustworthiness and $0 < \rho < 1$.
- **Case 2:** A_r has not interacted with A_g. A_r computes its total belief about A_g for decision making. A_r will trust A_g if $\pi_{A_r}(\{T_{A_g}\}) - \pi_{A_r}(\{\neg T_{A_g}\}) \geq \rho$.
- **Case 3:** A_g is totally new to the society. In this case $\pi_{A_r}(\{T_{A_g}\}) = \pi_{A_r}(\{\neg T_{A_g}\} = 0$, and $\pi_{A_r}(\{T_{A_g}, \neg T_{A_g}\}) = 1$. If we set the value of ρ equal to 0, then A_r may cooperate with the new agent A_g.

3. EXPERIMENTAL RESULTS

Our experiments are based on a simulation testbed we have developed, which involves between 100 and 500 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, here 4. The length of each referral chain is limited to 6. Moreover, we introduce

a probability between 0 and 1 to model the cooperativeness of each agent A_i, denoted as C_{A_i}. Agent A_i will generate an answer from his expertise vector upon a query with the probability C_{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. The queries are generated as vectors by perturbing the interest vector of the requesting agent. When an agent receives a query, it will try to answer the query based on its expertise vector, or refer to other agents it knows. The originating agent collects all suggested referrals, and continues the process by contacting some of them. At the same time, each agent may keep track of more acquaintances than are its neighbors. In our case the size is 12). Periodically it decides which acquaintances to be kept as neighbors, i.e., which are worth remembering.

Recently Prietula and Carley [18] studied the effects of agent trust in a simulated organization task. Schillo *et al.* tested the performance of two groups of agents with different settings for honesty and dishonesty, and altruism and egotism [21]. We previously studied the reputation changes of a "non-cooperative" agent and a new agent in a group of 20 to 60 agents. However, our previous work did not consider the different environments of the agents. Using our simulation testbed, we studied the role of trust in the following three settings: reputation buildup, community size and ratio of non-cooperative agents.

3.1 Metrics

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

DEFINITION 6. *Suppose* $\{w_1, \ldots, w_L\}$ *are exactly* L *agents whose neighbors include* A_i. *Then* β_{A_i}, *the cumulative belief regarding agent* A_i *is computed as*

$$\beta_{A_i} = \tau_{w_1} \oplus \tau_{w_2}, \ldots, \oplus \tau_{w_L}$$

and the reputation of agent A_i *is defined as*

$$\Gamma(A_i) = \beta_{A_i}(\{T_{A_i}\}) - \beta_{A_i}(\{\neg T_{A_i}\}.$$

If $L = 0$ *then* $\Gamma(A_i) = 0$.

DEFINITION 7. *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.*

3.2 Bootstrapping

Following Watts and Strogatz [25], we begin from a ring but, unlike them, we allow for edges to be *directed*. We use a regular ring with 100 nodes, and 4 edges per node (to its nearest neighbors) as a starting point for the experiment. Since the simulation does not involve real users, the quality of service (QoS) is estimated based on how close of the answer to the interest vector.

The cooperativeness for each agent is set to 1 if not specified. For any two agents A_i and A_j, $\tau_{A_i}(\{T_{A_j}\}) = \tau_{A_i}(\{\neg T_{A_j}\}) = 0$, $\tau_{A_i}(\{T_{A_j}, \neg T_{A_j}\}) = 1$ in the beginning. We have a relative low value for the lower threshold Ω_i, since we want to model the cooperativeness of agents. For each agent A_i, we have $\omega_i = 0.1$ and $\Omega_i = 0.5$. Whether a neighbor will be kept as a neighbor depends on how close of the neighbor's expertise vector to the agent's interest vector. After every 100 round, we compute the reputation for the single agent, and whole group using the metrics defined above. The computation is not counted in the simulation cycle.

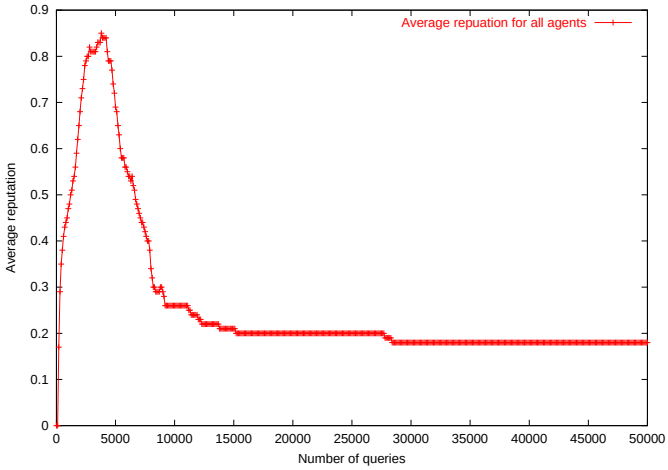

Figure 5: Reputations change of all agents in the bootstrapping stage (from a regular ring)

In the first simulation we evaluate the convergence of our approach. We hypothesize that the average reputation of all agents reaches equilibrium when each agent finds the right neighbors of itself. Consider the following example. Their initial average reputations are zero. During 50,000 simulation cycles, we found that the average reputation of the whole group agents changed rapidly in the very beginning, climbed to a peak, but then slowed down and stabilized at a low level. Figure 5 confirms our hypothesis.

