Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior

Chrysanthos Dellarocas
Sloan School of Management
Massachusetts Institute of Technology
Room E53-315, Cambridge, MA, 02139, USA
+1 (617) 258-8115
dell@mit.edu

ABSTRACT

Reputation reporting systems have emerged as an important risk management mechanism in online trading communities. However, the predictive value of these systems can be compromised in situations where conspiring buyers intentionally give unfair ratings to sellers or, where sellers discriminate on the quality of service they provide to different buyers. This paper proposes and evaluates a set of mechanisms, which eliminate, or significantly reduce the negative effects of such fraudulent behavior. The proposed mechanisms can be easily integrated into existing online reputation systems in order to safeguard their reliability in the presence of potentially deceitful buyers and sellers.

Keywords

Electronic markets, reputation mechanisms, online fraud, trust.

1. INTRODUCTION

The production of trust is an important requirement for forming and growing open online trading communities. The lack of a common history with potential trading partners as well as the relative ease with which buyers and sellers can change partners from one transaction to the next, gives incentives to both parties to provide inferior service quality or to hold back on their side of the exchange.

Reputation reporting systems have emerged as an important risk management mechanism in such online communities [4]. The goal of reputation systems is to encourage trustworthiness in transactions by using past behavior as a publicly available predictor of likely future behavior.

Most electronic marketplaces on the Internet support some form of reputation mechanism. eBay (www.ebay.com), for example, encourages both parties of each transaction to rate one another with either a positive (+1) or a negative (-1) rating. eBay makes the cumulative ratings of its members publicly known to every registered user.

Despite their widespread adoption and undeniable importance, very little work has been published so far on the reliability of various reputation mechanisms and, even more importantly, on ways in which these mechanisms can be compromised. Some notable exceptions include the work of Friedman and Resnick [2], which discusses risks related to the ease with which online community participants can change their identity. They conclude that the assignment of the lowest possible reputation value to newcomers is an effective mechanism for discouraging participants to misbehave and subsequently change their identity. Zacharia et. al. [8] propose a specific design for a reputation mechanism and touch upon some of the reliability desiderata, but do not attempt an exhaustive evaluation of their proposed design against such risks.

This paper aims to contribute in the construction of more robust online reputation systems by identifying, and proposing mechanisms for addressing, two important classes of reputation system fraud: scenarios where buyers intentionally provide unfairly high or unfairly low ratings for sellers, as well as scenarios where sellers attempt to "hide" behind their cumulative reputation in order to discriminate on the quality of service they provide to different buyers.

2. UNFAIR BEHAVIOR IN ONLINE TRADING COMMUNITIES

In most online trading communities participants can be distinguished into buyers and sellers. In this paper we further assume that only buyers can rate sellers. In the future, we will investigate additional issues that arise in systems, which allow bi-directional ratings. In a typical transaction, a buyer *b* contracts a seller *s* for the provision of a service. Upon conclusion of the transaction, *b* provides a numerical rating $R_b(s)$, reflecting the quality of service offered by *s* as perceived by *b* (we assume here that higher ratings reflect higher quality). In most situations, a rating would be a vector, reflecting a buyer's assessment of various different aspects of quality (for example, Zagat's

EC'00, October 17-20, 2000, Minneapolis, Minnesota.

restaurant guides rate restaurants according to three different quality indicators: Food, Décor and Service).

A reputation system's task is to use past ratings in order to calculate reliable *reputation estimates* $\hat{R}(s)$ for sellers. In this context, a reputation estimate acts as a predictor of a seller's future service quality.

The calculation of reliable reputation estimates is complicated by the fact that most attributes of quality can only be measured subjectively. For example, the food or décor quality of a restaurant is highly dependent on the tastes of individual raters. Since tastes can vary substantially from individual to individual, this can result in a corresponding dispersion of ratings for service, which, from the seller's perspective, is consistent over time. Taste differences introduce the need to *personalize* reputation ratings. In other words, the reputation system should provide a different estimate of a seller's expected service quality for each buyer, trying to compensate for each individual buyer's taste profile. A buyer's taste profile can be inferred from her past ratings using a family of techniques collectively known as *collaborative filtering* [3, 6]. Collaborative filtering techniques calculate a personalized reputation estimate $\hat{R}_b(s)$, as a weighted average of past ratings $R_u(s)$ given to s by other buyers. Weights are proportional to the *similarity* between buyers b and u. Buyer similarity is usually calculated as a function of the correlation between the past ratings of b and u for commonly rated sellers, although several alternative approaches have also been proposed [1].

