

Collaborative Reputation Mechanisms for Online Communities

by

Giorgos Zacharia

S.B. Computer Science and Engineering, 1998

S.B Mathematics, 1998

Massachusetts Institute of Technology

Submitted to the Program of Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences
at the

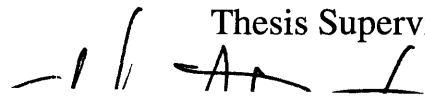
Massachusetts Institute of Technology

September, 1999

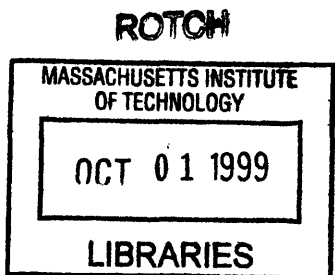
© Massachusetts Institute of Technology, 1999. All Rights Reserved.

Author _____
Program of Media Arts and Sciences
August 6, 1999

Certified by _____
Pattie Maes
Associate Professor of Media Technology
Thesis Supervisor



Accepted by _____
Stephen A. Benton
Allen Professor of Media Arts and Sciences
Chairperson
Departmental Committee on Graduate Students



Collaborative Reputation Mechanisms for Online Communities

by

Giorgos Zacharia

Submitted to the Program of Media Arts and Sciences,
School of Architecture and Planning, on August 6, 1999,
in partial fulfillment of the requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

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. I will investigate two complementary reputation mechanisms that 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, I believe that they may have applicability in other types of electronic communities such as chatrooms, newsgroups, mailing lists etc.

Thesis Supervisor: Dr Pattie Maes

Title: Associate Professor in Media Arts and Sciences

Collaborative Reputation Mechanisms for Online Communities

by
Giorgos Zacharia

The following people have served as readers for this thesis:

Thesis Reader _____

Judith Donath
Assistant Professor of Media Arts and Sciences
NEC Career Development Professor of Computers and Communications
MIT Media Lab

Thesis Reader _____

Mark Glickman
Assistant Professor
Department of Mathematics and Statistics
Boston University

Table of Contents

1. Introduction.....	15
2. Related Work.....	21
3. The problem of trust in online communities.....	29
3.1 Consumer-to-Consumer Electronic Marketplaces.....	29
3.2 Discussion forums.....	30
4. Desiderata for online reputation systems.....	33
5. Sporas: A reputation mechanism for loosely connected communities.....	37
Reliability of the Reputation value predictions.....	44
6. Histos: A reputation mechanism for highly connected communities.....	49
7. Implementation.....	55
8. Evaluation.....	61
8.1 Simulations.....	61
8.2 Evaluating Sporas on eBay user data.....	66
8.3 Survey of eBay users.....	71
9. Conclusion.....	77

List of Figures

<i>Number</i>	<i>Page</i>
Figure 1 Damping function.	40
Figure 2 Buildup of reputation.....	42
Figure 3 Rating paths between users A_1 and A_{11}	50
Figure 4 Example of a Histos query.	54
Figure 5 Histos Visualization.....	56
Figure 6 MarketMaker	57
Figure 7 Mailing List Experiment.....	58
Figure 8 Bootstrapping.....	62
Figure 9 Abuse of prior performance.....	64
Figure 10 Collusion between two users.....	65
Figure 11 Joint distributions of estimated differences.....	69
Figure 12 Estimated vs. Computed Reputation values for the eBay users.....	69
Figure 13 Estimated vs. Computed RD for the eBay users	70
Figure 14 Distribution of eBay Ratings.....	74
Figure 15 Distribution of Amazon Auction ratings.....	76

List of Tables

<i>Number</i>	<i>Page</i>
Table 1 Comparison of online reputation systems.	27
Table 2 Questions asked in the eBay Feedback Forum survey	72
Table 3 Responses to our survey on eBay's feedback forum.....	73

Acknowledgments

I am grateful to my advisor, professor Pattie Maes, for giving me the exciting opportunity to work in her group and on this project in particular. It all started in a brainstorming session for the next generation of our Agent mediated Electronic Commerce infrastructure. After the session, both Pattie and I knew that I would be working on "reputations" for my thesis. I need to thank her for her invaluable guidance in my research, and especially for the research questions she kept asking me. Those questions made this thesis possible. She has been the perfect advisor, the perfect supervisor, the perfect boss.

I also have to thank both my readers, Dr Mark Glickman and Dr Judith Donath for their valuable comments. The Glicko filter made sure that the thesis was mathematically precise and Judith made sure that the social issues were addressed properly. All the remaining errors and omissions are my fault.

I want thank my family for their love and support all these years and especially during the last few days I was writing this thesis. More particularly I have to thank my parents for their endless encouragement, my brother Γιάννης for the math brainstorming and my little sister Μαργαρίτα (she will always be the little one) for the wake up calls after allnighters.

I need to thank Özlem for her support, for proofreading the document, asking critical questions and for helping me keep my focus.

This section would be incomplete if I did not thank my αδράνεια/κβ buddies for making sure I would take my breaks on schedule.

Chapter 1

Introduction

"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" [15].

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, the 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 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 Kasbah [1], MarketMaker [18], eBay [7] and OnSale Exchange [20] 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 would be to incorporate in the system 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 [19]. In "Trust in a Cryptographic Economy" [21], 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.

We have developed methods through which we can automate the social mechanisms of reputation for electronic communities. We have implemented an early version of these reputation mechanisms in Kasbah [1]. 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 [1]. 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 Chapter of this thesis describes the related work in the domain of rating systems and reputation mechanisms. The third Chapter outlines the requirements for a successful reputation mechanism for online communities. The fourth Chapter describes problems specific to electronic marketplaces and online discussion forums. The fifth and sixth Chapters describe two reputation

mechanisms we have designed and evaluated. The seventh Chapter describes the implementation of our reputation mechanisms in the context of an Agent mediated Electronic Marketplace and an online discussion list. The eighth Chapter evaluates the mechanisms using simulations and user data from eBay and Amazon auctions. The last Chapter is the conclusion of the thesis and the outline of our future work.

Chapter 2

Related Work

We can divide the related work on reputation systems into two major categories: non-computational reputation systems like the Better Business Bureau Online [3] 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 chapter 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 [10] 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 [11], Weaving a web of Trust [15], and the Platform for Internet Content Selection (PICS)[24] (such as the Recreational Software Advisory Council [22]) 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 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 then can 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 [25] and GroupLens [23]. 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 [8] and the Glicko [14] 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 the 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 scores of the two players are positively related to their rating difference before the game.

BizRate [4] 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 web site. 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 Exchange¹ [20], eBay [7] and Amazon [1]. 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 six 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

¹ OnSale Exchange was later transformed to Yahoo Auctions, and Yahoo implemented the same rating mechanism as eBay.

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 the 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 the 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 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 [12].

Table 1 Comparison of online reputation systems. In the “Pairwise rating” column we indicate whether the ratings are bi-directional or one-directional, and who submits ratings. In the “Personalized Evaluation” column we indicate whether the ratings are evaluated in a subjective way, based on who makes the query.

