Applied Artificial Intelligence, 14:881–907

0883-9514/00 $12.00 + .00

TRUST MANAGEMENT THROUGH REPUTATION MECHANISMS

GIORGOS ZACHARIA and PATTIE MAES
MIT Media Laboratory,
Cambridge MA, U.S.A.

The members of electronic communities are often unrelated to each other; they may have never met and have no information on each other's reputation. This kind of information is vital in electronic commerce interactions, where the potential counterpart's reputation can be a significant factor in the negotiation strategy. Two complementary reputation mechanisms are investigated which rely on collaborative rating and personalized evaluation of the various ratings assigned to each user. While these reputation mechanisms are developed in the context of electronic commerce, it is believed that they may have applicability in other types of electronic communities such as chatrooms, newsgroups, mailing lists, etc.

"Although an application designer's first instinct is to reduce a noble human being to a mere account number for the computer's convenience, at the root of that account number is always a human identity." (Khare & Rifkin, 1997)

Online communities bring together people geographically and sociologically unrelated to each other. Online communities have traditionally been created in the context of discussion groups, in the form of newsgroups, mailing lists, or chatrooms. Online communities are usually either goal or interest-oriented. But other than that, there is rarely any other kind of bond or real-life relationship among the members of communities before the members meet each other online. The lack of information about the background, character, and especially the reliability of the members of these communities causes a lot of suspicion and mistrust among their members.

When a newcomer joins a chatroom, a newsgroup, or a mailing list, he/she does not know how seriously he/she should take each participant until he/she has formed an opinion about the active members of the group. Likewise, the old members of the group do not know how seriously they should take a newcomer until he/she establishes him/herself in the group. If

Address correspondence to Giorgos Zacharia, MIT Media Laboratory, 20 Ames Street, E15-305, Cambridge, MA, 02139, USA. E-mail: lysi@media.mit.edu

the group has a lot of traffic, the noise-to-signal ratio becomes too high, and the process of filtering out the interesting messages becomes increasingly difficult for a newcomer or an occasional reader of the group. If users did have an indication for the reputation of the author of each message, they could prioritize the messages according to their predicted quality.

Similar problems are encountered in other kinds of online communities. The recent development of online auction sites, and other forms of electronic marketplaces has created a new kind of online community, where people meet each other to bargain and transact goods. Online marketplaces like Amazon Auctions (Amazon), Kasbah (Chavez & Maes, 1996), MarketMaker (Wang, 1999), eBay (eBay), and OnSale Exchange (OnSale) introduce two major issues of trust:

- Potential buyers have no physical access to the product of interest while they are bidding or negotiating. Therefore, sellers can easily misrepresent the condition or the quality of their products.
- Additionally, sellers or buyers may decide not to abide by the agreement reached at the electronic marketplace, asking later to renegotiate the price, or even refuse to commit the transaction. Even worse, they may receive the product and refuse to send the money for it, or the other way around.

Although these problems of trust are also encountered in real-world experiences, the problem is more difficult in online communities, because one has very few cues about other people by which to evaluate them. Many of the signals that we use in real life are absent in online environments, and thus alternative methods of adjudicating reputation are needed.

One way of solving the above-mentioned problems in the system would be to incorporate a reputation brokering mechanism, so that each user can customize his/her pricing strategies according to the risk implied by the reputation values of his/her potential counterparts.

Reputation is usually defined as the amount of trust inspired by a particular person in a specific setting or domain of interest (Marsh, 1994). In "Trust in a Cryptographic Economy" (Reagle, 1996), reputation is regarded as asset creation and it is evaluated according to its expected economic returns.

Reputation is conceived as a multidimensional value. An individual may enjoy a very high reputation for his/her expertise in one domain, while having a low reputation in another. For example, a Unix guru will probably have a high rank regarding Linux questions, while he may not enjoy as high a reputation for questions regarding Microsoft's operating systems. These individual reputation standings are developed through social interactions among a loosely connected group that shares the same interest. Also, each

user has his/her personal and subjective criteria for what makes a user reputable. For example, in the context of a discussion group, some users prefer polite mainstream postings, while others engage in flame wars. Through this interaction, the users of online communities establish subjective opinions of each other.

Methods have been developed through which one can automate the social mechanisms of reputation for electronic communities. An early version of these reputation mechanisms has been implemented in Kasbah (Chavez & Maes, 1996). Kasbah is an ongoing research project to help realize a fundamental transformation in the way people transact goods- from requiring constant monitoring and effort, to a system where software agents do much of the bidding and negotiating on a user's behalf. A user wanting to buy or sell a good creates an agent, gives it some strategic direction, and sends it off into the marketplace. Kasbah agents proactively seek out potential buyers or sellers and negotiate with them on their creator's behalf. Each agent's goal is to make the "best deal" possible, subject to a set of user-specified constraints, such as a desired price, a highest (or lowest) acceptable price, and a date to complete the transaction (Chavez & Maes, 1996). In Kasbah, the reputation values of the individuals trying to buy/sell books/CDs are major parameters of the behavior of the buying, selling, or finding agents of the system.

The second section of this paper describes the related work in the domain of rating systems and reputation mechanisms. The third section outlines the requirements for a successful reputation mechanism for online communities. The fourth section describes problems specific to electronic marketplaces and online discussion forums. The fifth and sixth sections describe two reputation mechanisms that have been designed and evaluated. The seventh section evaluates the mechanisms using simulations and user data from eBay and Amazon auctions. The last section is the conclusion of the paper and the outline of future work.

RELATED WORK

The related work on reputation systems can be divided into two major categories: noncomputational reputation systems like the Better Business Bureau Online (BBB) and computational ones. The Better Business Bureau Online is a centralized repository of consumer and business alerts. They mainly provide information on how well businesses handle disputes with their clients. They also keep records of the complaints about local or online companies and even publish consumer warnings against some of them. They do not provide any kind of numerical ratings for business or consumer trustworthiness.

The computational methods cover a broad domain of applications, from rating of newsgroup postings and webpages, to rating people and their expertise in specific areas. This section focuses on the related computational methods and a comparison of their major features (Table 1).

One way of building a reputation mechanism involves having a central agency which keeps records of the recent activities of the users of the system, very much like the scoring systems of credit history agencies. The credit history agencies use customized evaluation mechanisms provided by the software of FairIsaac (FairIsaac) in order to assess the risk involved in giving a loan to an end consumer. The ratings are collected from the previous lenders of the consumers, and consumers are allowed to dispute those ratings if they feel they have been treated unfairly. The resolution of a rating dispute is a responsibility of the end consumer and the party that rated the particular consumer.

However useful a centralized approach may be, it requires a lot of overhead on behalf of the service providers of the online community. Furthermore, the centralized solutions ignore possible personal affinities, biases, and standards that vary across various users.

Other proposed approaches like Yenta (Foner, 1997), Weaving a Web of Trust (Khare & Rifkin, 1997), and the Platform for Internet Content Selection (PICS), such as the Recreational Software Advisory Council (RSAC), are more distributed. However, they require the users to rate themselves and to have either a central agency or other trusted users verify their trustworthiness. One major problem with these systems is that no user would ever label him/herself as an untrustworthy person. Thus, all new members would

TABLE 1 Comparison of Online Reputation Systems. In the "Pairwise Rating" Column It is Indicated Whether the Ratings are Bi-Directional or One-Directional, and Who Submits Ratings. In the "Personalized Evaluation" Column It is Indicated Whether the Ratings are Evaluated in a Subjective Way, Based on Who Makes the Query

System	Pair-wise rating	Personalized Evaluation	Textual comments
Firefly	Rating of recommendations	Yes	Yes
GroupLens	Rating of articles	Yes	No
Web of Trust	Transitive ratings	Yes	No
eBay	Buyers and sellers rate each other	No	Yes
Amazon	Buyers and sellers rate each other	No	Yes
OnSale	Buyers rate sellers	No	Yes
Credit history	Lenders rate customers	No	Yes
PICS	Self-rating	No	No
Elo & Glicko	Result of game	No	No
Bizrate	Consumers rate businesses	No	Yes

need verification of trustworthiness by other trustworthy users of the system. In consequence, a user would evaluate his/her counterpart's reputation by looking at the numerical value of his/her reputation as well as the trustworthiness of his/her recommenders.

Yenta and Weaving a Web of Trust introduce computational methods for creating personal recommendation systems, the former for people and the latter for webpages. Weaving a Web of Trust relies on the existence of a connected path between two users, while Yenta clusters people with common interests according to recommendations of users who know each other and can verify the assertions they make about themselves. Both systems require prior existence of social relationships among their users, while in online marketplaces, deals are brokered among people who may have never met each other.

