Collaborative Reputation Mechanisms for Electronic Marketplaces

Giorgos Zacharia, Alexandros Moukas and Pattie Maes

MIT Media Laboratory

20 Ames Street

Cambridge MA 02139 USA

1-617-253-9833

lysi,moux,pattie@media.mit.edu

ABSTRACT

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. Collaborative reputation mechanisms can provide personalized evaluations of the various ratings assigned to each user to predict their reliabilities. While these reputation mechanisms are developed in the context of electronic commerce, they are applicable in other types of electronic communities such as chatrooms, newsgroups, mailing lists etc.

Keywords

Reputation mechanisms, collaborative filtering, ecommerce, autonomous agents

INTRODUCTION

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 cause 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 as well. 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 [14], eBay [4] and OnSale Exchange [16] 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.

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 [15]. In "Trust in a Cryptographic Economy" [17], 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 form subjective opinions of each other. These opinions may differ greatly between different users, and their variance is most of the time large enough to make the average opinion a rather unreliable prediction.

We have developed methods through which we can automate the social mechanisms of reputation for electronic marketplaces. We have already 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 agent marketplace. Kasbah agents pro-actively 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 first section of this paper outlines the problem we are trying to solve and the problems we faced during the initial implementation of the system in Kasbah. The second section describes related work and the third section outlines specific problems inherent in online marketplaces. Finally the fourth and fifth sections describe the proposed solutions.

Related Work

We can divide the related work into two major categories: non-computational reputation systems like the local Better Business Bureaus [2] and computational ones. 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 the 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 [7]. This central agency also keeps records of the complaints about its users and even publishes warnings against possibly malicious users, much like the local Better Business Bureaus in the U.S. [2]. However useful this 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 which vary across the various users.

Other proposed approaches like Yenta [8], Weaving a web of Trust [12], and the Platform for Internet Content Selection (PICS) (such as the Recreational Software Advisory Council [18]) are more distributed. However, they require the users to rate themselves and to have either a central agency or other trusted users to 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 two 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 and it 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 [20] and GroupLens [19]. 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.

In the context of electronic marketplaces, the most relevant computational methods are the reputation mechanism of OnSale Exchange, eBay and Amazon.com. In OnSale, which allows its users to rate sellers, the overall reputation value of a seller is calculated as the average of 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, the newcomers have no reputation until someone eventually rates them, while in eBay they start with zero feedback points. Bidders in the OnSale Exchange auction system are not rated at all. OnSale tries to ensure the bidders' integrity through a rather psychological measure: bidders are required to register with the system by submitting a credit card. OnSale believes that this requirement helps to ensure that all bids placed are legitimate, which protects the interests of all bidders and sellers. In both systems, the reputation value of a seller is available, with any textual comments that may exist, to the potential bidders. The mechanism of Amazon.com is exactly the same with OnSale Exchange, with the improvement that both the buyers and the sellers are rated after each transaction.

Desiderata for online reputation systems

While the above discussed reputation mechanisms 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 [6][13]. 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. It is economically efficient [9] to give the beginners the option to purchase reputation points for a monetary value. However, in such a model we need to either charge for names in the first place or enforce persistent pseudonymous identities. We therefore decided for the reputation mechanisms described in the following section that a beginner cannot start with an average reputation.

In addition, even the users who receive 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 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 [9] 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.

Furthermore, in systems like Kasbah and OnSale Exchange, 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 in the case of eBay, where a seller can cheat 20% of the time but still maintain a monotonically increasing reputation value.

In multiagent systems, where the agents' goal is the utility maximization of their human owners, the reputation values have to represent the expected probability for the completion of the delegated tasks. Reputation mechanisms have to be able to quantify the subjective [3] expectations of the owners of the agents, based on their past activity on the system.

Finally, we have to consider the memory of our system [15]. We know that the larger the number of ratings used in the evaluation of reputation values the higher 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, we ensure that the predicted reputation values are closer to the current behavior of the individuals rather than their overall performance.

Sporas: A reputation mechanism for loosely connected online communities

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

New users start with a minimum reputation value, and they build up reputation during their activity on the system.

The reputation value of a user never falls below the reputation of a new user.