In settings where legitimate taste differences may exist among buyers, there exist four scenarios where buyers and/or sellers can intentionally try to "rig the system", resulting in biased reputation estimates, which do not reflect the true expected quality of service provided by a seller:

a. Unfair ratings by buyers

- *Unfairly high ratings ("ballot stuffing"):* A seller colludes with a group of buyers in order to be given unfairly high ratings by them. This will have the effect of inflating a seller's reputation, therefore allowing that seller to receive more orders from buyers and at a higher price than she deserves.
- *Unfairly low ratings ("bad-mouthing"):* Sellers can collude with buyers in order to "bad-mouth" other sellers that they want to drive out of the market. In such a situation, the conspiring buyers provide unfairly negative ratings to the targeted sellers, thus lowering their reputation.

b. Discriminatory seller behavior

- *Negative discrimination:* Sellers provide good service to everyone except a few specific buyers that they "don't like". If the number of buyers being discriminated upon is relatively small, the cumulative reputation of sellers will be good and an externality will be created against the victimized buyers.
- *Positive discrimination:* Sellers provide exceptionally good service to a few select individuals and average service to the rest. The effect of this is equivalent to ballot stuffing. That is, if the favored group is sufficiently large, their favorable ratings will inflate the reputation of discriminating sellers and will create an externality against the rest of the buyers.

The observable effect of all four above scenarios is that there will be a dispersion of ratings for a given seller. In settings where legitimate taste differences may exist among buyers, it will be very difficult, or impossible to distinguish ratings dispersion due to taste differences from that which is due to unfair ratings or discriminatory behavior. This creates a *moral hazard*, which requires additional mechanisms in order to be either avoided, or detected and resolved.

3. THE PROPOSED MECHANISMS

This section proposes a set of "exception handling" mechanisms, which eliminate, or significantly reduce the effects of unfair ratings and discriminatory seller behavior on the predictive accuracy of reputation estimates.

3.1 Using controlled anonymity to avoid unfairly low ratings and negative discrimination

Bad-mouthing and negative discrimination are based on the ability to pick a few specific "victims" and give them unfairly poor ratings or provide them with poor service respectively. Both can be avoided if the marketplace conceals the identities of the buyers and sellers from each other.

In such a "controlled anonymity" scheme, the marketplace knows the identity of all market participants. In addition, it keeps track of all transactions and ratings. The marketplace publishes the estimated reputation of buyers and sellers but keeps their identities concealed from each other (or assigns them pseudonyms which change from one transaction to the next). In that way, buyers and sellers make their decisions solely based on the offered terms of trade as well as the published reputations. Because they can no longer identify their "victims", bad-mouthing and negative discrimination can be avoided.

It is interesting to observe that, while, in most cases, the anonymity of online communities has been viewed as a source of additional risks [2, 4], here we have an example of a situation where anonymity can be used to eliminate some transaction risks.

Concealing the identities of buyers and sellers is not possible in all domains. For example, concealing the identity of sellers is not possible in restaurant and hotel ratings (although concealing the identity of buyers is). In other domains, it may require the creative intervention of the marketplace. For example, in a marketplace of electronic component distributors, it may require the marketplace to act as an intermediary shipping hub that will help erase information about the seller's address. Nevertheless there are several domains where this approach can be applied effectively with relatively small cost (for example, eBay and other related C2C auction sites).

Generally speaking, concealing the identities of buyers is usually easier than concealing the identities of sellers. This means that negative discrimination is easier to avoid than "bad-mouthing". Furthermore, concealing the identities of sellers before a service is performed is usually easier than afterwards. In domains with this property, controlled anonymity can be used at the seller selection stage in order to, at least, protect sellers from being intentionally picked for subsequent bad-mouthing. For example, in the above-mentioned marketplace of electronic component distributors, one could conceal the identities of sellers until after the closing of a deal. Assuming that the number of distributors for a given component type is relatively large, this strategy would make it difficult, or impossible, for malevolent buyers to intentionally pick specific distributors for subsequent bad-mouthing.