Collaborative filtering is a technique for detecting patterns among the opinions of different users, which can then be used to make recommendations to people, based on opinions of others who have shown similar taste. This technique basically automates "word of mouth" to produce an advanced and personalized marketing scheme. Examples of collaborative filtering systems are HOMR, Firefly (Shardanand & Maes, 1995), and GroupLens (Resnick et al., 1994). GroupLens is a collaborative filtering solution for rating the contents of Usenet articles and presenting them to the user in a personalized manner. In this system, users are clustered according to the ratings they give to the same articles. These ratings are used for determining the average ratings of articles for that cluster.

The Elo (Elo, 1978) and the Glicko (Glickman, 1999) systems are computational methods used to evaluate the player's relative strengths in pairwise games. After each game, the competency score of each player is updated based on the result and previous scores of the two users. The basic principle behind ratings in pairwise games is that the ratings indicate which player is most likely to win a particular game. The probability that the stronger player will win the game is positively related to the difference in the abilities of the two users. In general, the winner of a game earns more points for his/her rating, while the defeated player loses points from his rating. The changes in the ratings of the two users depend on their rating difference before the game takes place. If the winner is the player who had a higher score before the game, the change in the ratings of the two users is negatively related to their rating difference before the game. If, however, the winner of the game is the player who had a lower score before the game took place, the changes in the sores of the two players are positively related to their rating difference before the game.

BizRate (BizRate) is an online shopping guide that provides ratings for the largest 500 companies trading online. The ratings are collected in two different ways. If BizRate has an agreement with an online company, the

company provides BizRate with transaction information so that BizRate can independently survey the satisfaction of every customer who makes a purchase from its website. The surveys measure the customer satisfaction in several categories, and BizRate provides an overall, as well as detailed report, on the performance of the rated company. If a company does not have an agreement with BizRate, then the staff of BizRate reviews the company and provides a report based on the editorial assessment of BizRate. BizRate rates different features for different categories of companies, based on BizRate's hierarchical ontology of online businesses. The scores in each category are computed as the average of the collected ratings, and they are given on a scale of 1 to 5. The consumer reviews are presented separately from the editorial reviews, and the companies that agree to have their customers rate them are labeled as "customer certified merchants."

In the context of electronic marketplaces, the most relevant computational methods are the reputation mechanism of online auction sites like OnSale Exchange[1] (OnSale), eBay (eBay), and Amazon Auctions (Amazon). In OnSale, which used to allow its users to rate sellers, the overall reputation value of a seller was calculated as the average of all his/her ratings through his/her usage of the OnSale system. In eBay, sellers receive $+1$, 0 or -1 as feedback for their reliability in each auction and their reputation value is calculated as the sum of those ratings over the last 6 months. In OnSale, newcomers had no reputation until someone eventually rated them, while in eBay they start with zero feedback points. Bidders in the OnSale Exchange auction system were not rated at all.

OnSale tried to ensure the bidders' integrity through a rather psychological measure: bidders were required to register with the system by submitting a credit card number. OnSale believed that this requirement helped to ensure that all bids placed were legitimate, which protected the interests of all bidders and sellers. However, the credit card submission method does not solve the multiple identities, problem, because users can have multiple credit cards in their names. In both the eBay and OnSale systems, the reputation value of a seller is available, with any textual comments that may exist to the potential bidders. The mechanism at Amazon auctions is exactly the same as OnSale's, with the improvement that both the buyers and sellers are rated after each transaction.

In online marketplaces like the auction sites, it is very easy for a user to misbehave, receive low reputation ratings, and then leave the marketplace, obtain another online identity, and come back without having to pay any consequences for the previous behavior. Therefore, newcomers to online marketplaces are treated with suspicion until they have been around long enough with a consistent trustworthy behavior. Thus, newcomers receive less attractive deals than older users that are equally trustworthy. However, this poor treatment to the newcomers creates an economic inefficiency,

because transactions with newcomers are underpriced, or even do not take place at all. This economic inefficiency could be removed if the online sites disallowed anonymity, or alleviated it if newcomers were allowed to pay fees for higher initial reputation values, and those users could be committed to lifetime pseudonyms so that anonymity is preserved, but identity switching is eliminated (Friedman & Resnick, 1998).

Recently, both Amazon and eBay allowed their users to become "eBay registered" users or "Amazon registered" users, respectively. What that means is that they can provide to the marketplace provider enough personal data, so that the marketplace provider can find out their real identities in case of a fraud. Therefore, the users can transact online using pseudonymous identities, whose link to their real identities is held by the marketplace provider alone. Thus, at the expense of their total anonymity, the newly registered users can enjoy increased levels of trust towards them, despite the fact that they do not have any transaction history to prove themselves. This approach makes transactions more efficient from a microeconomic perspective, because the pseudonymous users can achieve better deals than totally anonymous users since they are trusted more (Friedman & Resnick, 1998).

THE PROBLEM OF TRUST IN ONLINE COMMUNITIES

Consumer-to-Consumer Electronic Marketplaces

The emergence of large consumer-to-consumer electronic marketplaces has highlighted several problems regarding issues of trust and deception in these marketplaces. Unlike discussion-oriented online communities like mailing lists, WWW message boards and chatrooms, in these online marketplaces there is a financial cost when users are deceived. The major marketplace providers like eBay, OnSale, Yahoo, and Amazon, tried to tackle the problem by introducing simple reputation mechanisms. These reputation mechanisms try to give an indication of how trustworthy a user is, based on his/her performance in his/her previous transactions. Although there are several kinds of possible frauds or deceptions in online marketplaces, the users' trustworthiness is typically abstracted in one scalar value, called the feedback rating or reputation. The fact that users' trustworthiness is abstracted in this one-dimensional value has been instrumental in the success of these mechanisms, because it minimizes the raters' overhead from a time-cost and usability perspective.

Discussion Forums

Online communities, whether on mailing lists, newsgroups, IRC, or web-based message boards and chatrooms, have been growing very rapidly.

Many Internet users use chatrooms to exchange information on sensitive personal issues, like health-related problems, financial investments, seek help and advise on research and technical related issues, or even discuss and learn about pressing political issues. In all these cases, the reliability of the information posted on the discussion forums is a significant factor for the forum's popularity and success.

The comfort of anonymity is extremely necessary in several cases like controversial political discussions or health-related questions. However, the allowed anonymity makes reliability of the provided information questionable. Therefore, the reputations of the individuals participating in an online community are fundamental for the community's success (Donath, 1998).

However, the perceptions about the reputations of the users among themselves can be very different and subjective. One example of this phenomenon is the "Cyprus List," an English-speaking bicommunal mailing list hosted at MIT. This mailing list has been the only open communication forum between the two communities in Cyprus for the several decades[2] now. However, the Cyprus List allows Greek Cypriots and Turkish Cypriots to share their interpretations of history, perceptions, and misperceptions, and their goals and expectations from a future solution of the problems.

The mailing list includes individuals across the whole political spectra of both sides: from extreme Greek or Turkish nationalists, to moderate and reconciliatory individuals from both communities. Therefore, each one of the members of the list has different subjective opinions about the quality of the postings of everybody else. Naturally, each member views members who come from their own community highly, while they consider the members coming from the opposing side as fanatical and biased. However, moderate members of both communities will often disagree with their extremist compatriots and find themselves in agreement with moderates coming from the opposite community.

The major problem of trust among the members of the list is the question of reliability of the information presented to support the arguments of the two communities. There have been several examples of members quoting books or news articles found at their favorite political publications or websites, which ended up being plagiarism, pure fabrication, or even intentional paraphrasing in order to misrepresent the original quotation. However, in all those cases, several members of the list provided their unconditional belief and confidence to the truthfulness of the information, based on their affinity with the person presenting the information to the list. Therefore, if we ask the members of such an online community to rate how highly they think of each other, we expect to observe a major disparity among the ratings, which should be strongly correlated with the differences of the political biases between the raters and the rated persons.

DESIDERATA FOR ONLINE REPUTATION SYSTEMS

While the previous sections discussed reputation mechanisms that have some interesting qualities, it is believed that they are not perfect for maintaining reputations in online communities and especially in online marketplaces. This section describes some of the problems of online communities and their implications for reputation mechanisms.