After each transaction, the reputation values of the involved users are updated according to the feedback provided by the other parties to reflect their trustworthiness in the latest transaction.

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.

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 [5] and the Glicko [11] systems for pairwise ratings.

Each user has one reputation value, which is updated according to the formulae below:

$$R_i = R_{i-1} + \frac{1}{\theta}\Phi(R_i)R_i^{other}\left(W_i - E(W_i)\right)$$

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

$$E(W_i) = {R_{i-1}}/{D}$$

Equation 1 Sporas formulae

where

- θ is a constant integer greater than 1,
- W_i represents the rating given by the user i,
- R_i^{other} is the reputation value of the user giving the rating,

- D is the range of the reputation values,
- σ is the acceleration factor of the dumping function Φ (the smaller the value of σ, the steeper the dumping factor Φ),
- new users start with reputation equal to 0 and can advance up the maximum of 3000,
- the reputation ratings vary from 0.1 for terrible to 1 for perfect.

Equation 1 shows that the 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. In addition, since the reputation of a user in the community is the weighted average of non-negative values, it is guaranteed that no user can have a negative reputation value, thus no user can have a rating value lower than that of a beginner. Also, the weighed average schema guarantees that no user exceeds the maximum reputation value of 3000. However, if a user has a persistent real reputation value, iterations of Equation 1 over a large number of ratings will give an estimate very close to that value [Figure 1]

The expected 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 expected rating value, the reputation value of the user decreases.

The value of θ determines how fast the reputation value of the user changes after each rating. The larger the value of θ, the longer the memory of the system. Thus, just like credit card history [7], 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 initial poor behavior.

Reliability of ratings

Using a similar approach to the Glicko system we have incorporated into the system the measure of the reliability of the users' reputations. The reliability is measured by the standard deviation (SD) of the estimated reputations. A high SD indicates that the user is not that active in the online community and consequently he/she has received few ratings from the other people. Also the reliability of a user's reputation indicates how reliable the opinion of that user for his/her counterpart is. Therefore the SD of each user is updated after each rating, and is affected by both the reputation and the SD of the rater. The update functions are similar to the ones used in the **Glicko** system [11], with the difference that we consider each rating a rating period by itself. The beginners start with a SD of 300 and the minimum SD is set to 30. There is a mathematically rigorous analysis of these algorithms in Paired Comparison Models with Time-Varying Parameters [11].

Histos: A reputation mechanism for highly connected online communities

"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" [12].

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, these reputation values are not final since users of online marketplaces will eventually meet each other in order to commit the agreed transactions and personal biases will be created on the trust relationships between these users. The PGP web of Trust [10] 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 system compared to Sporas. In Weaving a Web of Trust [12], what matters is that there is a connected path of PGP signed webpages between two 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.

We can represent the pairwise ratings in the system as a directed graph [Figure 2], 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.

When user A 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 to A_L that are of length less than or equal to N. As described above we only care about the chronologically θ most recent ratings given to each user. Therefore, if we find more than θ connected paths taking us to user A_L, we are interested only in the most recent θ 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 before A_L in the path. Thus, we create a recursive step with at most θ paths with length at most N-1.

If the length of the path is only 1, it means that the particular user, say C, was rated by A directly. The direct rating given to user C is used as the personalized reputation value for user A. 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 described above. For each user A_k, with *m* connected paths going from A 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_1, and $R_k(n)$ denote the personalized reputation of user $A_k(n)$ from the perspective of user $A_{1.}$

At each level **n** away from user A_1 the users $A_k(n)$ have a reputation value given by:

$$R_k(\mathbf{n}) = D \bullet \sum \left(R_j(\mathbf{n}-1) \bullet W_{jk}(\mathbf{n})\right) / \sum R_j(\mathbf{n}-1)$$
$$\forall j\mathbf{k}, \exists W_{jk}(\mathbf{n})$$
$$m = \deg(\mathbf{A}_k(\mathbf{n})) = |W_{jk}(\mathbf{n})|$$
$$\theta' = \min(\theta, m)$$

Equation 2 Histos formulae

where deg (A_k(n)) is the number of connected paths from A_1 to A_k(n). The users A_k(n) who have been rated directly by user A_1 with a rating $W_{1k}(1)$ have a reputation value equal to:

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

Equation 3 Histos formulae

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

Implementation

The Reputation Server was implemented as a plugin to MarketMaker [Figure 3]. The same architecture was also used for an email experiment. 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 are asked to rate each other based on their performance in the particular deal. A buyer or seller may rate his/her counterpart 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 they 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. 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 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 reputation values would become too slow to be done in real time. In that case we enable 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 with which he can visualize the ratings structure used to evaluate the reputation of the user he is looking at.