3.2 Using cluster filtering to reduce the effect of unfairly high ratings and positive discrimination

Even when identities of buyers and sellers are concealed, buyers and sellers who have an incentive to signal their identities to each other can always find clever ways to do so. For example, sellers involved in a "ballot stuffing" scheme can use a particular pattern in the amounts that they bid (e.g. amounts ending in .33) in order to signal their presence to their conspirators. Therefore, while controlled anonymity can avoid bad-mouthing and negative discrimination, it cannot avoid "ballot stuffing" and positive discrimination.

Nevertheless, if we don't have to worry about unfair negative ratings, it becomes much easier to greatly reduce the effects of unfair positive ratings. We propose the following algorithm for doing so:

To calculate an unbiased personalized reputation estimate $\hat{R}_b(s)$, we first use collaborative filtering techniques to identify the *nearest neighbor set* N of *b*. N includes buyers who have previously rated *s* and who are the nearest neighbors of *b*, based on their similarity with *b* on commonly rated sellers. Sometimes, this step will filter out all the unfair buyers. Suppose, however, that the colluders have taken collaborative filtering into account and have cleverly picked buyers whose tastes are similar to those of *b* in everything else except their ratings of *s*. In that case, the resulting set N will include some fair raters and some unfair raters. The ratings will, therefore, form two clusters: the lower cluster N_l, which consists of fair ratings and the upper cluster N_u, which consists of unfair ratings. To eliminate the unfair ratings we apply a divisive clustering algorithm, such as the one proposed by Macnaughton-Smith et. al. ([5]; also see appendix), in order to separate the set of raters N into two clusters N_l and N_u based on some function[1] of their ratings of *s*. Finally, we calculate the final reputation estimate $\hat{R}_b(s)$ as a function of the ratings provided by members of cluster N_l only.

4. ANALYSIS AND ENHANCEMENTS OF CLUSTER FILTERING

As with all filtering approaches, it is important to ascertain that the cluster filtering mechanism introduced in the previous section does a good job of removing unwanted "noise" (unfair ratings) while not significantly distorting the useful "signal content" (fair ratings) in the final reputation estimate. The following paragraphs evaluate the efficiency of cluster filtering in a variety of important settings.

4.1 Efficiency when ratings are steady over time

The simplest scenario to analyze is one where quality ratings, both fair and unfair, given to a seller by a group of buyers with similar tastes, are steady over time. To simplify our analysis, we assume that ratings are scalar quantities, which range between 0 and 100. The analysis can be easily generalized to vector ratings of arbitrary ranges as well. Suppose that fair ratings follow a normal distribution $N(\mu, \sigma)$[2] and unfair ratings a distribution $N(\mu', \sigma')$[3]. In such a scenario, the *fair* reputation estimate $\hat{R}_b(s)_{fair} \approx \mu$.

We assume that due to the use of controlled anonymity, unfair *negative* ratings are not an issue in this community. Therefore, the goal of unfair raters is to strategically choose μ' and σ' in order to *maximize* the reputation estimate $\hat{R}_b(s)$ calculated by the reputation system. Assume that the initial collaborative filtering step constructs a nearest neighbor set N, which includes $(1-\delta)\cdot 100\%$ fair raters and $\delta\cdot 100\%$ unfair raters. Finally, assume that the reputation estimate $\hat{R}_b(s)$ is calculated as the average of the *most recent rating* given to *s* by each qualifying rater in N. In the absence of cluster filtering, the *biased* reputation estimate calculated by the above algorithm will approximate:

$$\hat{R}_b(s)_{nofilter} \approx (1-\delta)\mu + \delta\mu'$$

The strategy which maximizes the above reputation estimate is one where $\mu'=100$ and $\sigma'=0$, i.e. where all unfair buyers give the maximum possible rating to the seller. We define the *reputation bias B* to be the maximum difference (expressed in "quality units") between the biased and fair reputation estimates for a given pair of μ and δ and over all possible choices of μ' and σ'. In the above case, the reputation bias is given by:

$$B(\mu, \delta)_{nofilter} = \hat{R}_b(s)_{nofilter} - \hat{R}_b(s)_{fair} \approx \delta(100 - \mu)$$

The above formula can result in very significant inflation of a seller's reputation, especially for small μ and large δ (see also Figure 2).