In online communities, it is relatively easy to change one's identity (Kollock, 1999; Donath, 1998; Friedman & Resnick, 1998). Thus, if a user ends up having a reputation value lower than the reputation of a beginner, he/she would have an incentive to discard his/her initial identity and start from the beginning. Hence, it is desirable that while a user's reputation value may decrease after a transaction, it will never fall below a beginner's value. However, with such positive reputation mechanisms, the beginners are subject to mistreatment by the rest of the community, because nobody knows if they are in fact users or bad ones who just switched identities. Hence, trustworthy beginners will have to accept less attractive deals in the context of an ecommerce community, or the information they provide on a discussion community will be undervalued until they establish themselves. Therefore, the mistreatment of newcomers creates an inherent economic inefficiency, because the monetary or information transactions of the newcomers are undervalued. This economic inefficiency can be faced either by disallowing anonymity or by allowing users to purchase reputation points for a monetary value. However, in such a model one needs to charge for names in the first place and enforce persistent pseudonymous identities (Friedman & Resnick, 1998). Despite the benefits of this model, one decided against it because of the requirement for persistent pseudonymous identities. In some forms of online communities, it is desirable to allow users to have multiple personalities and/or switch identities. For example, in political discussions forums like the Cyprus List (Cyprus-L), it is very important to allow some users to maintain different personalities than the ones they use on their respective Greek or Turkish community mailing lists. Because of these reasons, a first desideratum was decided on for online reputation mechanisms, namely, that it is desirable that a beginner cannot start with a reputation above the minimum allowed by the system.

In addition, users who have very low reputation ratings should be able to improve their ratings at almost the same rate as a beginner. This implies that the reputation value of users should not be the arithmetic average of all of their ratings, since this would give the users who perform relatively poorly in the beginning an incentive to get rid of their bad reputation history by adopting a new identity.

Therefore, a successful online reputation mechanism has to be based on

a positive reputation system. However, having the users start with minimum reputation is not necessarily the only viable solution. An alternative approach (Friedman & Resnick, 1998) would be to allow newcomers to pay entry fees in order to be considered trustworthy. This approach would be very applicable in online marketplaces, where the interaction is clearly monetary-based. However, it would probably be unwelcome in other more casual forms of online communities like newgroups or mailing lists.

Another problem with systems like Kasbah and online auction sites is that the overhead of performing fake transactions is fairly low. This makes it possible for people to perform fake transitions with their friends, rating each other with perfect scores each time, so as to increase their reputation value. Likewise, in an online group, the marginal cost of sending a new message is zero. So a group of users may exchange messages for the sake of creating fresh unique ratings for each other. Notice that prohibiting each user from rating others more than once would not solve this problem since a user can still falsely improve his/her ratings by creating multiple fake identities, which can then rate the user's real identity with perfect scores. A good reputation system should avoid both of these problems.

In order to do this, one has to ensure that the ratings given by users with an established high reputation in the system are weighted more than the ratings given by beginners or users with low reputations. In addition, the reputation values of the user should not be allowed to increase ad infinitum as is the case with eBay, where a seller can cheat 20% of the time but still maintain a monotonically increasing reputation value.

Reputation mechanisms have to be able to quantify the subjective expectations (Castelfranchi & Falcone, 1998) of the users, based on their past experiences on the online community. Therefore, it is desirable that the reputation mechanisms can provide personalized evaluations, based on the subjective criteria of the users engaged in an online interaction.

Finally, we have to consider the memory of the reputation system (Marsh, 1994). We know that the larger the number of ratings used in the evaluation of reputation values, the better the predictability of the mechanism. However, since the reputation values are associated with human individuals and humans change their behavior over time, it is desirable to disregard very old ratings. Thus, it is desirable that the predicted reputation values are closer to the current behavior of the individuals rather than their overall performance.

The desiderata described here are by no means universally applicable to any kind of online community. For example, the requirement for minimal initial reputations can be relaxed if our online community consists of people who know each other (Winter, 1999).

SPORAS: A REPUTATION MECHANISM FOR LOOSELY CONNECTED COMMUNITIES

Keeping in mind the discussion presented in the previous section, Sporas provides a reputation service based on the following principles:

1. New users start with a minimum reputation value and they build up reputation during their activity on the system.
2. The reputation value of a user never falls below the reputation of a new user.
3. After each transaction, the reputation values of the involved users are updated according to the feedback provided by the other parties, which reflect their trustworthiness in the latest transaction.
4. Two users may rate each other only once. If two users happen to interact more than once, the system keeps the most recently submitted rating.
5. Users with very high reputation values experience much smaller rating changes after each update. This approach is similar to the method used in the Elo (Elo, 1978) and the Glicko (Glickman, 1999) systems for pairwise ratings.
6. The algorithm adapts to changes in the users' behaviors. Thus ratings must be discounted over time so that the most recent ratings have more weight in the evaluation of a users's reputation.

From an algorithmic perspective this system has to satisfy the following requirements:

1. It has to require small computational space and time for the updates of the reputation predictions.
2. The system has to be adaptively controlled, predicted, and supervised using the accuracy of the rating predictions. The ratings submitted after each interaction have to be compared with the predicted ones, and their difference used as an input to the recursive function.
3. Old predictions have to be discounted and the system has to be a biased estimator of the most recent behavior.

Based on these requirements it is proposed to estimate the time varying reputation of a user using the following algorithm.

New users start with reputation values equal to 0 and can advance up the maximum of 3000, so let's call the reputation range $D = 3000$. The reputation ratings, W_i, vary from 0.1 for terrible to 1 for perfect. The minimum reputation rating, W_i, is set to be above 0, unlike the beginners' reputations $R_o = 0$, so that once a user has received at least one rating, then the user's reputation value will be necessarily greater than zero, even if that rating was

the minimum one. That way, a user is always worse off if he/she switches identities. Suppose that at time t = 1, a user with reputation R_{i-1} is rated with a score W_i by another user with reputation R_i^{other}. Let $E_i = R_{i-1}/D$. At equilibrium, E_i can be interpreted as the expected value of W_i, though early in a user's activity it will be an underestimate. Let $\theta > 1$ be the effective number of ratings considered in the reputation evaluation. It is then proposed that the Sporas formula eq. (1), which is a recursive estimate of the reputation value of a user at time t = i, given the user's most recent reputation, R_{i-1}, the reputation of the user giving the rating, R_i^{other}, and the rating W_i:

$$R_i = R_{i-1} + \frac{1}{\theta} \cdot \Phi(R_{i-1}) R_i^{other}(W_i - E_i)$$

$$\Phi(R_{i-1}) = 1 - \frac{1}{1 + e^{-(R_{i-1} - D)\frac{}{\sigma}}}$$

$$E_i = R_{i=1}/D$$

Equation 1. Sporas formulae.
Recursive computation of the reputation value at time = t and computation of the damping function Φ.

The parameter σ is the acceleration factor of the damping function Φ, which slows down the changes for very reputable users. The smaller the value of σ, the steeper the damping factor Φ is. The behavior of the damping function Φ with different value of σ is shown in Figure 1, which plots Φ for 10 equidistant values of σ, ranging from D/100 to 10D/100. The value of σ is chosen so that the Φ remains above 0.9 for all users whose reputation is below 3/4 of D. Therefore, it can be calculated that $\sigma \leq (0.25/\ln 9)/D = 0.11$.

Equation (1) shows that the incremental change in the reputation value of a user receiving a rating of W_i for user R_i^{other}, is proportional to the reputation value R_i^{other} of the rater.

$$\begin{aligned}
R_i &= R_{i-1} + \frac{1}{\theta} \Phi(R_{i-1}) R_i^{other}(W_i - E_i) \\
&> R_{i-1} - \frac{1}{\theta} \Phi(R_{i-1}) R_i^{other} \frac{R_{i-1}}{D}, \quad \text{since } \frac{1}{\theta} \Phi(R_{i-1}) R_i^{other} W_i > 0 \\
&> R_{i-1} - \frac{1}{\theta} \Phi(R_{i-1}) D \frac{R_{i-1}}{D}, \quad \text{since } R_i^{other} \leq D \\
&> R_{i-1} - \frac{1}{\theta} R_{i-1} = \frac{\theta - 1}{\theta} R_{i-1}, \quad \text{since } \Phi(R_{i-1}) \leq 1 \\
&> 0, \quad \text{since } \theta > 1.
\end{aligned}$$

