A Social Mechanism of Reputation Management in Electronic Communities

Bin Yu and Munindar P. Singh*

Department of Computer Science
North Carolina State University
446 EGRC, 1010 Main Campus Drive
Raleigh, NC 27695-7534, USA

{byu, mpsingh}@eos.ncsu.edu

Abstract. Trust is important wherever agents must interact. We consider the important case of interactions in electronic communities, where the agents assist and represent principal entities, such as people and businesses. We propose a social mechanism of reputation management, which aims at avoiding interaction with undesirable participants. Social mechanisms complement hard security techniques (such as passwords and digital certificates), which only guarantee that a party is authenticated and authorized, but do not ensure that it exercises its authorization in a way that is desirable to others. Social mechanisms are even more important when trusted third parties are not available. 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 and between people and businesses. We define an electronic *community* as a set of interacting parties (people or businesses). The members of a community provide services as well as referrals for services to each other. Our notion of *services* is general in that they need not be business services provided for a fee, but may be volunteer services, or not even "services" in the traditional sense, e.g., just companionship or lively discussion.

We model an electronic community as a social network, which supports the participants' *reputations* both for expertise (providing good service) and helpfulness (providing good referrals). The social network is maintained by personal agents assisting different users. Agents and their users have full autonomy in deciding whether or how to respond to a request. The agents assist their users in evaluating the services and referrals provided by others, maintaining contact lists, and deciding whom to contact. In this manner, the agents assist their users

* 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.

in finding the most helpful and reliable parties to deal with. The recommendations by the personal agents are based on a representation of how much the other parties can be trusted. The agents build and manage these representations of trust. To do so, the agents not only take into account the previous experiences of their users, but also communicate with other agents (belonging to other users).

The notion of trust complements hard security, e.g., through cryptography. Hard security approaches help establish that the party you are dealing with is authenticated and authorized to take various actions. They don't ensure that that party is doing what you expect and delivering good service. In other words, the hard security approaches simply place a low hurdle of legality that someone must cross in order to participate, whereas trust management makes people accountable even for the legal actions that they perform.

This paper is organized as follows. Section 2 presents some related work in reputation management. Section 3 presents some necessary background on how to establish an electronic community. Section 4 introduces our approach, giving the key definitions and discussing some informal properties of trust. Section 5 presents our experimental model and some basic results that we have obtained. Section 6 concludes our paper with a discussion of the main results and directions for future research.

2 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. Yenta [3], weaving a web of trust [4], and Kasbah [2, 12] 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.

Trusted Third Parties (TTP) [7] 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.

Rasmusson & Janson proposed the notion of *soft security* based on social control through reputation [6]. In soft security, the agents police themselves

without ready recourse to a central authority. Soft security is especially attractive in open settings, and motivates our approach.

Marsh presents a formalization of the concept of trust [5]. 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) [8]. 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 each other.

Challenges. The following are some important challenges for any agent-based approach for reputation management: how to (1) give full control to the users in terms of when to reveal their ratings; (2) help an agent find trustworthy agents (veritable strangers) even without prior relationships; and, (3) speed up the propagation of information through the social network. Our social mechanism seeks to address the above challenges. In particular, ratings are conveyed quickly among agents, even across sub-communities. Therefore, undesirable agents can quickly be ruled out.

3 Electronic Communities

To better understand the notion of trust in 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 defect, thereby leading to a worse payoff for both agents than if both were to cooperate.

The prisoner's dilemma is intimately related to the evolution of trust. 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 have to repeatly interact with each other. Our observation is that a reputation mechanism sustains rational cooperation, because the good players are rewarded by society whereas the bad players are penalized. Both the rewards and penalties from a society are greater than from an individual.

The proposed approach builds on (and applies in) our work on constructing a social network for information gathering [10, 11]. In our architecture, each user is associated with a personal agent. Users pose queries to their agents. 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 its user, the agent may respond with referrals to other users.

A query includes the question as well as the requester's ID and address and a limit on the number of referrals requested. A response may include an answer or a referral, or both, or neither (in which case no response is needed). An agent answers only if it is reasonably confident that its expertise matches the incoming query. A referral depends on the query and on the referring agent's model of other agents; a referral is given only if the referring agent places some trust in the agent being referred.