Our goal is to determine to what extent cluster filtering is capable of reducing reputation bias relative to the above baseline case. To determine this, we have applied the cluster filtering algorithm of Macnaughton-Smith et. al. (see appendix) to a large number of test cases and have calculated the maximum positive reputation bias that unfair buyers can achieve in each case.

Figure 1 summarizes the parameters of our experiments. For each pair of values μ, δ considered, we have tested a large range of unfair ratings strategies $N(\mu', \sigma')$ and calculated the maximum reputation bias achieved by any of them, with and without cluster filtering. The results are plotted in Figures 2 and 3.

[1] That function will most often be either the average of all their ratings of s or the value of their most recent rating of s.

[2] More precisely ratings are given by: max(0, min(100, *z*)) where *z* follows a normal distribution.

[3] Due to the lack of space we restrict our discussion to normally distributed unfair ratings. We have tested other distributions as well with very similar results.

Cardinality of nearest neighbor set N	100
Fraction δ of unfair ratings in N	Five cases tested: $\delta = 0.75, 0.50, 0.25, 0.10, 0$
Fair ratings distributions tested for each δ	$N(\mu, 5)$ for $\mu = 10, 20, \ldots, 90$
Unfair ratings distributions tested for each pair (δ, μ)	$N(\mu', \sigma')$ for $\mu' = 0, 5, 10, \ldots, 90, 95, 100$ and $\sigma' = 0, 1, 2, \ldots, 99, 100$

Figure 1. Steady-state test case parameters.

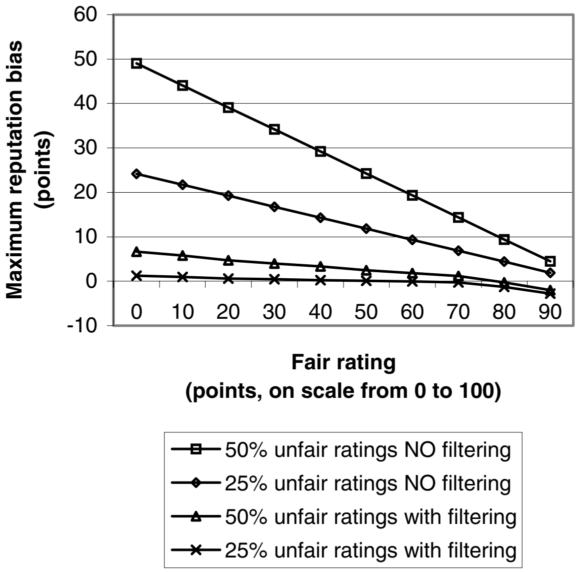

Figure 2. Effectiveness of cluster filtering in reducing reputation bias due to unfair ratings.

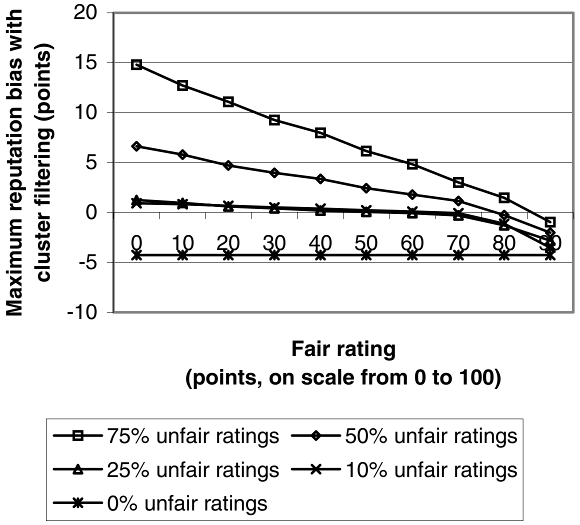

Figure 3. Reputation bias as a function of the percentage of unfair ratings.