3.3 Reputation Buildup

Clearly, a functioning social network cannot remain stable for long, because agents will continually introduce and remove themselves from the network. In the second simulation, we show that a *knowledgeable* agent A_g who accumulates a high reputation during the first simulation cycle of 25,000, behaves cooperatively with a cooperativeness factor 1 until it reaches a high reputation value, and then starts abusing its reputation by decreasing its responsiveness factor to 0.25. Thus its average reputation begins to drop, ultimately settling at a reputation of 0. Figure 6 illustrates this case. A reputation of 0 indicates that A_g is no longer a neighbor of any agent. That is, it ends up isolated from the other agents. However, in order to distinguish the agent with *zero* reputation (new agent) from the agent with -1 reputation, each agent can potentially blacklist agents for whom it has negative ratings. We defer this enhancement to future work.

3.4 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 hand, 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, with the number of agents

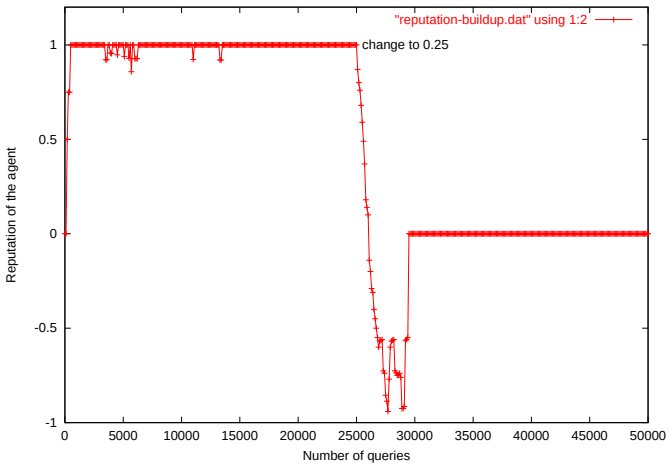

Figure 6: Reputation buildup and crash of a new agent

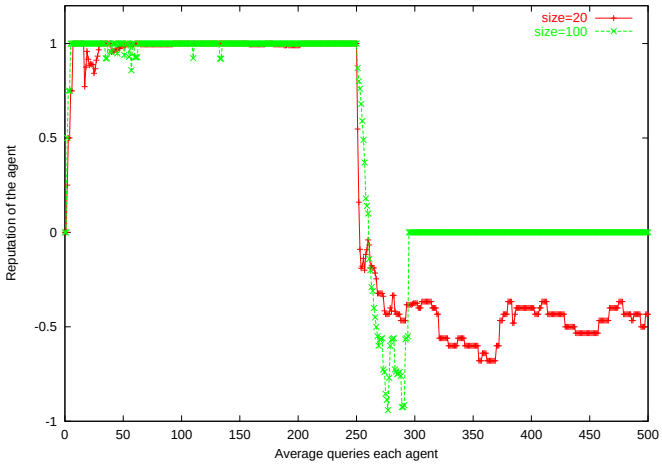

Figure 7: Non-cooperative agents in different community sizes

20 and 100, respectively. Suppose agent A_{g_1} and agent A_{g_2} are two cooperative agents in the beginning with the cooperativeness factors 1. After a series of simulation cycles, both of them decrease their cooperativeness factor to 0.25. Thus, their average reputation starts dropping because of their non-cooperative behaviors. Figure 7 shows that the reputation of agent A_{g_1} drops faster (measured by number of queries sent by each agent) since it is in a smaller community.

Another interesting phenomenon is that the reputation of agent A_{g_1} oscillates around -0.5, a quite low reputation level, but agent A_{g_2}'s reputation changes back to 0. This tells us that a bad agent is more easily forgotten in a big community than in a small group. For our simulation, each agent has 4 neighbors and 12 acquaintances. An agent has it easier selecting good neighbors from 100 agents than from 20 agents.

3.5 Ratio of Non-Cooperative Agents

We believe the ratio of honest to dishonest agents will affect the structure of the whole community. If several agents in the community are dishonest and agents don't trust each other, the community will collapse. We consider the problem of determining the percentage of non-cooperative agents at which a stable community is destroyed. Figure 8 shows the results, where the average reputation of agents converges to a high level when the ratio of non-cooperative agents is 20%. We also can find that agent group with 20% non-cooperative agents converges faster than other two. Since in a group where cooperative agents dominate, it is easier to find good neighbors and detect non-cooperative agents.