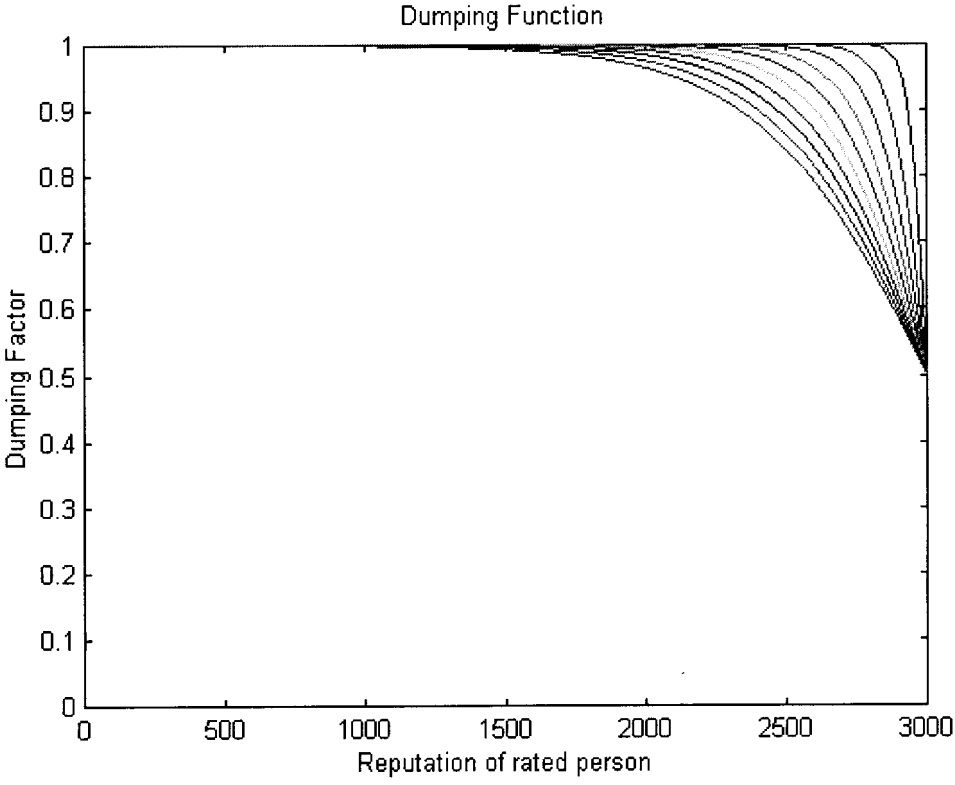

FIGURE 1. Damping function. The behavior of the damping function Φ with 10 different values of σ, ranging from D/100, to 10D/100.

Also, if $R_{i-1} = D - x$, and $x \geq 0$

$$
\begin{aligned}
R_i &= D - x + \frac{1}{\theta}\Phi(R_{i-1})R_i^{other}(W_i - (D - x)/D) \\
&\leq D - x + \frac{1}{\theta}\Phi(R_{i-1})R_i^{other}(1 - (D - x)/D), \quad \text{since } W_i \leq 1 \\
&\leq D - x + \frac{1}{\theta}\Phi(R_{i-1})Dx/D, \quad \text{since } R_i^{other} \leq D \\
&\leq D - x + \frac{x}{\theta}, \quad \text{since } \Phi(R_{i-1}) \leq 1 \\
&\leq D, \quad \text{since } \theta > 1, \quad \text{and } x \geq 0
\end{aligned}
$$

Equation 2. Proof of lower and upper bounds of the recursive estimates of R_i.

In addition, as one can see from eq. (2), the recursive estimates of R_i are always positive, thus no user can have a rating value lower than that of a beginner, and those estimates have an upper bound of D.

The predicted rating of a user is expressed as the current reputation value over the maximum reputation value allowed in the system. Thus, if the submitted rating for a user is less than his/her desired rating value, the reputation value of the user decreases.

Equation (1) is a simple machine-learning algorithm which guarantees that if W_i is a stationary time series of observations, then it will give asymptotic convergence of R_i to the actual $\bar{R}$ (Figure 2) and the speed of the convergence is controlled by the learning factor $1/\theta$ (Rumelhart et al., 1986).

The value of $1/\theta$ determines how fast the reputation value of the user changes after each rating. The smaller the value of $1/\theta$ the longer the memory of the system. Thus, just like credit card history (FairIsaac), even if a user enters the system with a very low reputation, if his/her reliability improves, his/her reputation value will not suffer forever from the past poor behavior.

Reliability of the Reputation Value Predictions

Using a similar approach to the Glicko system, a measure of the reliability of the users' reputations has been incorporated into the system. The reliability is measured by the reputation deviation (RD) of the estimated reputations. The recursively estimated RD of the algorithm is an indication of the predictive power of the algorithm for a particular user. Therefore, a high RD can mean either that the user has not been active enough to be able to make a more accurate prediction for his/her reputation, or that the user's behavior has indeed a lot of variation, or even that the user's behavior is too controversial to be evaluated the same way by his/her raters. As was explained in the previous sections, one assumes that the user's reputation is also an indication of how reputable the user's opinion about others is. Therefore, the change in the reputation of a person receiving a rating is positively related to the reputation of a user who submits the rating eq. (1). Thus, the RD of a user's reputation indicates the reliability of that user's opinion for the users he/she rates.

Since the reputation update function is computed according to eq. (1), if one ignores the damping factor Φ, then RD can be computed as a weighted LS problem (Madsen & Holst, 1998) defined by

$$RD_i^2 = \lfloor \lambda \cdot RD_{i-1}^2 + (R_i^{other}(W_i - E_i))^2 \rfloor / T_0 ,$$

Equation 3. Recursive computation of the reputation deviation (RD) at time $t = i$.

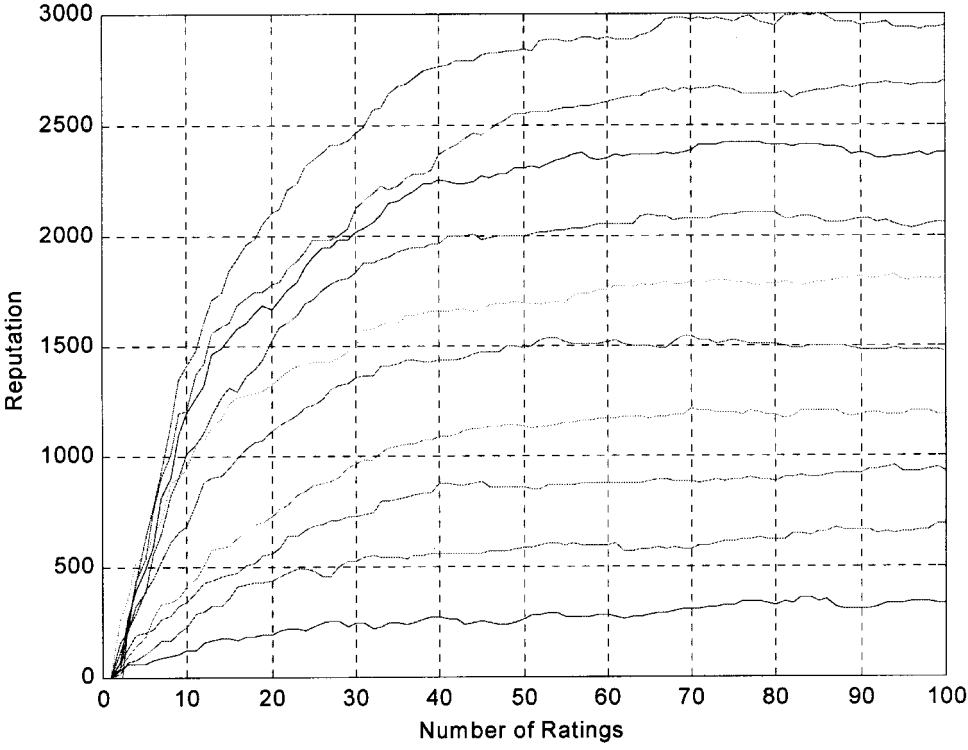

FIGURE 2. Buildup of reputation. Simulation with 10 different users over 100 ratings with $\theta = 10$.

Where $\lambda < 1$ is a constant and T_0 is the effective number of observations. Since λ is a constant, T_0, which will be set equal to θ of eq. (1), can be calculated as

$$T_0 = \sum_{i=0}^{\infty} \lambda^i = \frac{1}{1-\lambda}.$$

Equation 4. Computation of the effective number of observations with a forgetting factor of λ.

Equation (3) is a generic recursive estimation algorithm of recursive least squares (RLS) with a forgetting factor of λ, which can be used for online estimations (Madsen & Holst, 1998). So, eq. (3) estimates recursively the average square deviation of the predictions of eq. (1) over the last T_0 ratings. In fact, if $\lambda = 1$ and $T_0 = i$, then RD_i^2 is precisely the average square deviation of the predictions of eq. (1) over the last T_0 ratings. However, one incorporates the forgetting factor λ in order to ensure that the most recent ratings have more weight than the older ones. Note that eq. (1) is not the solution of the RLS eq. (3), as would be the case if one was trying to minimize RD for a given λ. However, eq. (3) is a recursive estimator of the RD, given eq. (1).