<i>System</i>	<i>Pair-wise rating</i>	<i>Personalized Evaluation</i>	<i>Textual comments</i>
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

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 [12].

Chapter 3

The problem of trust in online communities

3.1 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.

3.2 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 [9].

However, the perceptions about the reputations of the users among themselves can be very different and subjective. My favorite example of this phenomenon is the "Cyprus List", an English-speaking bicomunal mailing list hosted at MIT. This mailing list has been the only open communication forum between the two communities in Cyprus for the several decades² 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 problem.

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 highly members who come from their own community, while they consider the members coming from the opposing side as fanatic 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

² The Greek and the Turkish Cypriot communities have been estranged since the Turkish invasion in 1974. There are no direct phone lines between the two sides of the cease fire line.

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.

Chapter 4

Desiderata for online reputation systems

While the previous Chapters discussed reputation mechanisms that have some interesting qualities, we believe 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 [9][12][16]. 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 a positive reputation mechanisms, the beginners are subject to mistreatment by the rest of the community because nobody knows if they are in fact new 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 [12]. However, in such a model we need to charge for names in the first place and enforce persistent pseudonymous identities [12]. Despite the benefits of this model, we 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 [6], 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, we decided to a first desideratum 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 [12] 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 newsgroups 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 transactions 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, we have 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 users 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 [5] 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 [19]. 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 [27].

Chapter 5

Sporas: A reputation mechanism for loosely connected communities

Keeping in mind the discussion presented in the previous Chapter, 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 [8] and the Glicko [14] 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 user's reputation.

From an algorithmic perspective our 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 we propose 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 lets call our 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 $R_0=0$, so that once a user has received at least one rating, then the users 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 $t=i$, 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 our reputation evaluation. We then propose the Sporas formula [Equation 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 :

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

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

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

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 values 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 $\frac{3}{4}$ of D . Therefore, it can be calculated that $\sigma \leq \frac{0.25}{\ln 9} D = 0.11$

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

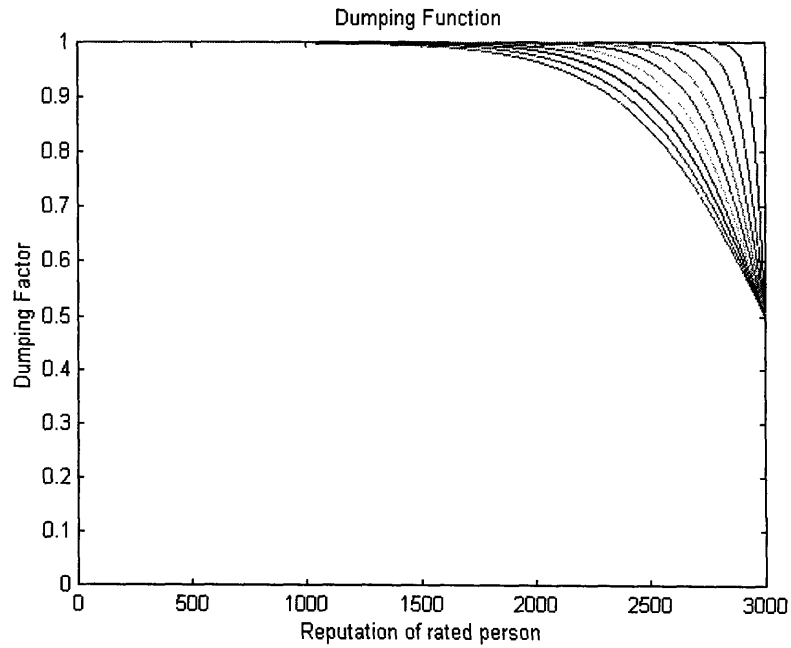


Figure 1 Damping function.

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

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

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

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

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

In addition, as we can see from Equation 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 .

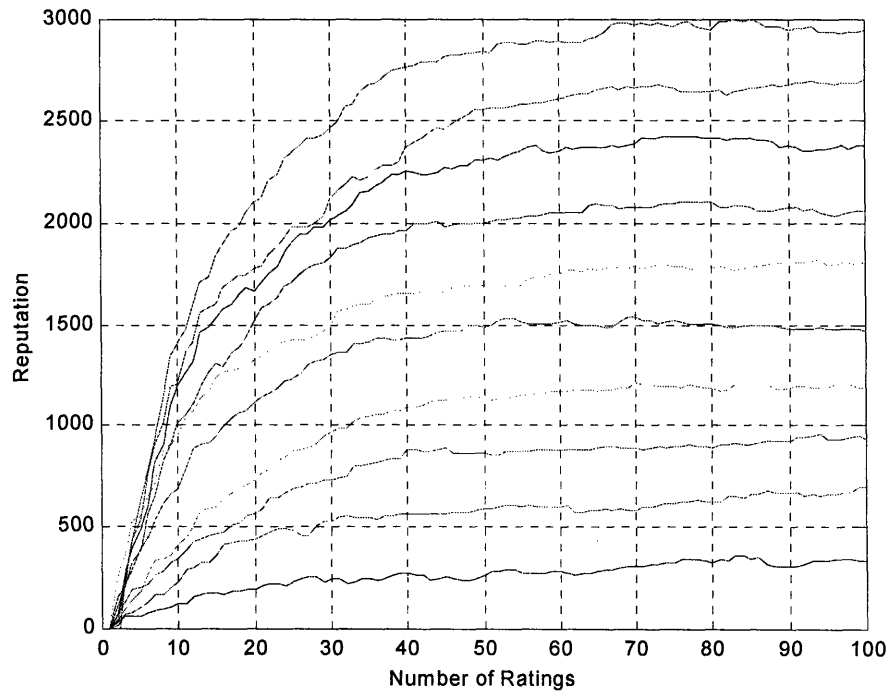


Figure 2 Buildup of reputation
Simulation with 10 different users over 100 ratings
with $\theta=10$

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 that guarantees that if W_i is stationary time series of observations, then it will give asymptotic convergence of R_i to the actual \bar{R} and the speed of the convergence is controlled by the learning

factor $\frac{1}{\theta}$. [25]

The value of $\frac{1}{\theta}$ determines how fast the reputation value of the user changes after each rating. The smaller the value of $\frac{1}{\theta}$ the longer the memory of the system. Thus, just like credit card history [10], 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 we have incorporated into the system a measure of the reliability of the users' reputations. 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 we explained in the previous Chapters, we assume 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 [Equation 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 Equation 1, if we ignore the damping factor Φ , then RD can be computed as a weighted LS problem [17] defined by:

$$RD_i^2 = \left[\lambda \bullet RD_{i-1}^2 + (R_i^{other} (W_i - E_i))^2 \right] / T_o$$

Equation 3 Recursive computation of the Reputation Deviation (RD) at time t=i.