Evaluation

To evaluate the reputation mechanisms we applied the algorithms in four simulations. In the first simulation we have 100 users with uniformly distributed real reputations. Each user starts with minimum reputation at 300, maximum SD of 300 and a minimum SD of 30. The users are matched randomly in each period of the simulation and get rated by each other according to their actual performance. 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 4 shows the reputation values for users 0, 1 and 8 over time. Users 0, 1 and 8 reached equilibrium after 15, 21 and 18 ratings respectively. Therefore our system can reach equilibrium very quickly.

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 falling because of his/her unreliable behavior. As we can see in Figure 5 it takes something less that 20 ratings to adjust the reputation of the user to his/her new performance. We plot on the same graph the reputation values that the user would have if he 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 we can see from the graph, the user can take advantage of his/her past good ratings for quite long and keep deceiving people about his/her actual reliability.

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.com's. The actual performance of the two users is .2 and .3 (out of 1) respectively. As we can see in Figure 6, on the simplistic reputation system they actually manage to raise their reputations to .54 and .69 respectively, while with our algorithms, their reputations reflect their actual performance.

Conclusion

Collaborative filtering methods have been around for some years now, but they have focused on content rating and selection. We have developed two collaborative reputation mechanisms that establish reputation ratings for the users themselves. The proposed solutions are able to face all the problems and fulfill all the desiderata described in Section 3. Incorporating reputation mechanisms in online communities may induce social changes in the way users participate in the community.

REFERENCES

[1] 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

[2] Better Business Bureau. http://www.bbb.org

[3] 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), Minneapolis, St. Paul, May 9-13, 1998

[4] eBay. http://www.ebay.com

[5] A. E., Elo, The Rating of Chessplayers, Past and Present, Arco Publishing, Inc. (New York) 1978.

[6] J. Donath, Identity and Deception in the Virtual Community, Communities in Cyberspace, Kollock, P. and Smith M. (eds). London: Routledge, 1998.

[7] Fair Isaak Co. http://www.fairisaac.com

[8] L. Foner, Yenta: A Multi-Agent, Referral Based Matchmaking System, First International Conference on Autonomous Agents (Agents '97), Marina del Rey, California, February 1997.

[9] E. Friedman, and P. Resnick, The Social Cost of Cheap Pseudonyms: Fostering Cooperation on the Internet, Proceedings of the 1998 Telecommunications Policy Research Conference.

[10] S. Garfinkel, PGP: Pretty Good Privacy, O'Reilly and Associates, 1994.

[11] M. E. Glickman, Paired Comparison Models with Time-Varying Parameters, PhD Thesis, Harvard University, May 1993.

[12] 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).

[13] 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.

[14] MarketMaker. http://maker.media.mit.edu

[15] S. P. Marsh, Formalising Trust as a Computational Concept, PhD Thesis, University of Stirling, April 1994.

[16] OnSale Exchange. http://www.onsale.com/exchange.htm

[17] J. M. Jr. Reagle, Trust in a Cryptographic Economy and Digital Security Deposits: Protocols and Policies, Master Thesis, Massachusetts Institute of Technology, May 1996.

[18] Recreational Software Advisory Council: http://www.rsac.org/

[19] 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

[20] 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.