With the proper choice of the initial values of a RLS algorithm, with or without a forgetting factor, the algorithm's predictions will coincide with the predictions of an offline least square fitting of the user's data, if the user's behavior has a stationary, nonperiodic mean and standard deviation (Madsen & Holst, 1998). In this case though, one will deliberately choose initial conditions that estimate a beginner's reputation to be minimal with a maximum standard deviation. One needs these initial conditions so that there is no incentive for a user to switch identities. Thus the beginners start with a RD of D/10 and the minimum RD is set to D/100, and, as it was explained above, their initial reputation value is set to 0.

With these initial values, one is ensured that the reputation value of any user will always be strictly higher than the reputation value of a beginner eq. (2). Therefore, user A, for example, who has been consistently receiving poor scores will end up having both a low reputation and a low RD, but the reputation value of A will always be higher than a beginners reputation.

However, the low RD of user A identifies him/her as an established untrustworthy person. Therefore, the combination of a low reputation value and a low RD may incite user A to switch identities. However, it is not clear that A will be better off by switching identities, because although he will start with a larger RD, due to the uncertainty about his/her trustworthiness, A's reputation will be lower than before switching identities. Therefore, if A intends to improve him/herself, he/she is better off by preserving his identity, because he/she can grow it faster. If A intends to keep behaving improperly, he/she does not really have a big incentive to switch identities, because as a beginner, he/she will be treated equally unfavorably.

The major limitation of Sporas is that it treats all the new users very unfavorably. This unfavorable treatment is a necessary trade-off, if one wants to allow total anonymity for the users of an online community.

HISTOS: A REPUTATION MECHANISM FOR HIGHLY CONNECTED COMMUNITIES

Sporas, described in the previous section, provides a global reputation value for each member of an online community. This information is associated with the users as a part of their identity. However, different people groups have different standards and they tend to trust the opinions of the people who have the same standards with themselves. For example, if I am about to transact online with someone I have never interacted before, if a trusted friend of mine has transacted with the same user before, I am probably willing to trust my friend's opinion about that user more than the opinions of a few people I have never interacted with before. Likewise, the PGP Web of Trust (Garfinkel, 1994) uses the idea that we tend to trust someone trusted by someone we trust more than we trust a total stranger.

Following a similar approach, it was decided to build Histos, which is a more personalized reputation system compared to Sporas. In Weaving a Web of Trust (Khare & Rifkin, 1997), entities are trusted if there is a connected path of PGP-signed webpages between every pair of users. In the case of Histos, which is a pairwise rating system, one also has to consider the reputation ratings connecting the users of the system. So unlike Sporas, the reputation of a user in Histos depends on who makes the query, and how that person rated other users in the online community.

One can represent the pairwise ratings in the system as a directed graph (Figure 3), where nodes represent users and weighted edges represent the most recent reputation rating given by one user to another, with the arrow pointing towards the rated user. If there exists a connected path between two users, say from A to A_L, then one can compute a more personalized reputation value for A_L.

When user A_0 submits a query for the Histos reputation value of user A_L, one performs the following computation:

The system uses a Breadth First Search algorithm to find all the directed paths connecting A_0 to A_L that are of length less than or equal to N. As described above, one only cares about the chronologically q most recent ratings given to each user. Therefore, if one finds more than q connected paths taking one to user A_L, one is interested only in the most recent q paths with respect to the last edge of the path.

One can evaluate the personalized reputation value of A_L if one knows all of the personalized reputation ratings of the users connecting to A_L in the path. Thus, one creates a recursive step with at most q paths with length at most $N - 1$.

If the length of the path is only 1, it means that the particular user, A_L, was rated by A_0 directly. Then, the direct rating given to user A_L is used as

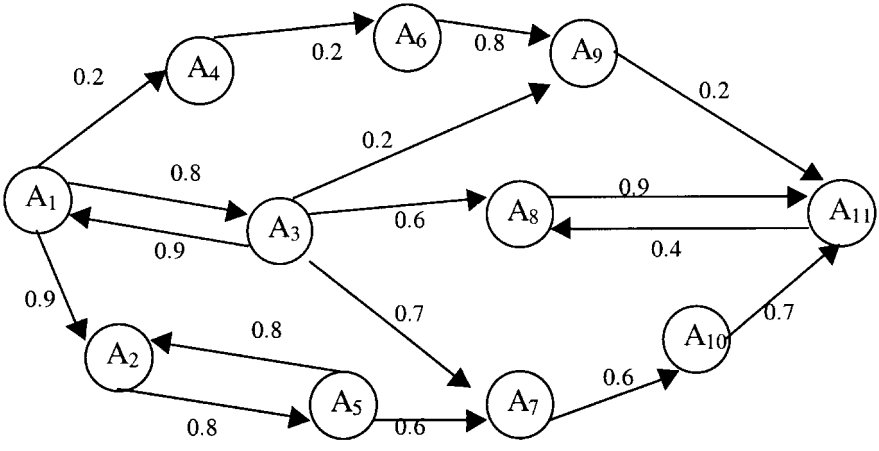

FIGURE 3. Rating paths between users A_1 and A_{11}.

the personalized reputation value for user A_0. Thus, the recursion terminates at the base case of length 1.

For the purpose of calculating the personalized reputation values, one uses a slightly modified version of the reputation function of Sporas eq. (1). For each user A_k, with $m_k(n)$ connected paths going from A_0 to A_k, one calculates the reputation of A_k as follows:

Let $W_{jk}(n)$ denote the rating of user A_i for user $A_k(n)$ at a distance n from user A_0, and $R_k(n)$ denote the personalized reputation of user $A_k(n)$ from the perspective of user A_0.

At each level n away from user A_0, the users $A_k(n)$ have a reputation value given by

$$R_k(n) = D \cdot \sum_j (R_j(n-1) \cdot W_{jk}(n)) \Big/ \sum_j R_j(n-1)$$

$\forall jk$, such that $W_{jk}(n) \geq 0.5$

$$m_k(n) = \deg(A_k(n)) = |W_{jk}(n)|,$$

Equation 5. Histos formulae.

where deg $(A_k(n))$ is the number of connected paths from A_0 to $A_k(n)$ and D is the range of reputation values eq. (1). The users $A_k(n)$, who have been rated directly by user A_0 with a rating $W_{1\,k}(1)$, have a reputation value equal to

$$R_k(0) = D \cdot W_{1\,k}(0).$$

Equation 6. Histos formulae.

As was explained above one is only interested in the q most recent ratings for each user, so if $m_k(n)$ is larger than q, one picks from those edges the subset with the q most recent ratings.

Consider, for example, Figure 4 at level 2. The personalized reputation of user $A_1(3)$ will be

$$R_1(3) = D \cdot (R_1(2) \cdot W_{1\,1}(2) + R_2(2) \cdot W_{2\,1}(2) + R_3(2) \cdot W_{3\,1}(2)) / \times (R_1(2) + R_2(2) + R_3(2)).$$

Equation 7. Histos query for user $A_1(3)$ in Figure 4.

Since all the paths at both Level 0 and Level 1 have rating contributions from only one source per target, it means that the personalized reputation of $A_1(3)$ is

$$R_1(3) = D \cdot (W_{1\,1}(1) \cdot W_{1\,1}(2) + W_{2\,2}(1) \cdot W_{2\,1}(2) + W_{3\,3}(1) \cdot W_{3\,1}(2)) / \times (W_{1\,1}(1) + W_{2\,2}(1) + W_{3\,3}(1)).$$

Equation 8. Result of a Histos query for user $A_1(3)$ in Figure 4.