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

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

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 [17]. So Equation 3 estimates recursively the average square deviation of the predictions of Equation 1, over the last T_o ratings. In fact, if $\lambda = 1$ and $T_o = i$, then RD_i^2 is precisely the average square deviation of the predictions of Equation 1, over the last T_o ratings. However, we incorporate the forgetting factor λ in order to ensure that the most recent ratings have more weight than the older ones. Note that Equation 1 is not the solution to the RLS Equation 3, as would be the case if we were trying to minimize RD for a given λ . However, Equation 3 is a recursive estimator of the RD, given Equation 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, non-periodic mean and standard deviation [17]. In our case though, we will deliberately choose initial conditions that estimate a beginner's reputation to be minimal with a maximum standard deviation. We need these initial conditions so that there is no incentive for a user to switch identities. So 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, we ensure that the reputation value of any user will always be strictly higher than the reputation value of a beginner [Equation 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 very unfavorably all the new users. This unfavorable treatment is a necessary trade off, if we want to allow total anonymity for the users of an online community

Chapter 6

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 [13] 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, we decided to build Histos, which is a more personalized reputation system compared to Sporas. In *Weaving a Web of Trust* [15], 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, we also have 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.

We 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 we can compute a more personalized reputation value for A_L .

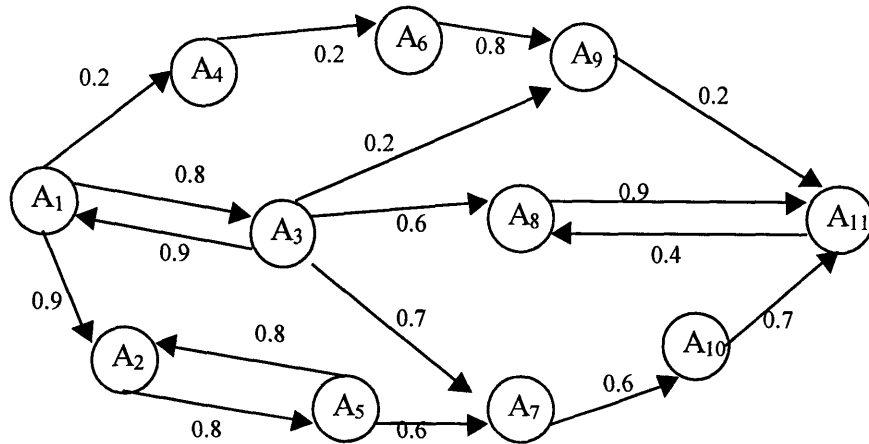


Figure 3 Rating paths between users A_1 and A_{11}

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

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

We can evaluate the personalized reputation value of A_L if we know all of the personalized reputation ratings of the users connecting to A_L in the path. Thus, we create 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_o directly. Then, the direct rating given to user A_L is used as the personalized reputation value for user A_o . Thus, the recursion terminates at the base case of length 1.

For the purpose of calculating the personalized reputation values, we use a slightly modified version of the reputation function of Sporas [Equation 1]. For each user A_k , with $m_k(n)$ connected paths going from A_o to A_k , we calculate the reputation of A_k as follows:

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

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

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

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

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

Equation 5 Histos formulac

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

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

Equation 6 Histos formulac

As we explained above we are interested only in the q most recent ratings for each user, so if $m_k(n)$ is larger than q , we pick 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:

$$R_1(3) = D \cdot (R_1(2) \cdot W_{11}(2) + R_2(2) \cdot W_{21}(2) + R_3(2) \cdot W_{31}(2)) / (R_1(2) + R_2(2) + R_3(2))$$

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:

$$R_1(3) = D \cdot (W_{11}(1) \cdot W_{11}(2) + W_{22}(1) \cdot W_{21}(2) + W_{33}(1) \cdot W_{31}(2)) / (W_{11}(1) + W_{22}(1) + W_{33}(1))$$

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

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

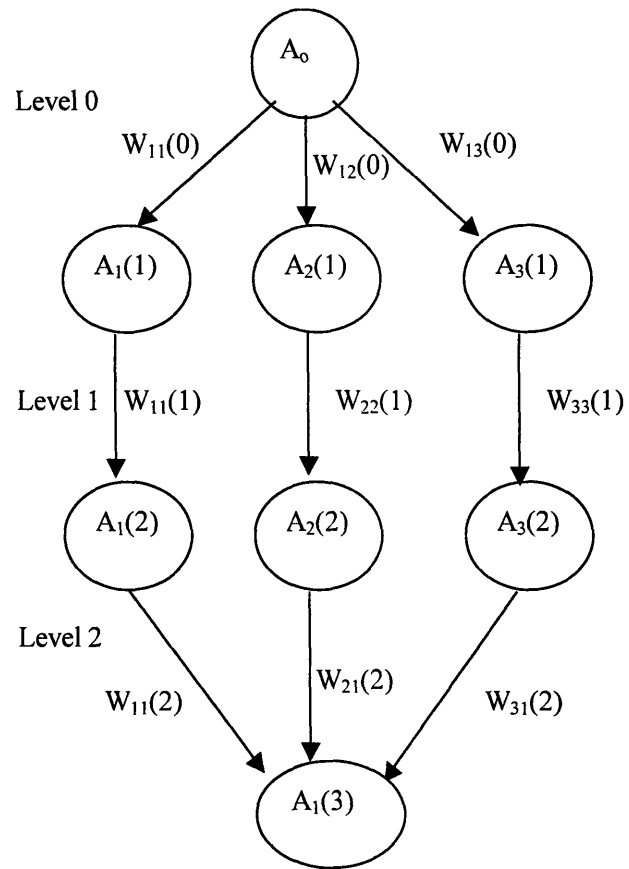


Figure 4 Example of a Histos query.
 User A_0 makes a Histos query for user $A_1(3)$. The query finds 3 unique paths of reputable ratings and evaluates the personalized reputation of $A_1(3)$ from the perspective of A_0 .

Chapter 7

Implementation

The Reputation Server was implemented as a plugin to the MarketMaker [Figure 6] [17], a Consumer-to-Consumer Ecommerce site at MIT. The same architecture was also used for an email experiment [Figure 7]. As described above, MarketMaker is a web-based Agent Mediated Marketplace. Whenever two agents make a deal on the marketplace, the users are notified about the terms of the deal and the contact information of their counterpart and the buyer and the seller are then asked to rate each other based on their performance in the particular deal. The ratings may be submitted within 30 days from the moment the deal was reached, and the two users are prompted to rate each other whenever they login on the marketplace. The user may rate his/her counterpart as Horrible, Difficult, Average, Good or Great. The ratings are translated to their respective numerical values of 0.2, 0.4, 0.6, 0.8 and 1 and are submitted to the backend database.

When the user browses the marketplace he can see the various buying or selling agents with the reputation values of their owners [Figure 6]. The reputation values are presented both as numerical values and as colored bars. Since the number of users of MarketMaker is still very small, the personalized reputation

values are evaluated in real time, through a wrapper called by the CGI script, which generates the html of the webpage. However, if we had more users the calculation of the personalized reputation values might become too slow to be done in real time. In that case we would use a daemon, which updates the reputations of all users who are affected by a newly submitted rating, and caches the results in the database so that it can return the requests faster. When the user clicks on the image giving the reputation score, he/she is given a directed graph [Figure 5] with which he/she can visualize the ratings structure used to evaluate the reputation of the user he is looking at. The global (Sporas) reputation values can always be calculated in real time, because of the recursive nature of its update functions.