System	Computational	Pair-wise rating	Personalized	Textual comments
GroupLens	Yes	Rating of articles	Yes	
Elo & Glicko	Yes	Result of game		
OnSale	Yes	Buyers rate sellers		
FairIsaac	Yes		Yes	
Local BBB's				Yes
Web of Trust	Yes	Self rating of cost	Yes	
MarketMaker	Yes	Yes	Yes	
Firefly	Yes	Rating of recommendations	Yes	
eBay	Yes	Buyers rate sellers		Yes

Table 1 Comparison of reputation systems

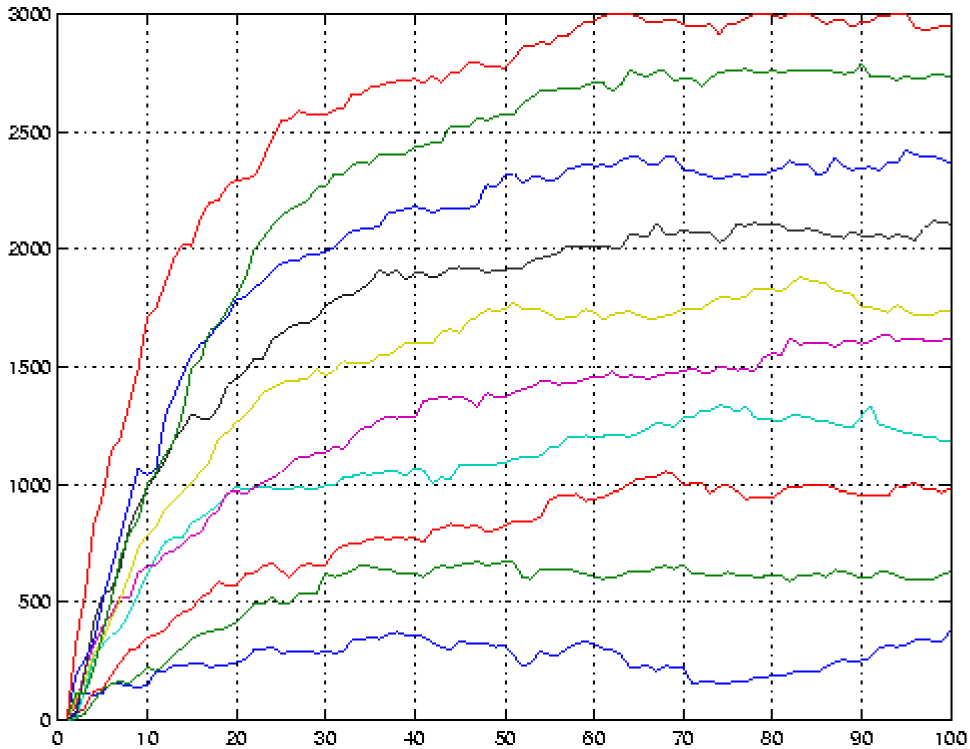

Figure 1 Change of reputation for 10 different users over 100 ratings with $\theta=10$

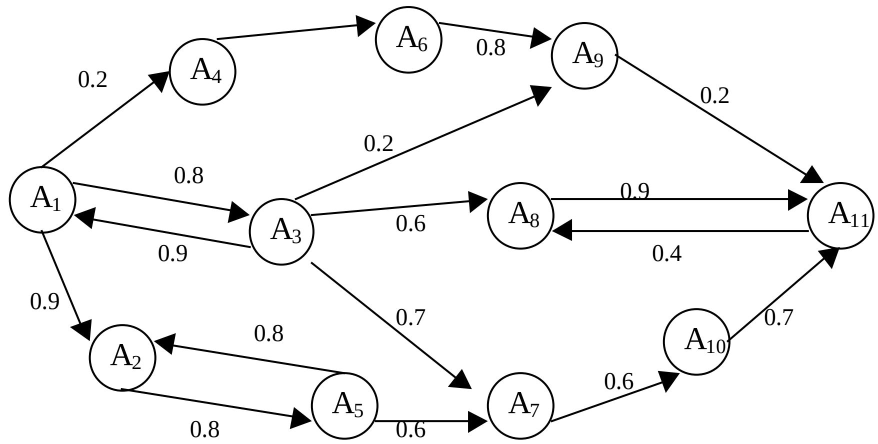

Figure 2 A directed graph representing the rating paths between user A_1 and A_{11}