When the originating agent receives a referral, it decides whether to follow it up. When the agent receives an answer, it uses the answer as a basis for evaluating the expertise of the agent who gave the answer. This evaluation affects its model of the expertise of the answering agent, and its models of any agent who may have given a referral to this answering agent. In general, the originating agent may keep track of more peers than his neighbors. Periodically he decide which peers to keep as neighbors, i.e., which are worth remembering.

4 Reputation Rating and Propagation

In our approach, agent A assigns a rating to agent B based on (1) its direct observations of B *as well as* (2) the ratings of B as given by B's neighbors, and A's rating of those neighbors. The second aspect makes our approach a social one and enables information about reputations to propagate through the network.

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 reputations of the witnesses themselves. Clearly, the weight assigned to a rating should depend on the reputation of the rater. Moreover, reputation ratings cannot be allowed to increase ad infinitum. To achieve the above, we first define an agent's rating of another agent. Initially, the rating is zero.

Definition 1. $T_i(j)^t$ is the trust rating assigned by agent i to agent j at time t. We require that $-1 < T_i(j)^t < 1$ and $T_i(j)^0 = 0$.

Each agent will adapt its rating of another agent based on its observation. Cooperation by the other agent generates a positive evidence α and defection a negative evidence β. Thus $\alpha \geq 0$ and $\beta \leq 0$. To protect those who interact with an agent who cheats some of the time, we take a conservative stance toward reputations, meaning that reputations should be hard to build up, but easy to tear down. This contrasts with Marsh [5], where an agent may cheat a sizable fraction (20%) of the time but still maintain a monotonically increasing reputation. We can achieve the desired effect by requiring that $|\alpha| < |\beta|$. We use a simple approach to combine in evidence from recent interactions.

Definition 2. After an interaction, the updated trust rating $T_i(j)^{t+1}$ is given by the following table and depends on the previous trust rating.

$T_i(j)^t$	**Cooperation by j**	**Defection by j**
> 0	$T_i(j)^t + \alpha(1 - T_i(j)^t)$	$(T_i(j)^t + \beta)/(1 - \min\{\lvert T_i(j)^t\rvert, \lvert\beta\rvert\})$
< 0	$(T_i(j)^t + \alpha)/(1 - \min\{\lvert T_i(j)^t\rvert, \lvert\alpha\rvert\})$	$T_i(j)^t + \beta(1 + T_i(j)^t)$
$= 0$	α	β

Following Marsh [5], we define for each agent an upper and a lower threshold for trust.

Definition 3. For agent i: $-1 \leq \omega_i \leq 1$ and $-1 \leq \Omega_i \leq 1$, where $\omega_i \geq \Omega_i$.

$T_i(j) \geq \omega_i$ indicates that i trusts j and will cooperate with j; $T_i(j) \leq \Omega_i$ indicates that i mistrusts j and will defect against j; $\Omega_i < T_i(j) < \omega_i$ means that i must decide on some other grounds.

4.1 Propagation of Reputation Rating

Each agent has a set of potentially changing *neighbors* with whom it may directly interact. How an agent evaluates the reputations of others will depend in part on the testimonies of its neighbors. This naturally leads to the idea of a referral chain.

Definition 4. $\chi = \langle A_0, \ldots, A_n \rangle$ is a (possible) referral chain from agent A_0 to agent A_n, where A_{i+1} is a neighbor of A_i.

A_0 will use a referral chain to A_n to compute its rating $T_0(n)$ towards A_n. We define a trust propagation operator, $\otimes$.

Definition 5. $x \otimes y =$ if $(x \geq 0 \wedge y \geq 0)$ then $x \times y$ else $-|x \times y|$

In other words, the level of trust propagated over a negative link in a referral chain is negative. Below, let $\chi = \langle A_0, \ldots, A_n \rangle$ be a referral chain from agent A_0 to agent A_n at time t. We now define trust propagation over a referral chain.