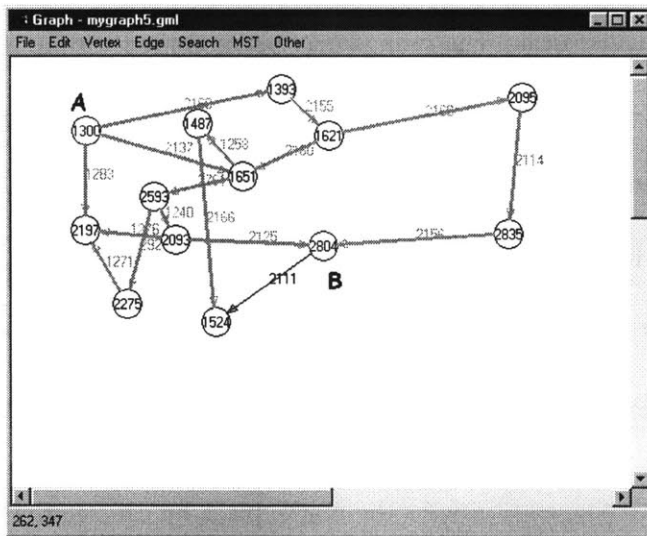


Figure 5 Histos Visualization.
 User A with reputation makes a query about the Reputation of user B. The query is broadcasted across A's network of trusted users.

Music

5 Items forsale:

Record Type: CD
 Genre: Misc.
 Title: Christoph Stuebbe live act
 Artist:
 Condition: Good to own but not to play
 Description:
 Reputation: 2132 out of 3000

[Create agent to buy this item.](#)

Record Type: CD
 Genre: Rock
 Title: rush
 Artist: rush
 Condition: Used but not damaged
 Description:
 Reputation: 620 out of 3000

[Create agent to buy this item.](#)

Record Type: CD
 Genre: Misc.
 Title: Rush
 Artist: Rush
 Condition: Used but not damaged
 Description:
 Reputation: 1754 out of 3000

[Create agent to buy this item.](#)

Figure 6 MarketMaker

A user browses a category of CD products being sold on MarketMaker. The reputation of each user is included with the description of the product. A colored bar coding is used to visually represent the relative trustworthiness of each user. The length of the blue bar is proportional to the reputation of the user, and the length of the yellow bar is what that user is missing to achieve a perfect reputation.

[Show reputations](#) | [Hide reputations](#)

Sort by: [Author](#) | [Date](#) | [Topic](#)
[Chronologically](#) | [Most recent first](#)

Options: [Show author](#) | [Hide author](#)
[Back to main kwlobares page](#)

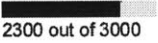
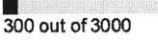
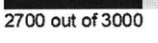
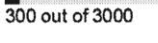
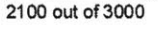
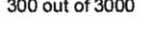
1. [Συγαρητρία Stous Diaskedazovtes](#) (25 lines)
From: Dimitrios Konstantakos <Konstand@MIT.EDU>

2300 out of 3000
2. [Χρῶνια Polla se Kwstavtivous & Elenes!](#) (12 lines)
From: Pantelis A Pittas <pantelis@MIT.EDU>

300 out of 3000
3. [Patates - Moment of Silence @ MIT entrance](#) (42 lines)
From: Petros Komodromos <petros@MIT.EDU>

2700 out of 3000
4. [Re: Housing/Sales - May 21](#) (21 lines)
From: Pantelis A Pittas <pantelis@MIT.EDU>

300 out of 3000
5. [Re: Housing/Sales - May 21](#) (29 lines)
From: Georgios Sarakinos <gsarakin@FAS.HARVARD.EDU>

2100 out of 3000
6. [\(no subject\)](#) (14 lines)
From: Nikolaos Prezas <prezas@MELBOURNE-CITY-STREET.MIT.EDU>

300 out of 3000

Figure 7 Mailing List Experiment.
This is a web interface of the Greek jokes list at MIT. Each user's posting is augmented with the user's reputation in the particular list. We used the same scale and color-coding as in the case of MarketMaker.

The backend and the interface of the reputation server were implemented in Visual C++ and the reputation values are stored in a Microsoft SQL server. The html of MarketMaker is created on the fly using servlets. Therefore the queries about the user's reputations are passed from the servlet being called to the reputation database. Likewise the submissions of fresh ratings are made through calls in the servlet code. In the case of the email experiment, the html is created by a collection of CGI-scripts that access the plain text archives of the mailing list. The queries and submissions of reputation scores from and to the database are called through the CGI-scripts themselves. In both the case of MarketMaker and the email experiment, the reputation data were stored in a standalone database, so a separate table was maintained with the necessary authentication and identification information of the marketplace transactions and the message postings respectively.

Chapter 8

Evaluation

8.1 Simulations

To evaluate the reputation mechanisms we applied the algorithms in four simulations. In the first simulation we evaluate the convergence speed of the algorithm. We have 100 users with uniformly distributed real reputations. Each user starts with 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. We assume that we have 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 8 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,

21 and 18 ratings respectively. Therefore our system can reach equilibrium very quickly. As we can see from the results of the three users and Figure 8, the users with high reputations are estimated with a better precision than users with low reputations.