The effectiveness of cluster filtering in significantly reducing reputation bias is evident from those results. The most important conclusions of this analysis can be summarized as follows:

- The optimal "ballot stuffing" strategy is to use a distribution of unfair ratings where μ' is very large (at or near the top of the ratings scale) while σ' is between $0.2\mu'$ and $0.5\mu'$ (actual values depend on the fair rating mean μ; exact values are available from the author). Intuitively, this means that unfair buyers will give the highest possible ratings to the seller, but with a sufficient dispersion in order to "confuse" the clustering algorithm into incorporating at least some of them into the lower cluster N_l.
- The maximum reputation bias rises with the fraction δ of unfair raters in the nearest neighbor set and is inversely proportional to the fair reputation estimate μ. In other words, the worse the actual quality of a seller, the biggest the impact of unfair ratings in inflating her reputation
- For percentages of unfair raters of 25% and below, cluster filtering practically eliminates reputation bias due to "ballot stuffing". For percentages up to 50% it limits reputation bias to 6 points or less (on a scale from 0 to 100).
- In cases where there are no unfair ratings, cluster filtering results in a negative reputation bias, roughly equal to the standard deviation of fair ratings σ. If collaborative filtering been successful in bringing together buyers with "similar tastes" in the nearest neighbor set N, then one expects that fair ratings will be relatively consistent and therefore the negative bias σ will be small. This issue is discussed more fully in Section 4.4.

4.2 Additional considerations when ratings vary over time

This section expands our analysis by discussing some additional considerations, which arise in environments where seller quality, and therefore ratings, may vary over time. We identify some additional "ballot stuffing" strategies that can be very disruptive in such environments. Section 4.3 then proposes an enhancement to the original cluster filtering algorithm introduced in Section 3.2, which practically eliminates the negative effects of these new strategies.

In real-life trading communities, sellers may vary their service quality over time, improving it, deteriorating it, or even oscillating between phases of improvement and phases of deterioration. In his seminal analysis of the economic effects of reputation [7], Shapiro proved that, in such environments, the most economically efficient way to estimate a seller's reputation (i.e. the way that induces the seller to produce at the highest quality level) is as a time discounted average of recent ratings. Shapiro went even further to prove that efficiency is higher (1) the higher the weight placed on recent quality ratings and (2) the higher the discount factor of older ratings.

In this paper we are basing our analysis on an approach with approximates Shapiro's desiderata, but is simpler to implement and analyze. The principal idea is to calculate time-varying personalized reputation estimates $\hat{R}_b(s,t)$ as averages of ratings submitted within the most recent time window W=[t-ε, t] only. This is equivalent to using a time discounted average calculation where weights are equal to 1 for ratings submitted within W and 0 otherwise. More specifically, in order to calculate time-varying personalized reputation estimates $\hat{R}_b(s,t)$, we first use collaborative filtering in order to construct an initial nearest neighbor set $N_{initial}$. Following that we construct the *active* nearest neighbor set N_{active}, consisting only of those buyers $u \in N_{initial}$ who have submitted at least one rating for s within W. Finally, we base the calculation of $\hat{R}_b(s,t)$ on ratings $R_u(s,t)$ where $u \in N_{active}$ and $t \in W$.

The conclusions of Section 4.1 make it clear that the maximum reputation bias due to unfair ratings is proportional to the ratio δ of unfair ratings, which "make it" into the active nearest neighbor set N_{active}. Therefore, an obvious strategy for unfair buyers is to try to increase δ by "flooding" the system with unfair ratings. Zacharia et. al. [8] touch upon this issue and propose keeping only the *last* rating given by a given buyer to a given seller as a solution. In environments where reputation estimates use all available ratings, this simple strategy ensures that eventually δ can never be more than the actual fraction of unfair raters in the community, usually a very small fraction. However, the strategy breaks down in environments where reputation estimates are based on ratings submitted within a relatively short time window (or where older ratings are heavily discounted). The following paragraph explains why.