Definition 6. For any k, $0 \leq k \leq n$, $T_0^\chi(k)^t = T_0^\chi(1)^t \otimes \ldots \otimes T_{k-1}^\chi(k)^t$

The penultimate agent on a referral chain has direct evidence of the last agents on the chain. For this reason, we term the penultimate agent the *witness.*

Definition 7. A testimony for agent 0 from agent k relative to a chain χ is defined as $E_0^\chi(k)^t = T_0^\chi(k)^t T_k^\chi(k+1)^t$. Here k is the witness of this testimony.

Testimony from a witness is used when the witness is considered sufficiently reliable. So as to allow testimony from weak agents to be combined in, we consider witnesses reliable as long as they have a positive trust rating.

Definition 8. For agent i at time t, a testimony from agent k is *reliable* if and only if agent k is trusted, i.e., $T_i^\chi(k)^t > 0$.

Two referral chains χ_1 and χ_2 may pass through the same agent k. In this case, we choose a referral chain that yields the highest trust rating for k.

Definition 9. For agent i, a testimony from agent k with respect to referral chain χ_1 is more reliable than with respect to referral chain χ_2 if and only if χ_1 yields a higher trust rating for agent k, i.e., $T_i^{\chi_1}(k) \geq T_i^{\chi_2}(k)$.

4.2 Incorporating Testimonies from Different Witnesses

We now show how testimonies from different agents can be incorporated into the rating by a given agent. First, to eliminate double counting of witnesses, we define *distinct* sets of testimonies. (E_w refers to the witness of testimony E).

Definition 10. A set of testimonies $\mathcal{E} = \{E_1, \ldots, E_L\}$ towards agent n is *distinct* if and only if the witnesses of all testimonies in $\mathcal{E}$ are distinct, i.e., $|\{E_{1w}, \ldots, E_{Lw}\}| = L$.

The *maximally reliable distinct (MRD)* subset of a set of testimonies contains all the trustable testimonies, and for any witness, it contains the best testimony from that witness. Notice that the individual witnesses do not have to be trusted greater than ω_i for their testimony to be used.

Definition 11. $\mathcal{V}$ is a MRD subset of a set of testimonies $\mathcal{E}$ if and only if $\mathcal{V}$ is distinct, $\mathcal{V} \subseteq \mathcal{E}$, and $(\forall E : (E \in \mathcal{E} \wedge T_i^{\chi_E}(E_w) > 0) \Rightarrow (\exists V : V \in \mathcal{V} \wedge V_w = E_w \wedge T_i^{\chi_V}(V_w) \geq T_i^{\chi_E}(E_w)))$.

Given a set of testimonies $\mathcal{E}$ about A_n, we first find its MRD subset $\mathcal{V}$. Next we compute the average of testimonies from $\mathcal{V}$: $\overline{E} = 1/L \sum_{i=1}^{|V|} V_i$. Therefore, agent A_0 will update its trust rating of agent A_n as follows (all ratings are at time t except where specified).

when	**then** $T_0(n)^{t+1} =$
$T_0(n)$ and $\overline{E}$ are positive	$T_0(n) + \overline{E}(1 - T_0(n))$
one of $T_0(n)$ and $\overline{E}$ is negative	$T_0(n) + \overline{E}/(1 - min\{\|T_0(n)\|, \|\overline{E}\|\})$
$T_0(n)$ and $\overline{E}$ are negative	$T_0(n) + \overline{E}(1 + T_0(n))$

4.3 Gossip

If an agent A encounters a bad partner B during some exchange, A will penalize B by decreasing its rating of B by β and informing its neighbors. An agent who receives this information can combine it into its trust model of B.

Gossip is different from the usual referral process, because an agent can propagate a rumor without having been explicitly queried. For this reason, gossip is processed incrementally.