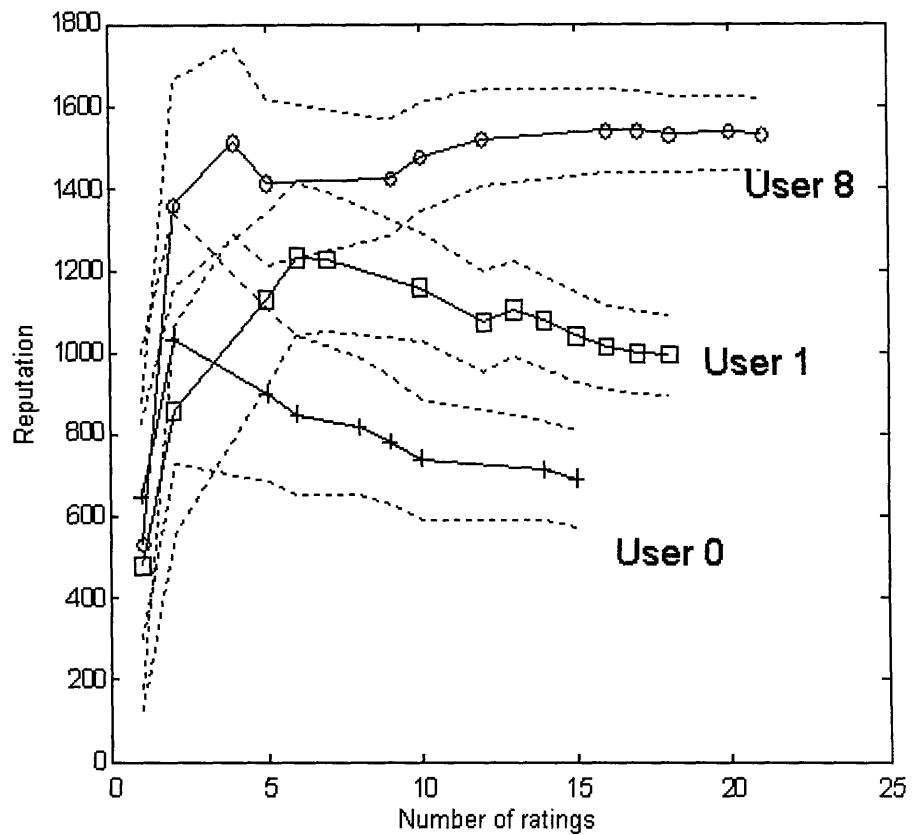


Figure 8 Bootstrapping.
Simulation of 100 users with uniformly distributed reputations. The simulation achieves an average square error in 1603 ratings. The dotted lines around each one of the 3 curves, shows the RD of that user.

In the second simulation we show 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 $\theta=30$. We plot 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 reputation mechanism of Amazon auctions. As we 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 preserves a reputation of $0.6D$, if he 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 a quite long time and keep deceiving people about his/her actual reliability. However, as we can see in Figure 9, 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.

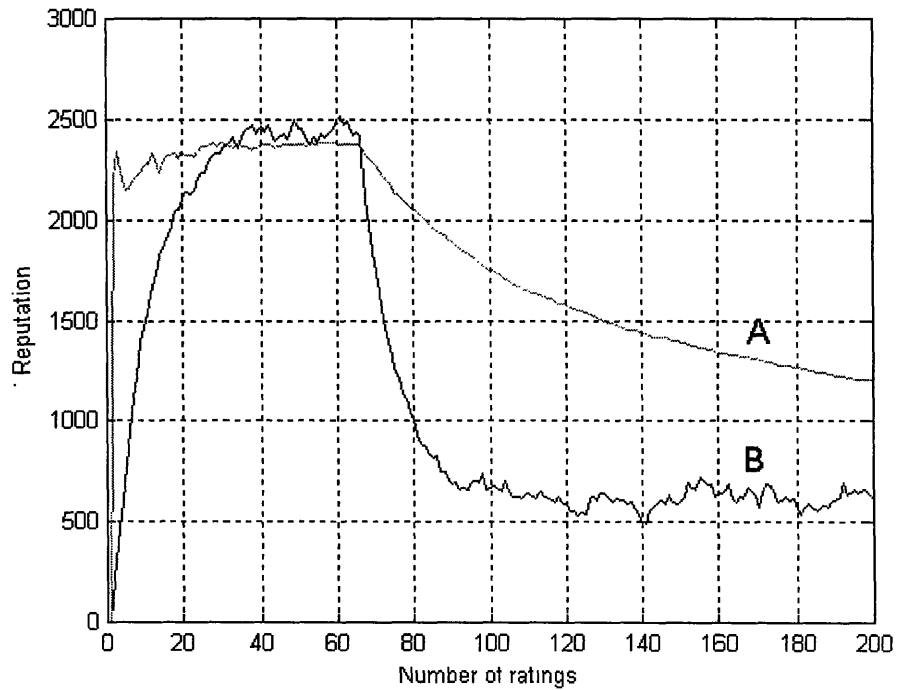


Figure 9 Abuse of prior performance.

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.

In the third simulation we present the effect of collusion by two users. In this experiment both users get rated every other time by one of their friends with a perfect score. Like the previous experiment we plot the reputations of both users evaluated on our system and on a system like Amazon's. The actual performance of the two users is 900 and 600 (out of 3000) respectively. As we can see in Figure 10, 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 respectively. The reputations of the other users and the ratings they submit are created the same way as in the previous experiment [Figure 9].

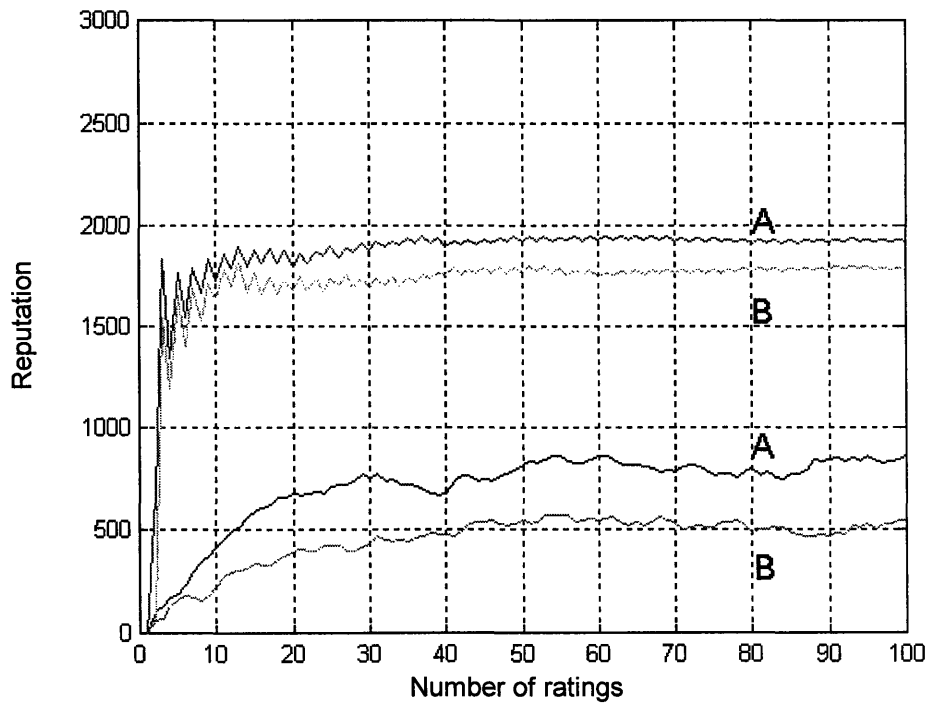


Figure 10 Collusion between two users.
A and B collude and rate each other perfectly every other transaction. User A has a real reputation of 900 and User B a reputation of 600. With simple averages they achieve reputations of 1781 and 1921, while with Histos, for a user who has never interacted with them before directly they achieve reputations of 619 and 960 respectively.

8.2 Evaluating Sporas on eBay user data

To evaluate the Sporas algorithm with real user data, we decided to spider the Feedback Forum of eBay, and use the actual eBay ratings with our algorithm. We spidered feedback pages for 7269 eBay users using a recursive spidering tool. We initiated the spidering process 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 we terminated the process.