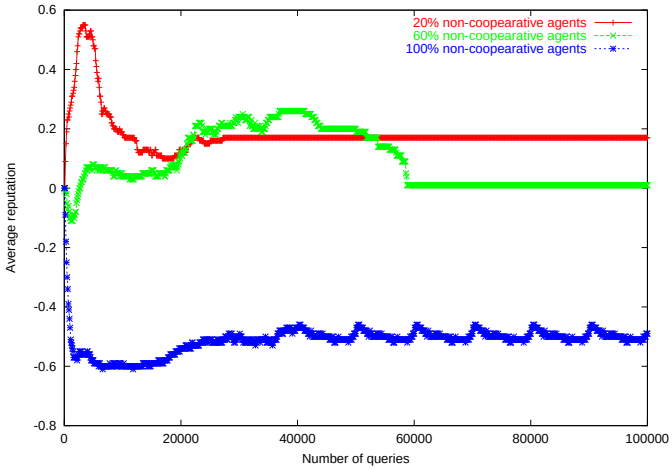

Figure 8: Different ratio of non-cooperative agents

4. RELATED WORK

OnSale Exchange and eBay are important practical examples of reputation management systems. OnSale Exchange 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 are completely centralized and require users to explicitly make and reveal their ratings of others. However, it is questionable if the reputation ratings reflect the trustworthy behavior of sellers, since in the online marketplaces, it is very likely for a user to misbehave, receive low reputation ratings and obtain another online identity.

Some prototype approaches are relevant, like Yenta [9] and Weaving a Web of Trust [14]. Yenta clusters people with common interests according to referrals of users who know each other and verify the assertions they make about themselves, while Weaving a Web of Trust relies on the existence of a connected path between two users. 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.

A social mechanism of reputation management was implemented in Kasbah [8, 28] which require that users give a rating for themselves and either have a central agency (di-

rect 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 in a directed graph. However, it is not clear how the agents collect the ratings in an open environment where the number of agents grows to very large.

Trusted Third Parties (TTP) [19] are often employed to facilitate trust in commercial transactions. Typical TTP services for electronic commerce include certification, time-stamping and notarization. TTPs act as a bridge between buyers and sellers in electronic marketplaces. However, TTP is most appropriate for closed marketplaces. In loosely federated, open systems a TTP may either not be available or have limited power to enforce good behavior.

One of the first works that tried to give a formal treatment of trust was that of Marsh [16]. His model attempted to integrate all the aspects of trust taken from sociology and psychology. Since Marsh's model has strong sociological foundations, the model is rather complex and cannot be easily used in today's electronic communities. Moreover the model only considers 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) [20]. In SIF, an agent evaluates the reputation of another agent based on direct observations as well through other *witnesses*. Moreover Michael Schillo *et al.* tested the performance of two groups of agents with different honesty/dishonesty for altruism/egotism [21]. Schillo's work motivates some of our experiments for reputation management. 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 each other.

Yu and Singh developed an approach for social reputation management [26, 27], in which they used a scalar value to represent an agent's belief ratings about another and combine them with testimonies using combination schemes similar to the certainty factor model. The drawbacks of the certainty factor models, discussed in Section 2, led us to consider alternate approaches.

Rahman and Hailes [1] proposed an approach in virtual communities. Basically it is a kind of adaptation of Marsh's work and some concepts were simplified (for example, trust can have only four possible values) and some were kept (such as situation or contexts). The main problem with their approach is that every agent must keep rather complex data structures that represent a kind of global knowledge about the whole network. Usually maintaining and updating these data structures can be laborious and time-consuming. Also it is not clear how the agents get needed information and how well the model will scale when the number of agents grows.

Aberer and Despotovic [2] simplified our model and use that to manage trust in a peer-to-peer network where no central database is available. Their model is based on binary trust, i.e., an agent is either trustworthy or not. In case a dishonest transaction happened, the agents can forward their complaints to other agents. They used a special data structure, namely P-Grid, to store the complaints in a peer -to-peer network. In order to evaluate the trustworthiness of another agent B, an agent A searches the leaf level of the P-Grid for complaints on agent B.

Barber and Kim [4] discussed a multiagent belief revision algorithm based on belief networks. In their model the agent is able to evaluate incoming information and generate a consistent knowledge base, and avoid fraudulent information from unreliable or deceptive information source or agents. Their research focused on modeling the reliability of information source and maintaining the knowledge base of each agent, while we tried to effectively detect untrustworthy agents in a group.