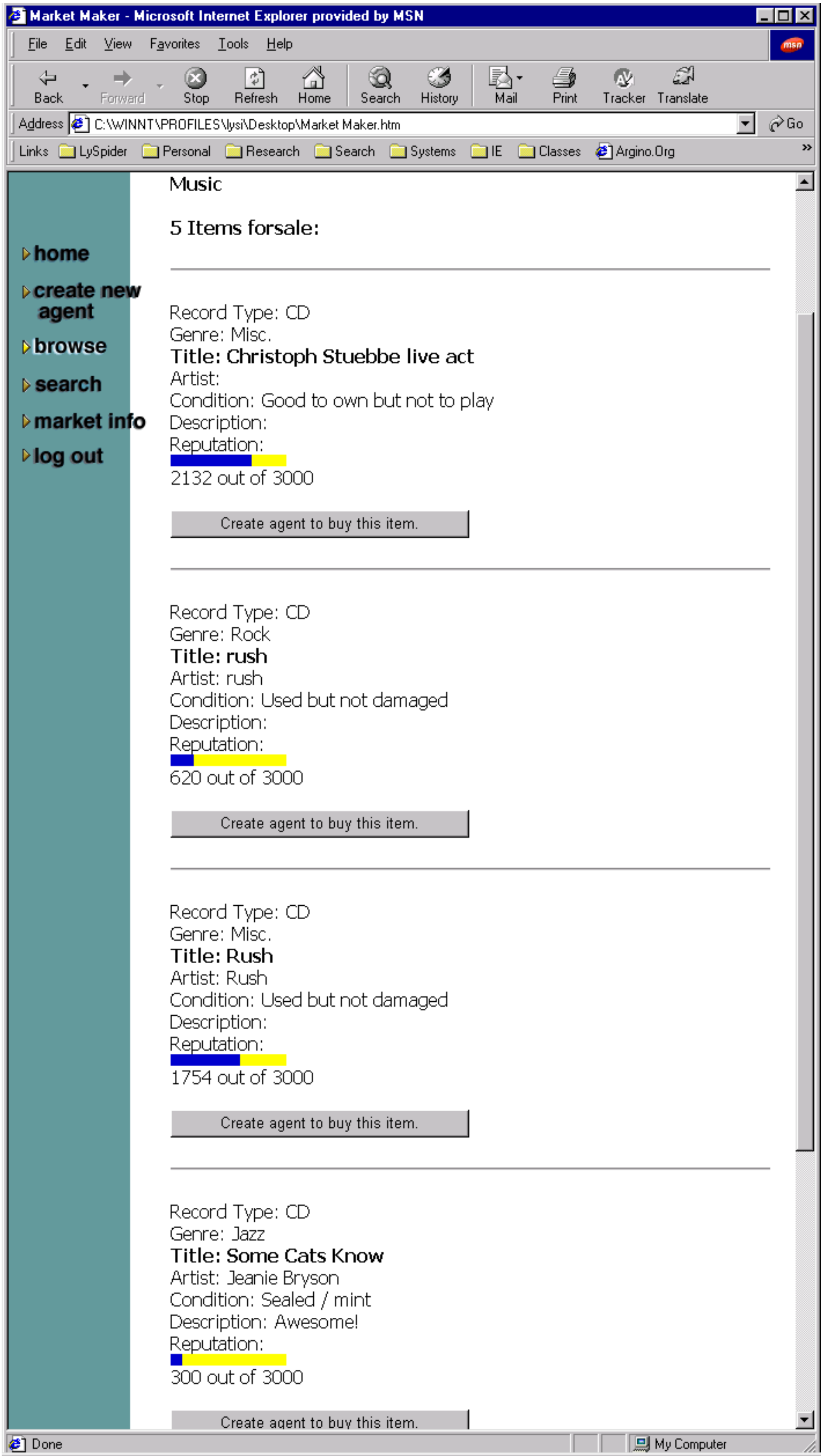

Figure 3 MarketMaker

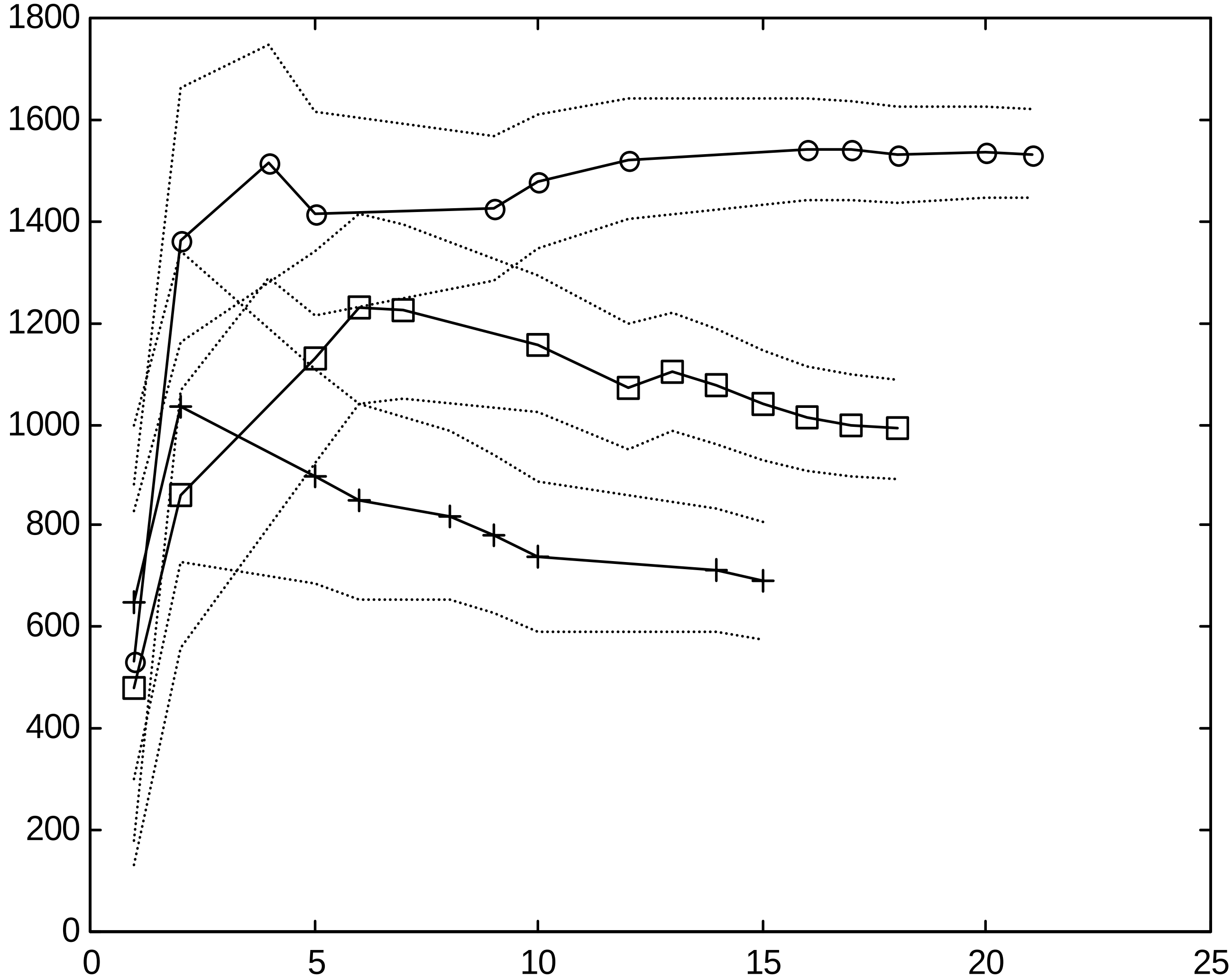