The spidering tool, kept in its memory a queue of the extracted feedback URLs, and explored those URLs in a Breadth First Search manner. Due to the design of the eBay feedback forum, for many of these users we only managed to spider only a fraction of their actual feedback forum data, because the additional pages were considered one level below in the tree structure. Therefore, instead of using eBay's summary data, we recomputed the total number of transactions, positive, neutral and negative comments, based on the data we managed to collect through the spidering process. Thus, in our calculations we are missing some of the old data for several of our users, because the feedback pages on eBay are sorted in reverse chronological order. Each feedback page on eBay has a at most 25 comments, and our incomplete data are for users with more than one page, therefore even without the missing data we 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, we had to scale the ratings to a [0,1] interval so we replaced them with 1, 0.5 and 0 respectively. For each one of the users, we calculated the mean and the standard deviation of his/her performance in the data we have collected. Then for each one of those users, we applied the Sporas algorithm and tried to predict the Reputation and Reputation deviation (RD) in a recursive manner as described in Chapter 4.

Figure 11 shows the joint distribution of $\hat{R} - \bar{R}$ and $\hat{RD} - \bar{RD}$, where \hat{R} is the Reputation value and \hat{RD} the Reputation Deviation estimated using Sporas, and \bar{R} is the average Reputation value and \bar{RD} the Reputation Deviation computed from the sampled transactions of the same user. Figure 12 shows \hat{R} vs. \bar{R} and Figure 13 shows \hat{RD} vs. \bar{RD} .

As we can see from Figure 11 and Figure 12, the Sporas algorithm, in general, underestimates the sampled Reputation of a user. This is clearly seen in Figure 12 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 transactions 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 started getting low feedback scores lately, even if their linear average is exactly the same.

In addition, as we can see from Figure 11 and Figure 13, 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. We expected to observe this result, 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 we do 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 we are trying make our predictions based more on the recent data, the overestimation of the Reputation Deviation in these cases is the desired behavior.

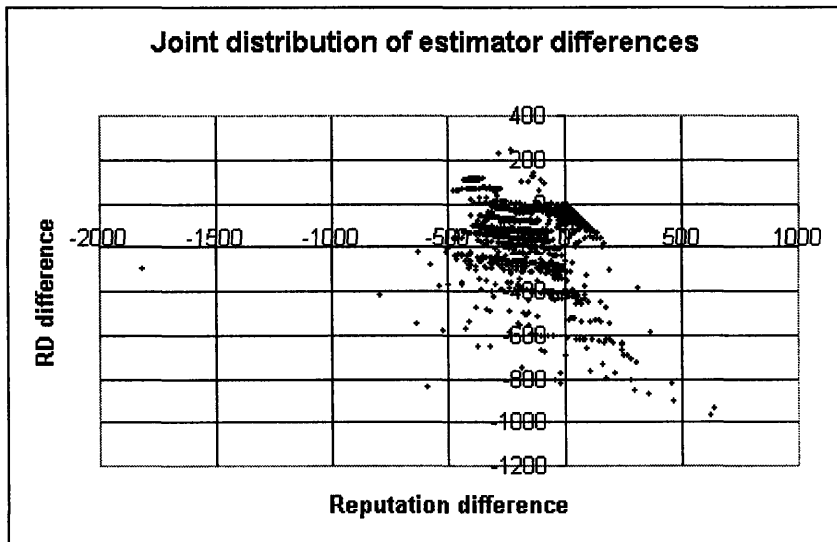


Figure 11 Joint distributions of estimated differences.
 The difference of the estimated Reputations from the computed Reputations, and the estimated vs. \hat{RD} s and computed RD s for each one of the eBay users

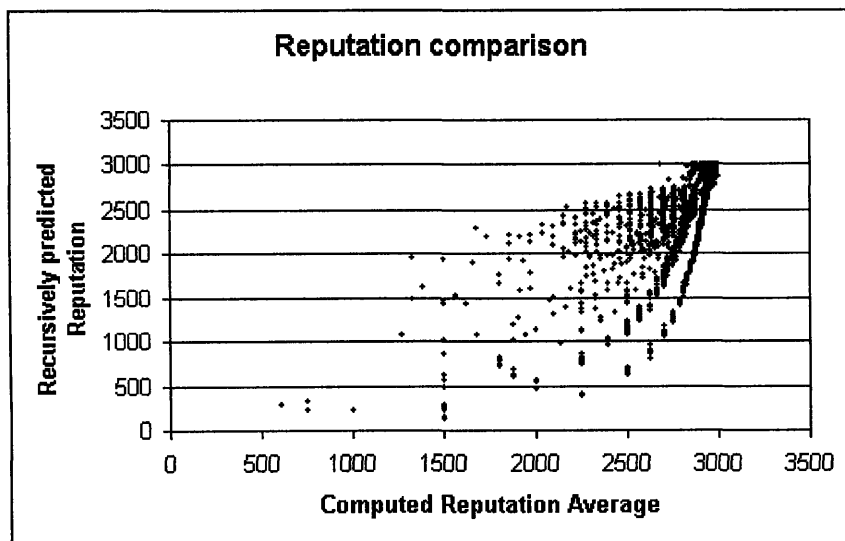


Figure 12 Estimated vs. Computed Reputation values for the eBay users

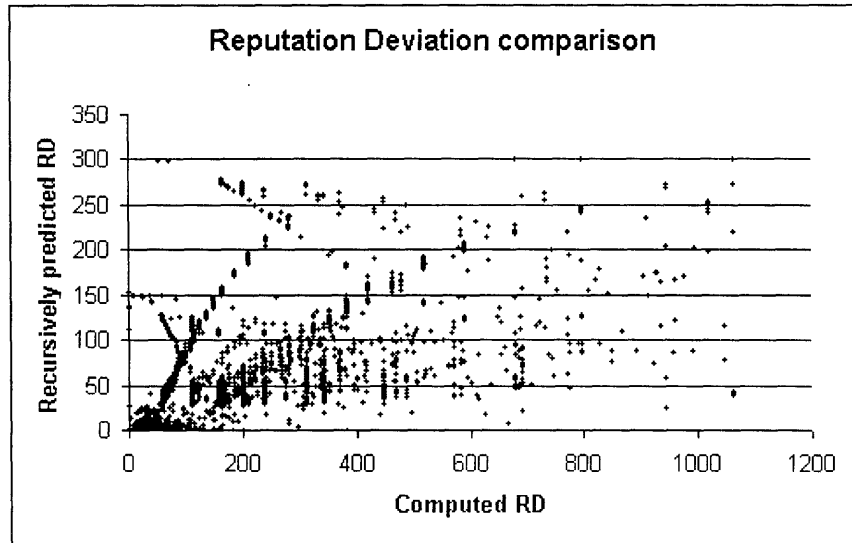


Figure 13 Estimated vs. Computed RD for the eBay users

8.3 Survey of eBay users

For completeness purposes we decided to conduct a survey on eBay users asking them key questions about the usefulness and their perception of eBay's feedback system. In order to pick participants for the survey, we spidered the feedback pages of eBay and selected randomly users whose eBay handle was an email address. Then we emailed those users a questionnaire of 7 questions. The questions can be found in Table 2. The users could respond either through email or through an html form. We received 2390 responses through the html form and 1263 responses through email. Due to time limitations we were able to analyze only the responses given through the html form because the ones received through email, had to be transformed manually to a machine-readable format. The results of the survey are shown on Table 3. The most common answer to question 6 was that the most important problem with eBay's Feedback is that users hesitate to give negative feedback because of fear of retaliation from their transaction partner. Therefore many users choose either to give positive feedback to a problematic transaction, or just not to leave any feedback at all. There was also very frequently the suggestion that in order to fix this problem, eBay should include in the feedback forum another statistic, namely the percentage of the transactions for which a user has received any rating at all. That way they would assume that a user with a very small number of ratings is in

fact an untrustworthy person, who has not been rated negatively in fear of retaliation.