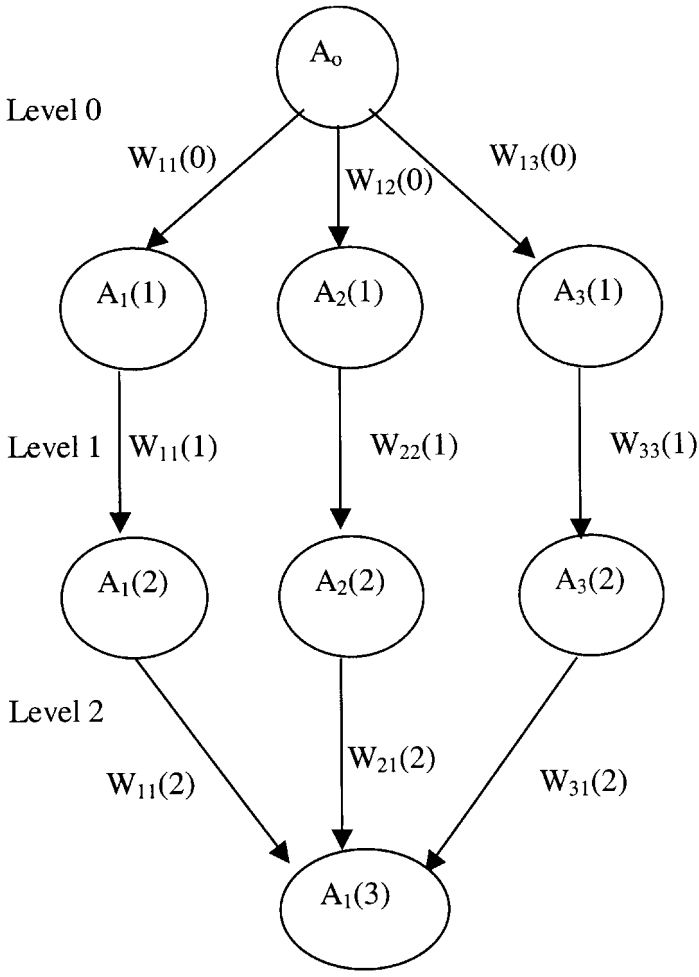

FIGURE 4. Example of a Histos query. User A_0 makes a Histos query for user $A_1(3)$. The query finds three unique paths of reputable ratings and evaluates the personalized reputation of $A_1(3)$ from the perspective of A_0.

Histos needs a highly connected graph. If there does not exist a path from A_0 to A_L with length less than or equal to N, one falls back to the simplified Sporas reputation mechanism.

EVALUATION

Simulations

To evaluate the reputation mechanisms, one applies the algorithms in four simulations. In the first simulation one evaluates the convergence speed

of the algorithm. One has 100 users with uniformly distributed real reputations. Each user starts with a minimum reputation at 300, initial RD of 300, and can have a minimum RD of 30. The users are matched randomly in each period of the simulation and get rated by each other according to their actual performance. Each user's performance is drawn from a normal distribution with a mean equal to its real reputation and a standard deviation of 100. One assumes that one has reached equilibrium when the average square error of the reputation scores of users from their real reputations falls below 0.01. In this specific simulation, the system reached equilibrium after 1603 ratings- in other words after each user has made on average 16 transactions. Figure 5 shows the reputation values for users 0, 1, and 8 over time until the average square error becomes $0.01D^2$. At the time of equilibrium, users 0, 1, and 8 with real reputations 327.1, 1458.1, and 746.8, respectively, had reached reputation values of 691.6, 1534.1, and 991.0, with RD's 116.5, 86.7, and 103.4, respectively. The equilibrium was reached after receiving 15,

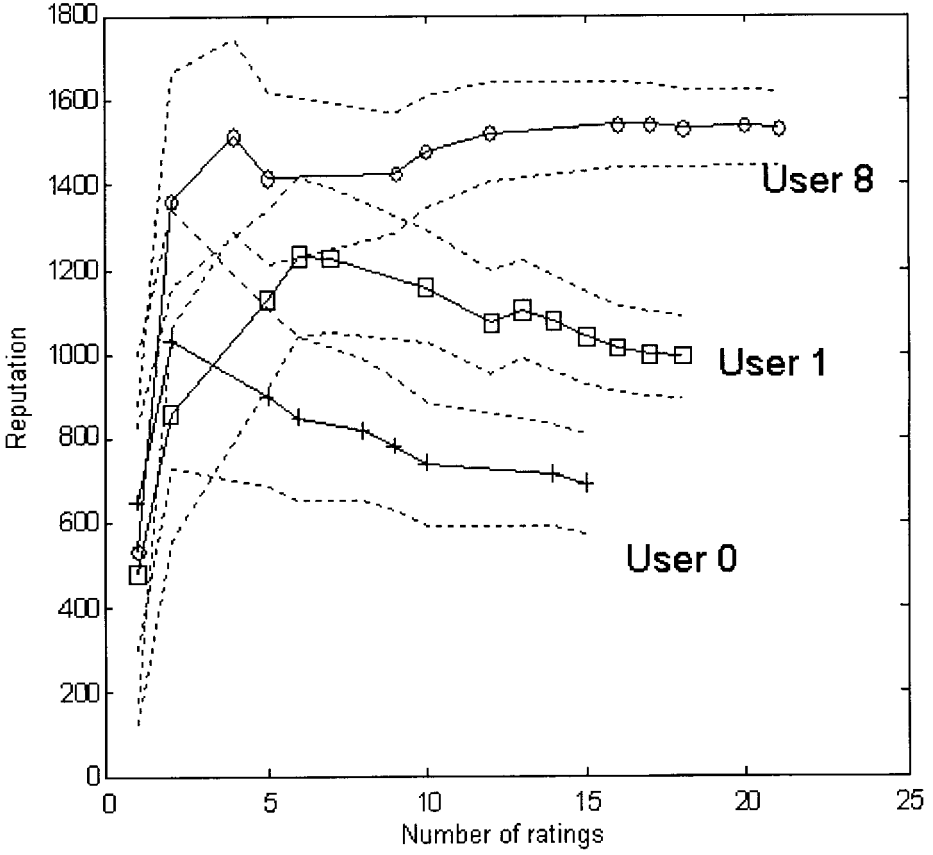

FIGURE 5. Bootstrapping. Simulation of 100 users with uniformly distributed reputations. The simulation achieves an average square error in 1603 ratings. The dotted lines around each one of the three curves, shows the RD of that user.

21, and 18 ratings, respectively. Therefore, this system can reach equilibrium very quickly. As one can see from the results of the three users and Figure 5, the users with high reputations are estimated with a better precision than users with low reputations.

In the second simulation, one shows a user who joins the marketplace, behaves reliably until he/she reaches a high reputation value, and then starts abusing his/her reputation to commit fraud. Thus, the user's ratings start dropping because of his/her unreliable behavior. During the first 1/3 of his/her interactions, the user performs with a reputation of 0.8D. During the last 2/3 of his/her interactions, the user behaves with a reputation of 0.3. The user receives ratings, which are normally distributed around his/her actual performance, with a standard deviation of 0.1. The reputations of the raters of the user are drawn from a uniform distribution with a range D. The effective number of ratings in Sporas is $\theta = 30$. One plots on the same graph the reputation values that the user would have if he/she received the same ratings in a simplistic reputation system, where the reputations are evaluated as the average of all the ratings given to the user, as is the case with the

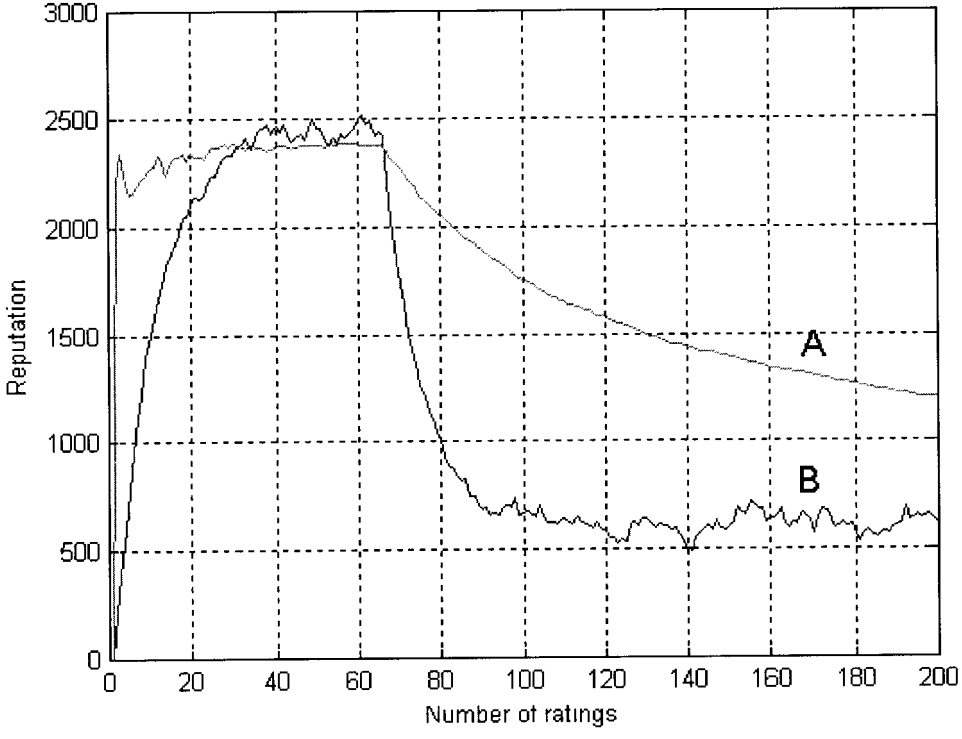

FIGURE 6. Abuse of prior performance. The curve A, shows the computed average reputation value of a user who starts very reputable and then starts behaving as an untrustworthy person. The curve B shows the effect of the same behavior using the Sporas reputation mechanism.

reputation mechanism of Amazon auctions. As one can see from the graph, although the user keeps receiving consistently lower scores for a time period twice as long his/her reputable period, he/she still perceives a reputation of 0.6D if he/she is evaluated using the averages method of Amazon.com. Hence, in this case, the user can take advantage of his/her past good ratings for quite a long time and keep deceiving people about his/her actual reliability. However, as one can see in Figure 6, if the user is evaluated using Sporas, it takes less than 20 ratings to adjust the reputation of the user to his/her new performance.