Figure 4 Bootstraping

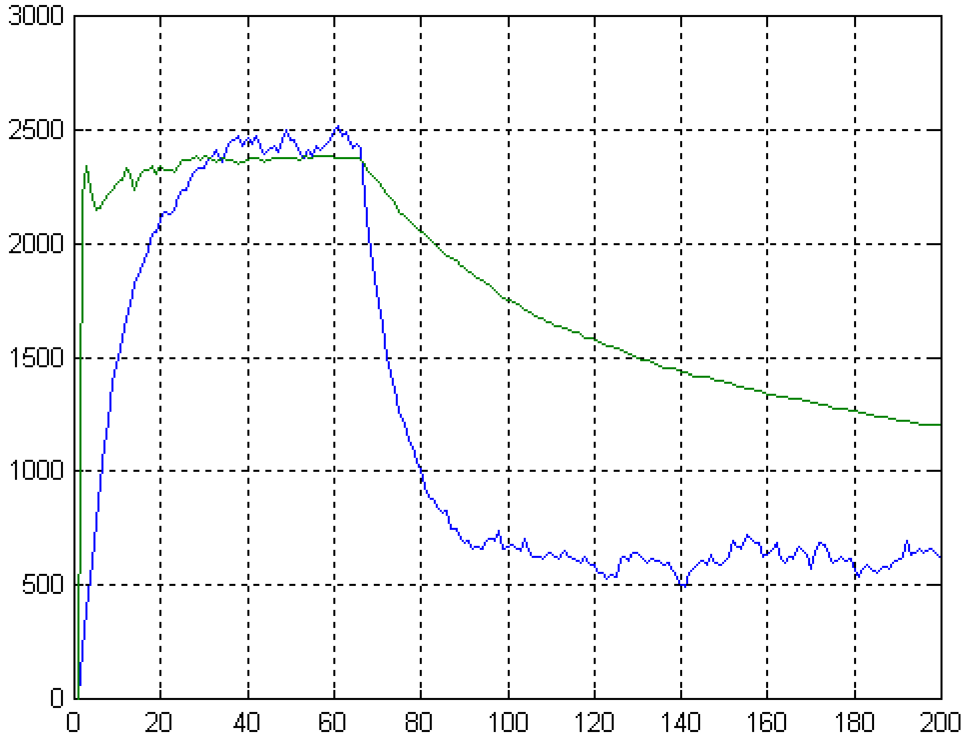