Table 2 Questions asked in the eBay Feedback Forum survey

Q1	Do you check the eBay Feedback Forum before transacting with another eBay user? <input type="checkbox"/> ALWAYS <input type="checkbox"/> SOMETIMES <input type="checkbox"/> NEVER
Q2	Do you pay more attention to the eBay Feedback Forum information when you are buying/selling an expensive item? <input type="checkbox"/> YES <input type="checkbox"/> NO
Q3	Do you think the eBay Feedback Rating represents an accurate estimate of another user's trustworthiness? <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> SO-SO
Q4	Do you consider your own eBay Feedback Rating a valuable asset? <input type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> SO-SO
Q5	Do you wish you could bring your eBay Feedback Rating to other auction sites? <input type="checkbox"/> YES <input type="checkbox"/> NO
Q6	Any related suggestions or comments?
Q7	Do you want us to send you the results of the survey? <input type="checkbox"/> YES <input type="checkbox"/> NO

Table 3 Responses to our survey on eBay's feedback forum

	Q1		Q2		Q3		Q4		Q5		Q7	
Yes/Always	861	42%	1753	86%	1273	62%	1830	98%	1399	71%	1131	47%
So-So/ Sometimes	1076	53%			743	36%						
No/Never	106	5%	289	14%	23	1%	30	2%	567	39%		

This reluctance to submit negative scores is actually verified by the data we have collected by spidering the feedback forum of eBay. Out of 8342407 ratings, we found 8250012 (98.9%) positive ratings, 70398 (0.84%) neutral and 21997 (0.26%) negative scores.

This problem could be alleviated if the rating scale allowed more fine grained scores, for example 1 through 5 like the feedback system on Amazon's auctions [1]. If a user A engages in an unsuccessful transaction with user B, A would be willing to give a non perfect score, if the expected cost of retaliation from B is small enough not to affect A's future transactions severely. Therefore, the level of the expected retaliation is a critical factor in A's decision to give a non-perfect score to B. If the rating scale is binary, or even tertiary, then a non-perfect score is necessarily a much worse score than a perfect one. While if the rating system allows better score fine-grain, a non-perfect score can still be above average, and therefore will be considered better than the second best one in the binary or tertiary rating system. For example, if user A can rate B on a scale of 1 to 5, it

will be more willing to give a non-perfect score of 4 out of 5 (compared to 0 on a scale of -1, 0, 1), which is also the expected retaliation from B.

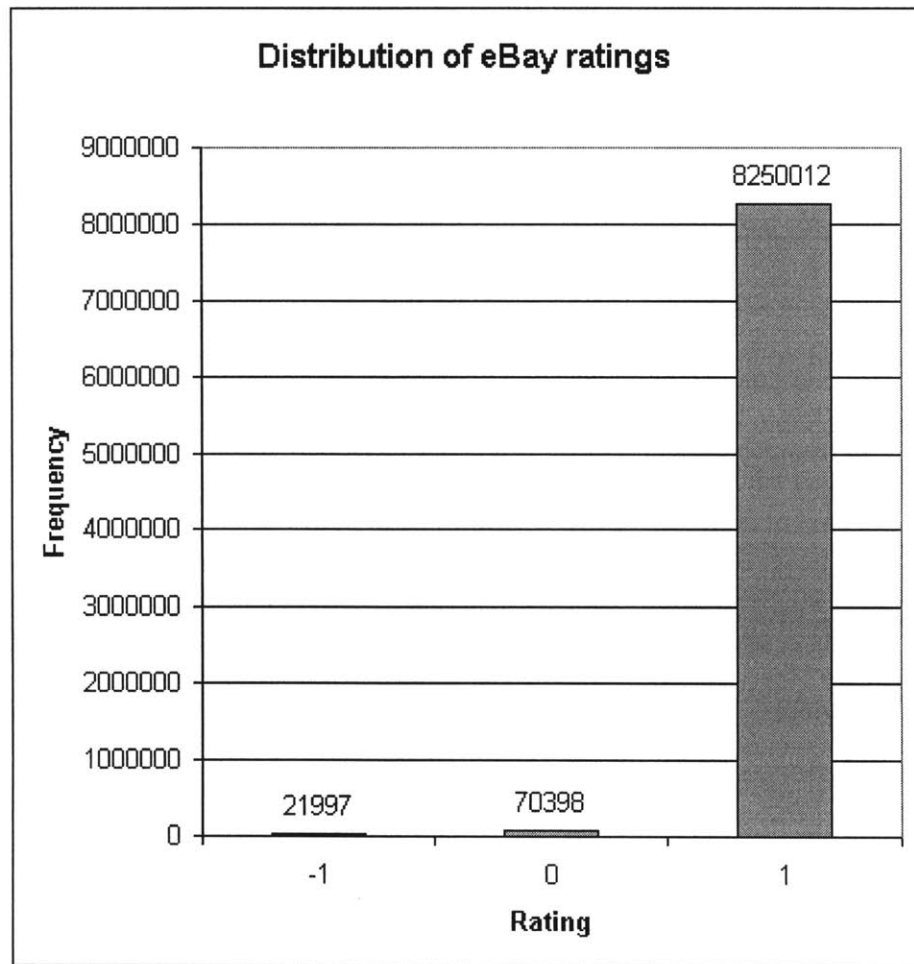


Figure 14 Distribution of eBay Ratings

In order to compare the difference in behavior when users can give ratings over a wider range, we spidered the Feedback forum of Amazon's auctions. We collected 13304 ratings, whose breakdown was 9607 (72.2%) 5's, 1893 (14.2%)

4's, 586 (4.4%) 3's, 392 (2.9%) 2's and 826 (6.2%) 1's. The difference is very significant; on eBay it would seem that 98.9% of the transactions were carried out without any problem, while on Amazon's auctions only 72.2% of the transactions were carried out without any problem.

These results suggest that fine-grained ratings may provide a solution to the fear of retaliation problem. As it was expected, the Amazon rating system resulted in significantly more 4's out of 5 than the total of -1's and 0's in the eBay rating system. Interestingly enough, there are also significantly more 1's, 2's and 3's out of 5 on Amazon than -1's and 0's on eBay. This may also suggest that the actual values of the ratings have a psychological effect on the users' perception about how bad a non-perfect rating is. In the case of eBay, the fact that the lowest score is actually negative may make it sound too harsh for the eBay users. If this is actually true, then we might expect a better distribution of ratings if the scale of ratings of eBay was from 1 to 3 instead of -1 to 1. Unfortunately, we do not have empirical data to examine this argument.