Definition 12. Suppose agent i receives a message $T_k(n)$ (from agent k about agent n). If $T_i(k)$ is negative, then i ignores the message. If $T_i(k)$ is positive, then agent i updates its trust rating of agent n as follows.

when $T_i(n)$ **and** $T_k(n)$	**then** $T_i(n)^{t+1} =$
are both positive	$T_i(n) = T_i(n) + T_i(k)T_k(n)(1 - T_i(n))$
are both negative	$T_i(n) + T_i(k)T_k(n)(1 + T_i(n))$
have opposite signs	$(T_i(n) + T_i(k)T_k(n))/(1 - min\{\|T_i(n)\|, \|T_i(k)T_k(n)\|\})$

4.4 Properties of Trust

We now describe and formalize some important properties of trust.

1. *Symmetry*
   In general, symmetry will not hold, because an agent may trust another more than it is trusted back. However, when the agents are trustworthy, through repeated interactions, they will converge to high mutual trust. Conversely, if one of the agents doesn't act in a trustworthy manner, the other agent will be forced to penalize it, leading to low mutual trust. For this reason, we have for any two agents A_x and A_y, $T_x(y)^t \approx T_y(x)^t$ when $t \to \infty$.
2. *Transitivity*
   Trust is not transitive, but the following will hold if x is a rational agent:
   $(T_x(y)^t > T_x(z)^t) \wedge (T_x(z)^t > T_x(w)^t) \Rightarrow (T_x(y)^t > T_x(w)^t)$
3. *Self-reinforcement*
   Trust is self-reinforcing, because agents act positively with those whom they trust. The converse is true, as below a certain trust, individuals tend to confirm their suspicions of others [9]. The first part of the following rule is based on the idea that if trust between two agents is initially above ω, then the trust between those two agents will not decrease below that threshold. The converse is true, since if both agents trust each other below Ω, they will tend not to cooperate with each other whatever the situation, thus reinforcing the other's opinion about them as non-cooperative and unhelpful. Between ω and Ω, anything can happen [5].
   - If $(T_x(y)^t > \omega_x) \wedge (T_y(x)^t > \omega_y)$ then
     $(T_x(y)^{t+1} \geq T_x(y)^t) \wedge (T_y(x)^{t+1} \geq T_y(x)^t)$
   - If $(T_x(y)^t < \Omega_x) \wedge (T_y(x)^t < \Omega_y)$ then
     $(T_x(y)^{t+1} \leq T_x(y)^t) \wedge (T_y(x)^{t+1} \leq T_y(x)^t)$
4. *Propagation*
   Consider three agents x, y, and z. If x knows y and y knows z, but x does not know z. How much x trusts z should depend on how much x trusts y, and how much y trusts z. The following rule will hold if x is rational.

   $$(T_x(z)^{t+1} \leq T_x(y)^t) \wedge (T_x(z)^{t+1} \leq T_y(z)^t)$$

   A simple formula for determining trust that satisfies the above constraint, is

   $$T_x(z)^{t+1} = T_x(y)^t T_y(z)^t$$

5 Experiments and Results

In our simulated setup, each agent has an *interest* vector, an *expertise* vector, and models of several *neighbors*. 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 reputation.

An agent's queries are generated based on its interest vector. The queries are generated as vectors by perturbing the interest vector of the given agent. The motivation for this is to capture the intuition that an agent will produce queries depending on its interests.

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 involve between 20 and 60 agents with interest and expertise vectors of dimension 5. The agents send queries, referrals, and responses to one another, all the while learning about each others' interest and expertise vectors. The agents are limited in the number of neighbors they may have—in our case the limit is 4.

5.1 Metrics

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

Definition 13. The average reputation of an agent A_i from the point of other agents is given by $\overline{r(A_i)}$:

$$\overline{r(A_i)} = 1/n \sum_{j=1}^{n} T_j(A_i)$$

where n is the number of agents who know agent A_i. We say that agent A_k knows agent A_i if and only if A_i is a neighbor of A_k.

Definition 14. The average reputation of all agents is: $\overline{R} = 1/N \sum_{i=1}^{N} \overline{r(A_i)}$, where N is the total number of agents.

This average is a metric for determining the stabilization of a community.