Figure 5 Abuse of prior performance

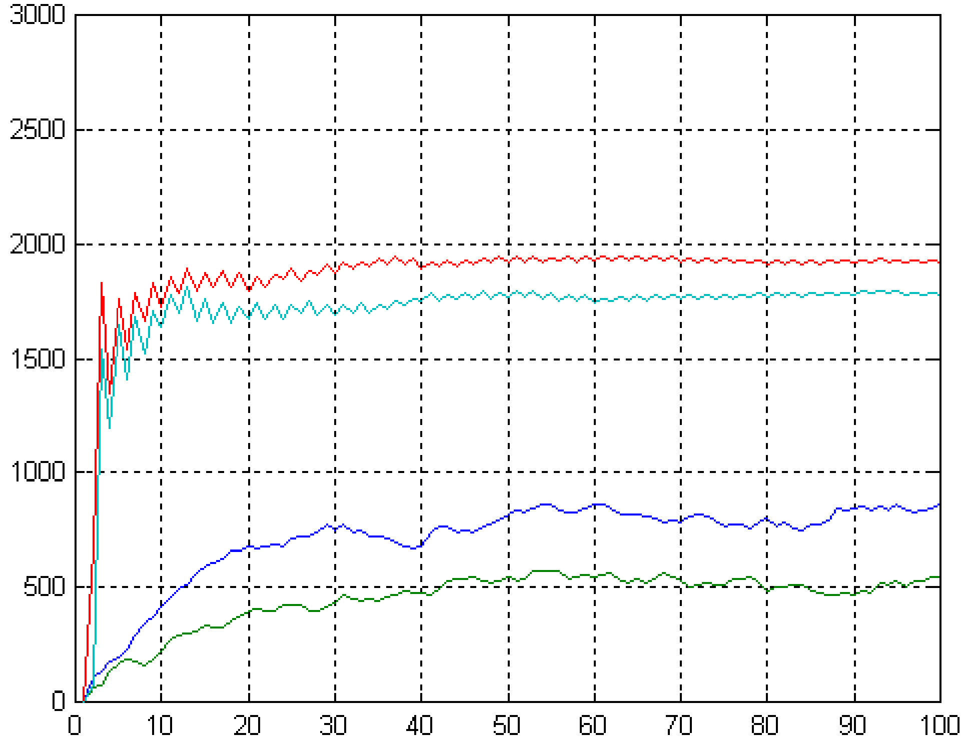

Figure 6 Collusion