In the third simulation, the effect of collusion by two users is presented. In this experiment, both users get rated every other time by one of their friends with a perfect score. Like the previous experiment, one plots the reputations of both users evaluated on this system and on a system like Amazon's. The actual performance of the two users is 900 and 600 (out of 3000), respectively. As one can see in Figure 7, on the simplistic reputation system they actually manage to raise their reputations to 1781 and 1921, respectively, while with our algorithms, their reputations reflect their actual performance by letting them achieve reputation values of 619 and 960,

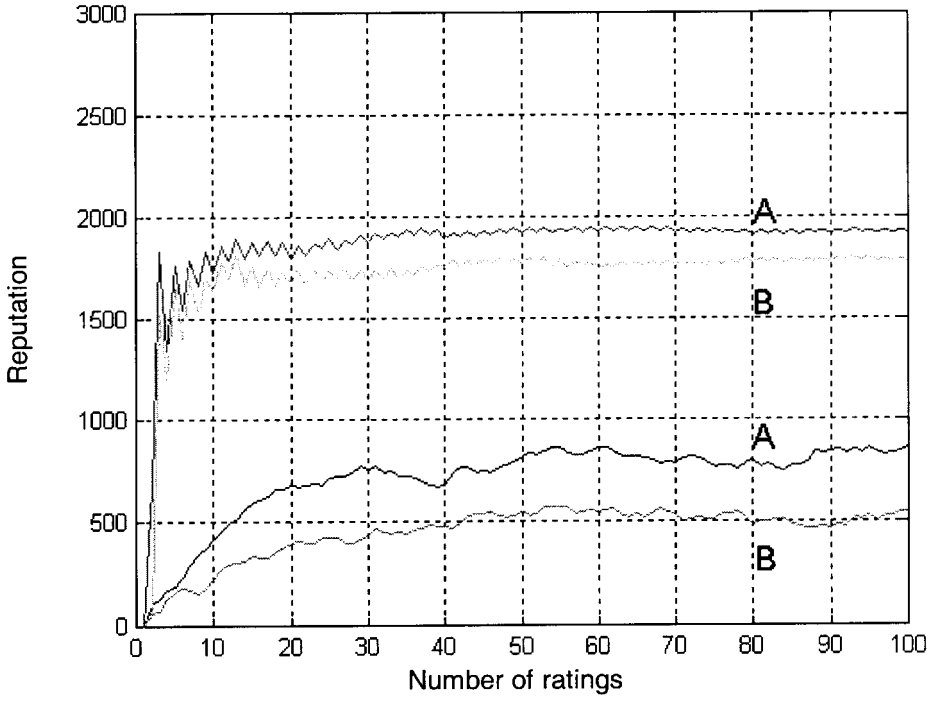

FIGURE 7. Collusion between two users. A and B collude and rate each other perfectly on every other transaction. User A has a real reputation of 900 and User B a reputation of 600. With simple averages they achieve reputations of 1921 and 1781, while with Histos, for a user who has never interacted with them before directly, they achieve reputations of 960 and 619, respectively.

respectively. The reputations of the other users and the ratings they submit are created the same way as in the previous experiment (Figure 6).

Evaluating Sporas on eBay User Data

To evaluate the Sporas algorithm with real-user data, it was decided to spider the Feedback Forum of eBay, and use the actual eBay ratings with the algorithm. Feedback pages were spidered for 7269 eBay users using a recursive spidering tool. The spidered process was initiated from the most recent feedback page of a random eBay user, and from there on it recursively downloaded the feedback pages of everyone who rated that user and kept going like that until the process was terminated.

The spidering tool kept a queue of the extracted feedback URLs in its memory, and explored those URLs in a Breadth First Search manner. Due to the design of the eBay feedback forum, for many of these users only a fraction of their actual feedback forum data managed to be spidered, because the additional pages were considered one level below in the tree structure. Therefore, instead of using eBay's summary data, the total number of transactions was recomputed, positive, neutral, and negative comments, based on the data that one managed to collect through the spidering process. Thus, in the calculations some of the old data is missing for several of the users because the feedback pages on eBay are sorted in reverse chronological order. Each feedback page on eBay has at most 25 comments, and the incomplete data are for user with more than one page; therefore, even

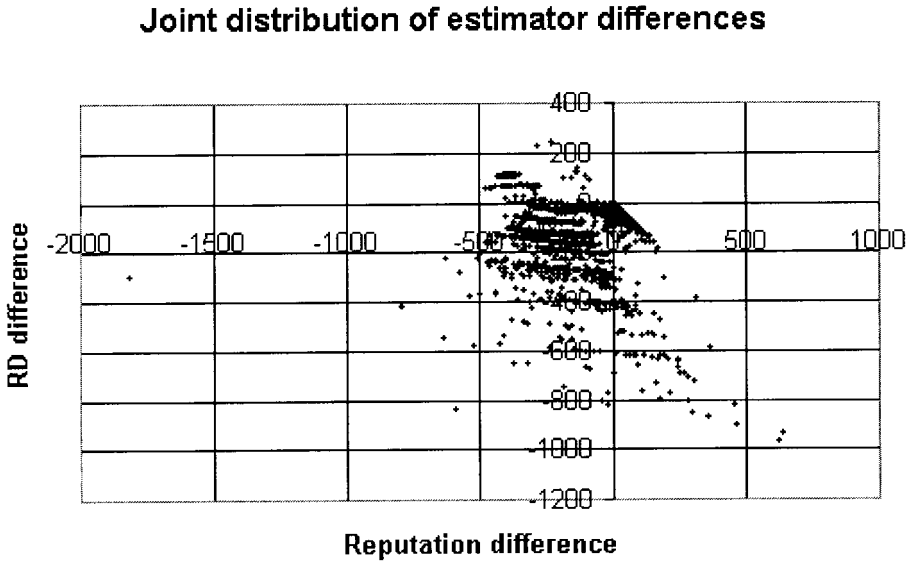

FIGURE 8. Joint distributions of estimated differences. The difference of the estimated reputations from the computed reputations, and the estimated vs. $\bar{R}$Ds and computed $\hat{R}$Ds for each one of the eBay users.

without the missing data one had at least 25 ratings for each one of those users. In the evaluation process below, the effective number of observations was set to 10, so the 25 most recent ratings of the users with missing data was a good enough sample for their most recent behavior.

Since users on eBay are rated with either 1 or 0 or -1, the ratings had to be scaled to a [0,1] interval so they were replaced with 1, 0.5, and 0, respectively. For each one of the users, the mean and the standard deviation of his/her performance was calculated in the data that was collected. Then for each one of those users, the Sporas algorithm was applied and attempts to predict the reputation and reputation deviation (RD) was tried in a recursive manner as described in the fourth section.

Figure 8 shows the joint distribution of $\hat{R} - \bar{R}$ and $\widehat{RD} - \overline{RD}$, where $\hat{R}$ is the reputation value and $\widehat{RD}$ the reputation deviation estimate using Sporas, and $\bar{R}$ is the average reputation value and $\overline{RD}$ the reputation deviation computed from the sampled transactions of the same user. Figure 9 shows $\hat{R}$ versus $\bar{R}$ and Figure 10 shows $\widehat{RD}$ versus $\overline{RD}$.