5.2 Selection of Rewards and Penalties

Figure 1 illustrates the change of trust ratings depending on different values of α and β. Part A applies to a new agent who initially has a trust of 0, but builds up the rating through positive interactions; Part B applies to a cooperative agent who is already well-trusted; Part C applies to an untrusted agent who through repeated positive interactions becomes trusted; Part D applies to a new agent whose rating falls because of negative interactions; Part E describes a trusted agent who becomes untrusted because of defections; and, Part F applies to an untrusted agent who becomes further untrusted because of defections.

Consider an agent who cooperates and defects on different interactions. Let θ be the ratio between the number of cooperations and defections. By appropriately selecting the ratings of α and β, we can let $\theta \to \infty$. Assume the initial trust rating of agent A_i is 0.6. Let $\theta = 5, 10, 20$. Figure 2 displays the change of trust rating. Notice that trust built up through several positive interactions is lost through even a single defection.

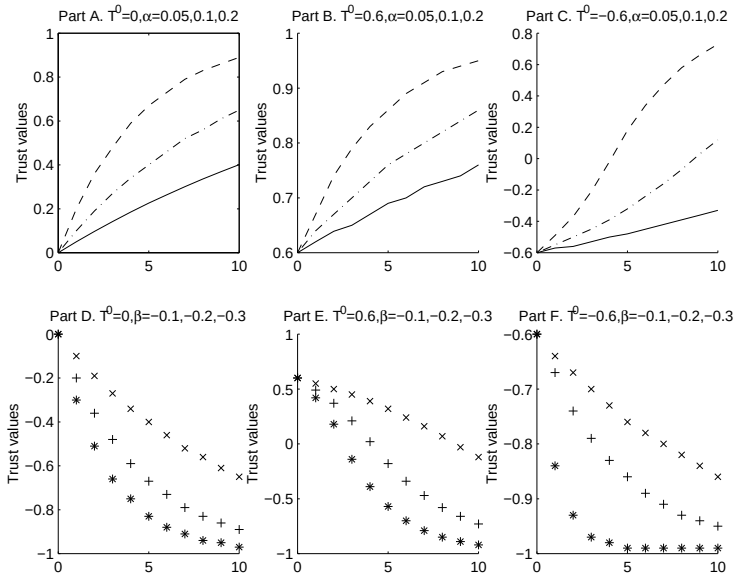

Fig. 1. Selection of α and β, where $\alpha = 0.05('-'),\ 0.1('-.'),\ 0.2('--')$ and $\beta = -0.1('x'),\ -0.2('+'),\ -0.3('*')$

5.3 Avoiding Undesirable Agents

Our mechanism quickly lowers the reputations of selfish agents. Consider the following example. Assume agent A_w is a non-cooperative agent, and only three agents A_x, A_y, and A_z know him. Their initial ratings towards A_w are 0.4, 0.5, and 0.6, respectively.

So the average reputation of agent A_w at time 0 is 0.5. However, say at time 1, agent A_w defects against agent A_x. Let $\alpha = 0.05$ and $\beta = -0.3$. According to the formula for updating trust, $T_x(w) = (0.4 + (-0.3))/(1 - min|0.4|, |-0.3|) = 0.1/0.7 = 0.1429$. The new reputation of the agent is $r(A_w) = 0.413$. Moreover, agent A_x will disseminate its observation of agent A_w throughout the social network. Eventually the average reputation of agent A_w may decrease to a low level. This is the *power of referrals*. Figure 3 experimentally confirms our hypothesis.

5.4 Introducing New Agents

Clearly, a social network will not remain stable for long, because agents will continually introduce and remove themselves from the network. To evaluate how our approach accommodates changes of this variety, we begin with a stable network and introduce a new agent randomly into it. The new agent is given random neighbors, and all of their trust ratings towards this new agent are zero.

Assume $\overline{R} = 0.637$ at time t. In order to be embedded into the social network, the new agent would have to keep cooperating reliably or else be ostracized early.