There has been much work on social abstractions for agents, e.g., [6, 10]. 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 [7]. However, our proposed approach goes beyond their approach in the kinds of representations of trust, propagation algorithms, and formal analysis.

5. CONCLUSION

This paper examines trust in an application- and domain-independent manner and emphasizes the key properties of trust. For this reason, we directly consider how agents may place trust in other agents and finesse the ways in which a principal may convey its trustworthiness to another principal, for example, with various subtle actions and social moves. The explicit reputation management can help the agents detect selfish, antisocial, or unreliable agents and leads to more robust multiagent systems.

The iterated, multi-player prisoners' dilemma is intimately related to the evolution of trust [3, 5]. On the one hand, if the players trust each other, they can both cooperate and avert a mutual defection where both suffer. On the other hand, such trust can only build up in a setting where the players must repeatedly interact with each other. Our observation is that a reputation mechanism sustains rational cooperation, because the better players are rewarded by society whereas the bad players are penalized. Both the rewards and penalties can be greater from a society than from an individual [23].

Our present approach does not fully protect against spurious ratings generated by malicious agents. It relies upon 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 in extensions of the present framework. We also plan to study 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.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant IIS-9624425 (Career Award) and ITR-0081742. We are indebted to the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] A. Abdul-Rahman and S. Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd Hawaii International Conference on Systems Science*, 2000.

[2] K. Aberer and Z. Despotovic. Managing trust in a peer-2-peer information system. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 310–317, 2001.

[3] R. Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[4] K. S. Barber and J. Kim. Belief revision process based on trust: Simulation experiments. In *Proceedings of Autonomous Agents '01 Workshop on Deception, Fraud, and Trust in Agent Societies*, pages 1–12, May 2001.

[5] R. Boyd and J. P. Borderbaum. No pure strategy is evolutionary stable in the repeated prisoner's dilemma game. *Nature*, 327:58–59, 1987.

[6] C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.

[7] C. Castelfranchi and R. Falcone. Principle of trust for MAS: cognitive anatomy, social importance, and quantification. In *Proceedings of 3rd International Conference on MultiAgent Systems*, pages 72–79, 1998.

[8] A. Chavez and P. Maes. Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the 1st International Conference on the Practical Application of Intelligent Agents and Multiagent Technology (PAAM)*, pages 75–90, 1996.

[9] L. Foner. Yenta: A multi-agent, referral-based matchmaking system. In *Proceedings of the 1st International Conference on Autonomous Agents*, pages 301–307, 1997.

[10] L. Gasser. Social conceptions of knowledge and action: DAI foundations and open systems semantics. *Artificial Intelligence*, 47:107–138, 1991.

[11] J. Gordon and E. H. Shortliffe. A method for managing evidential reasoning in a hierarchical hypothesis space. *Artificial Intelligence*, 26:323–357, 1985.

[12] D. Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In *Uncertainty in Artificial Intelligence*, pages 167–196, 1986.

[13] H. Kautz, B. Selman, and A. Milewski. Agent amplified communication. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 3–9, 1996.

[14] R. Khare and A. Rifkin. Weaving a web of trust. *World Wide Web*, 2(3):77–112, 1997.

[15] H. E. Kyburg. Bayesian and non-bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

[16] S. P. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Department of Computing Science and Mathematics, University of Stirling, Apr. 1994.

[17] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Mateo, California, 1988.

[18] M. Prietula and K. M. Carley. Exploring the effects of agent trust and benevolence in a simulated organization task. *Applied Artificial Intelligence*, 13:321–338, 1999.

[19] T. Rea and P. Skevington. Engendering trust in electronic commerce. *British Telecommunications Engineering*, 17(3):150–157, 1998.

[20] M. Schillo and P. Funk. Who can you trust: Dealing with deception. In *Proceedings of the Autonomous Agents Workshop on Deception, Fraud and Trust in Agent Societies*, pages 95–106, 1999.

[21] M. Schillo, P. Funk, and M. Rovatsos. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence*, 14:825–848, 2000.

[22] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[23] S. P. Shapiro. The social control of impersonal trust. *The American Journal of Sociology*, 93(3):623–658, 1987.

[24] M. P. Singh, B. Yu, and M. Venkatraman. Community-based service location. *Communications of the ACM*, 44(4):49–54, 2001.

[25] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, June 1998.

[26] B. Yu and M. P. Singh. A social mechanism of reputation management in electronic communities. In *Proceedings of the 4th International Workshop on Cooperative Information Agents*, pages 154–165, 2000.

[27] B. Yu and M. P. Singh. Trust and reputation management in a small-world network. In *Proceedings of the 4th International Conference on MultiAgent Systems*, pages 449–450, 2000. Poster.

[28] G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14:881–908, 2000.