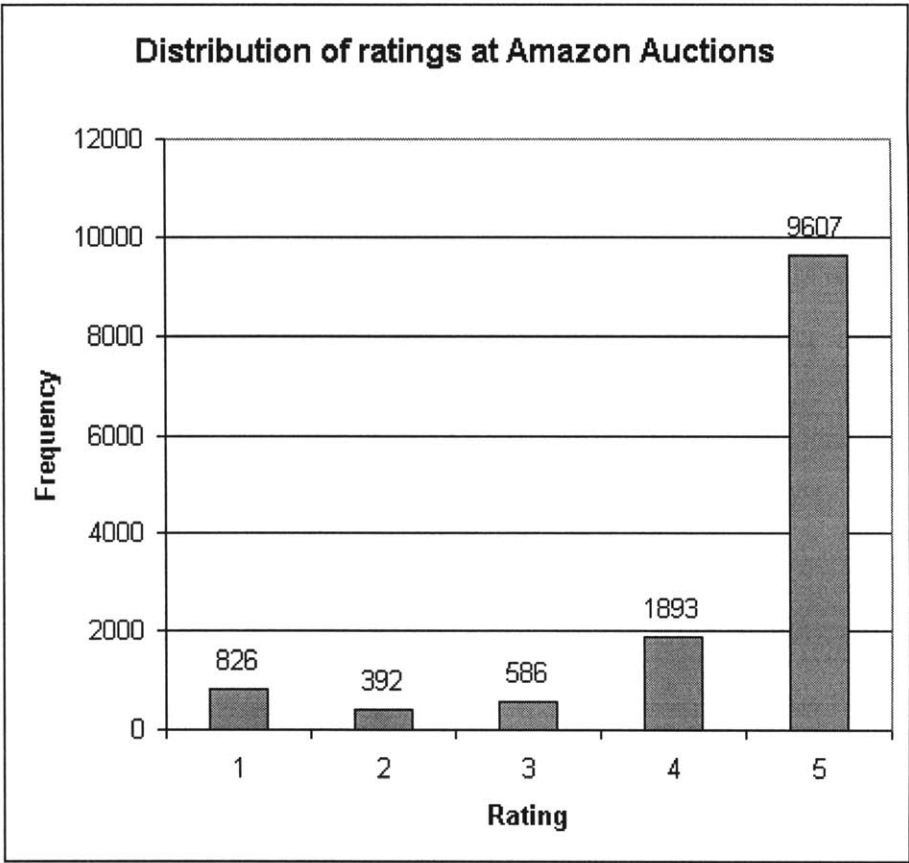


Figure 15 Distribution of Amazon Auction ratings

Chapter 9

Conclusion

We have developed two collaborative reputation mechanisms that establish reputation ratings for users of online services. The proposed solutions are able to face the problems and fulfill the desiderata described in Chapter 4. Incorporating reputation mechanisms in online communities may induce social changes in the way users participate in the community. As we have 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 our future work we 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. We want to study how intelligent the pricing algorithms of the agents have to be, so that we achieve economic efficiency in conjunction with pairwise reputation mechanisms.

References

- [1] Amazon.com Auctions. <http://auctions.amazon.com>
- [2] A. Chavez, and P. Maes, Kasbah: An Agent Marketplace for Buying and Selling Goods. Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM'96). London, UK, April 1996
- [3] Better Business Bureau. <http://www.bbb.org>
- [4] Bizrate. <http://www.bizrate.com>
- [5] C. Castelfranchi, R. Falcone, 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, Minneapolis, May 9-13, 1998
- [6] The Cyprus List. <http://kypros.org/Lists/Cyprus>
- [7] eBay. <http://www.ebay.com>
- [8] A. E., Elo, The Rating of Chessplayers, Past and Present, Arco Publishing, Inc. (New York) 1978.
- [9] J. Donath, Identity and Deception in the Virtual Community, Communities in Cyberspace, Kollock, P. and Smith M. (eds). London: Routledge, 1998.
- [10] FairIsaac Co. <http://www.fairisaac.com>
- [11] L. Foner, Yenta: A Multi-Agent, Referral Based Matchmaking System, First International Conference on Autonomous Agents (Agents '97), Marina del Rey, California, ACM Press, February 1997.
- [12] E. Friedman, and P. Resnick, The Social Cost of Cheap Pseudonyms: Fostering Cooperation on the Internet, Proceedings of the 1998 Telecommunications Policy Research Conference.
- [13] S. Garfinkel, PGP: Pretty Good Privacy, O'Reilly and Associates, 1994.
- [14] M. E. Glickman, Parameter estimation in large dynamic paired comparison experiments, Applied Statistics, 48, 377-394, 1999.
- [15] R. Khare, and A. Rifkin Weaving a Web of Trust, summer 1997 issue of the World Wide Web Journal (Volume 2, Number 3, Pages 77-112).
- [16] P. Kollock, The Production of Trust in Online Markets, Advances in Group Processes (Vol. 16), edited by E. J. Lawler, M. Macy, S. Thyne, and H. A. Walker. Greenwich, CT: JAI Press. 1999.
- [17] Henrik Madsen and Jan Holst: Lecture Notes in Non-linear and Non-stationary Time Series Analysis, IMM, DTU, 1998
- [18] D. Wang, Market Maker: an Agent-Mediated Marketplace Infrastructure, MEng Thesis, Massachusetts Institute of Technology, May 1999

- [19] S. P. Marsh, Formalising Trust as a Computational Concept, PhD Thesis, University of Stirling, April 1994.
- [20] OnSale. <http://www.onsale.com>
- [21] J. M. Jr. Reagle, Trust in a Cryptographic Economy and Digital Security Deposits: Protocols and Policies, Master Thesis, Massachusetts Institute of Technology, May 1996.
- [22] Recreational Software Advisory Council: <http://www.rsac.org/>
- [23] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl GroupLens: An Open Architecture for Collaborative Filtering of Netnews, from Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: Pages 175-186
- [24] P. Resnick and J. Miller, PICS: Internet Access Controls Without Censorship, Communications of the ACM, 1996, vol. 39(10), pp. 87-93.
- [25] Rumelhart, D. E., Hinton, G. E., & Williams, R. J., Learning internal representations by error propagation. In Rumelhart, D. E., & McClelland, J. L. (Eds.), Parallel Distributed Processing, Volume 1: Foundations, pp. 318-362. MIT Press Bradford Books, Cambridge MA, 1986.
- [26] U. Shardanand, and P. Maes, Social Information Filtering: Algorithms for Automating 'Word of Mouth', Proceedings of the CHI-95 Conference, Denver, CO, ACM Press, May 1995.
- [27] Winter Mike, 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, 1999.