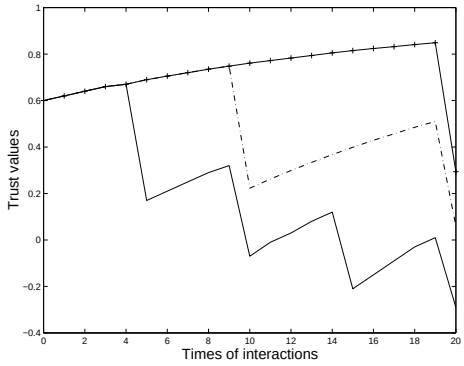

Fig. 2. Change of trust for $\theta = 5('-')$, $10('-.')$, $20('-+')$ when $\alpha = 0.05$ and $\beta = -0.3$

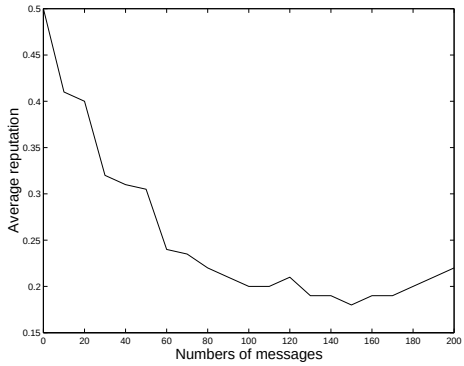

Fig. 3. Average reputation of agent A_w for $N = 30$, $\alpha = 0.05$ and $\beta = -0.3$

Its initial threshold for cooperating is low. By frequently cooperating with other agents, the new agent can have its average reputation increase steadily. Figure 4 confirms this hypothesis.

6 Discussion

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. 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.

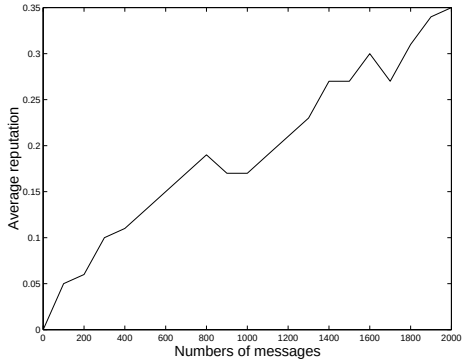

Fig. 4. Average reputation of new agent A_{new} for $N = 30$, $\alpha = 0.05$ and $\beta = -0.3$

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. This is not ideal, but not any worse than democratic rule in human societies. Democratic societies cannot guarantee that a malicious ruler won't be elected, but they reduce the chance of such an event by engaging a large fraction of the population in the rating process.

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. Anthony Chavez and Pattie 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'96)*, 1996.
3. Lenny Foner. Yenta: A multi-agent, referral-based matchmaking system. In *Proceedings of the 1st International Conference on Autonomous Agents*, pages 301–307, 1997.
4. Rohit Khare and Adam Rifkin. Weaving a web of trust. *World Wide Web*, 2(3):77–112, 1997.
5. P. Steven Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, Department of Computing Science and Mathematics, University of Stirling, April 1994.
6. Lars Rasmusson and Sverker Janson. Simulated social control for secure Internet commerce. In *Proceedings of the Workshop on New Security Paradigms*, 1996.

7. Tim Rea and Peter Skevington. Engendering trust in electronic commerce. *British Telecommunications Engineering*, 17(3):150–157, 1998.
8. Michael Schillo and Petra Funk. Who can you trust: Dealing with deception. In *Proceedings of the workshop Deception, Fraud and trust in Agent Societies at the Autonomous Agents Conference*, pages 95–106, 1999.
9. Susan P. Shapiro. The social control of impersonal trust. *The American Journal of Sociology*, 93(3):623–658, 1987.
10. Bin Yu and Munindar P. Singh. An multiagent referral system for expertise location. In *Working Notes of the AAAI Workshop on Intelligent Information Systems*, pages 66–69, 1999.
11. Bin Yu, Mahadevan Venkatraman, and Munindar P. Singh. An adaptive social network for information access: Theoretical and experimental results. *Applied Artificial Intelligence*, 2000. To appear.
12. Giorgos Zacharia, Alexandros Moukas, and Pattie Maes. Collaborative reputation mechanisms in electronic marketplaces. In *Proceedings of the HICSS-32 Minitrack on Electronic Commerce Technology*, 1999.