As we can see from Figure 8 and Figure 9, the Sporas algorithm, in general, underestimates the sampled reputation of a user. This is clearly seen in Figure 9, where we can see that users with the same sampled reputation $\bar{R}$, end up having different estimations for $\hat{R}$. This difference depends on how recently the user committed his/her transitions with low scores. Therefore, the time dependency of our recursive estimation, ensures that users who have been trustworthy in their latest transactions, rather than their earliest ones, will have higher scores than others who performed well in the past, but

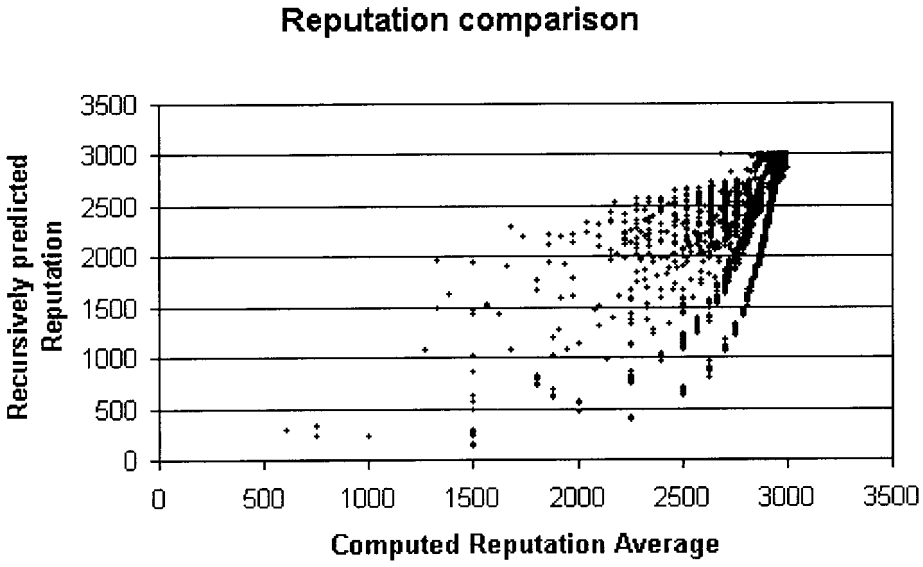

FIGURE 9. Estimated vs. computed reputation values for the eBay users.

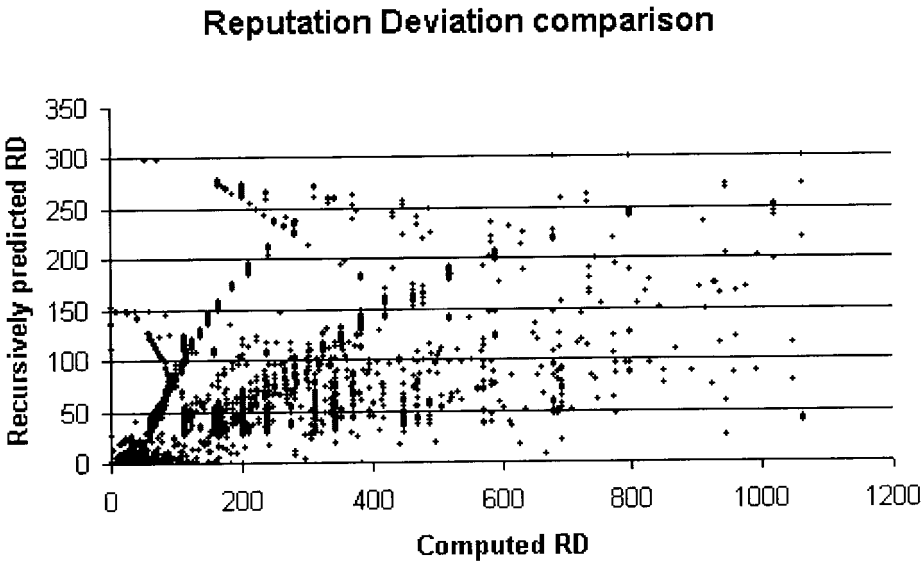

FIGURE 10. Estimated vs. computed RD for the eBay users.

started getting low feedback scores lately, even if their linear average is exactly the same.

In addition, as one can see from Figure 8 and Figure 10, the Sporas algorithm, in general, underestimates the sampled reputation deviation of a user, compared to the reputation deviation computed from the sample of the user's transactions. Observing this result was expected, because the recursive estimation of the reputation deviation discounts older deviations and tries to make its predictions based on the most recent performance. However, in some cases one does estimate a larger reputation deviation than the one observed over the whole sample. This happens when the user exhibits a varying performance during his/her most recent transactions rather than his/her earlier ones. Since attempts are being made to make predictions based on the more recent data, the overestimation of the reputation deviation in these cases is the desired behavior.

CONCLUSION

Two collaborative reputation mechanisms have been developed that establish reputation ratings for users of online services. The proposed solutions are able to face the problems and fulfill the desiderata described in the fourth section. Incorporating reputation mechanisms in online communities may induce social changes in the way users participate in the community. As one has seen in the case of eBay, the scale of its rating system made the users reluctant to give low scores to their trading partners, which reduces the

value of the rating system. Thus, a successful reputation mechanism, besides having high prediction rates and being robust against manipulability, has to make sure that it does not hurt the cooperation incentives of the online community.

In future work, it is the plan to build a reputation brokered agent mediated knowledge marketplace, where buying and selling agents will negotiate for the exchange of intangible goods and services on their owner's behalf. The agents will be able to use current reputation scores to evaluate the utility achieved for a user under each candidate contract. The author want to study how intelligent the pricing algorithms of the agents have to be, so that one achieves economic efficiency in conjunction with pairwise reputation mechanisms.

NOTES

1. OnSale Exchange was later transformed to Yahoo Auctions, and Yahoo implemented the same rating mechanism as eBay.
2. The Greek and Turkish Cypriot communities have been estranged since the Turkish invasion in 1974. There are no direct phone lines between the two sides of the seize-fire line.

REFERENCES

Amazon.com Auctions. http://auctions.amazon.com

Better Business Bureau. http://www.bbb.org

Bizrate. http://www.bizrate.com

Castelfranchi, C., and R. Falcone. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. *Workshop in Deception, Fraud and Trust in Agent Societies, Second International Conference on Autonomous Agents* (*Agents '98*), St. Paul, MI, May 9–13.

Chavez, A., and P. Maes. 1996. Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology* (*PAAM'96*), London, UK.

Cyprus List. http://kypros.org/Lists/Cyprus

Donath, J. 1998. *Identity and deception in the virtual community, communities in cyberspace*, Kollock, P. and Smith, M. (eds.). London: Routledge.

eBay. http://www.ebay.com

Elo, A. E. 1978. *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing, Inc.

FairIsaac Co. http://www.fairisaac.com

Foner, L. 1997. Yenta: A multi-agent, referral based matchmaking system. *First International Conference on Autonomous Agents* (*Agents '97*), Marina del Rey, CA. New York: ACM Press.

Friedman, E., and P. Resnick. 1998. The social cost of cheap pseudonyms: Fostering cooperation on the Internet. In *Proceedings of the 1998 Telecommunications Policy Research Conference*, Alexandria, VA. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Garfinkel, S. 1994. *PGP: Pretty good privacy*. Sebastopol, CA: O'Reilly and Associates.

Glickman, M. 1999. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48(3):377–394. Oxford: Blackwell Publishers.

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

Kollock, P. 1999. *The production of trust in online markets, advances in group processes*, eds. Lawler, E. J., Macy, M., Thyne, S., and Walker, H. A. Vol. 16. Greenwich, CT: JAI Press.

Madsen, H., and J. Holst. 1998. *Lecture Notes in Non-linear and Non-stationary Time Series Analysis*, Institute of Mathematical Modelling (IMM), Technical University of Denmark, Lyngby, Denmark.

Marsh, S. P. 1994. Formalising trust as a computational concept. Ph.D. Thesis, University of Stirling, Stirling, Scotland, UK.

OnSale. http://www.onsale.com

Reagle, J. M., Jr. 1996. Trust in a cryptographic economy and digital security deposits: Protocols and policies. Master Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Recreational Software Advisory Council: http://www.rsac.org/

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, Chapel Hill, NC, 175–186.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. In Rumelhart, D. E., and McClelland, J. L. (eds.), *Parallel distributed processing, Volume 1: Foundations*, pp. 318–362. Cambridge MA: MIT Press Bradford Books.

Shardanand, U., and P. Maes. 1995. Social information filtering: Algorithms for automating "word of mouth." Human factors in computing systems. In *CHI'95 Conference Proceedings*, 210–217. New York: ACM.

Wang, D. 1999. Market maker: An agent-mediated marketplace infrastructure. MEng Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Winter, M. 1999. The role of trust and security mechanisms in an agent-based peer help system. *Workshop in Deception, Fraud and Trust in Agent Societies, Third International Conference on Autonomous Agents* (*Agents '99*), Seattle, WA, May 1–4.