Let us assume that the initial nearest neighbors set $N_{initial}$ contains m fair raters and n unfair raters. In most cases $n << m$. Assume further that the average interarrival time of fair ratings for a given seller is λ and that personalized reputation estimates $\hat{R}_b(s,t)$ are based only on ratings for s submitted by raters $u \in N_{initial}$ within the time window $W = [t - k\lambda, t]$. Based on the above assumptions, the average number of fair ratings submitted within W would be equal to k. To ensure accurate reputation estimates, the width of the time window W should be relatively small; therefore k should generally be a small number (say, between 3 and 10)[4]. For $k << m$ we can assume that every rating submitted within W is from a distinct fair rater. Assume now that unfair raters flood the system with ratings at a frequency much higher than the frequency of fair ratings. If the unfair ratings frequency is high enough, every one of the n unfair raters will have submitted at least one rating within the time window W. As suggested by Zacharia et. al., we keep only the last rating sent by each rater. Even using that rule, however, the above scenario would result in an active nearest neighbor set of raters where the fraction of unfair raters is given by $\delta = n/(n+k)$. This expression results in $\delta \geq 0.5$ for $n \geq k$, independent of how small n is relative to m. For example, if n=10 and k=5, then $\delta = 10/(10+5) = 0.67$. We therefore see that, for relatively small time windows, even a small (e.g. 5-10) number of colluding buyers can successfully use unfair ratings flooding to significantly increase δ and, therefore, the reputation bias.

Figure 3 shows that for $\delta > 0.5$, even when cluster filtering is used, significant reputation biases are possible. This fact, together with the above arguments, has prompted us to consider the possibility of unfair ratings flooding as a serious issue. Fortunately, it is relatively straightforward to extend the cluster filtering algorithm proposed in Section 3.2 in order to practically eliminate the effects of "flooding". The next section describes and evaluates the proposed enhancements.

4.3 Using cluster filtering in the frequency domain to eliminate unfair ratings flooding

The cluster filtering approach introduced in Section 3.2 attempts to separate fair and unfair ratings by clustering the members of the nearest neighbor set according to the *values* of their ratings. In order to counter attempts to inflate a seller's reputation using unfair ratings flooding, in this section we are proposing to use cluster filtering based on the *frequency* of ratings as well. The extended algorithm, which combines cluster filtering in the value and frequency domains, is described in Figure 4.

In order to experimentally evaluate the effectiveness of performing cluster filtering in the frequency domain, we have simulated an electronic marketplace with m=90 fair and n=10 unfair buyer agents. Fair ratings follow an exponential distribution with mean interarrival time λ. Reputation estimates are based on a time window whose width is equal to 5λ. This means that, k, the average number of fair ratings per time window, is equal to 5.

The goal of our experiment is to observe the maximum reputation bias that can be achieved by unfair raters through the use of flooding and to evaluate the effectiveness of frequency-based cluster filtering in reducing that bias. From the analysis of Section 4.1 we know that the maximum reputation bias is inversely proportional to the fair reputation estimate μ (see Figures 2 and 3). Therefore, the maximum observable effects occur when μ=0. For μ=0, the most effective "ballot stuffing" strategy was experimentally found to be one where unfair ratings follow a normal distribution N(μ=100, σ=48).

We have completed several simulation runs with the above parameters, varying the frequency of unfair ratings relative to the frequency of fair ratings. In each run we have calculated the biased and unbiased reputation estimate on 100 successive time windows with and without frequency-based cluster filtering (value-based cluster filtering was used in all test runs).

The results are plotted in Figure 5. From that figure we can observe that, even in the presence of value-based cluster filtering, 10 unfair raters can effectively flood the system with unfairly high ratings in order to considerably increase the average reputation bias (and the maximum reputation bias even more due to the relatively small number of fair ratings per time window). On the other hand, the addition of frequency-based cluster filtering was successful in almost completely neutralizing the effects of unfair ratings flooding.

[4] Making the width of the time window small is approximately equivalent to using a higher discount factor for older ratings, which, according to Shapiro, results in more efficient reputation mechanisms.

```
Initialization at time 0:
1. Pick the desired average number of fair ratings per time window $k$
2. For each seller $s$, calculate an initial estimate of $\lambda_s$, the average interarrival time of fair ratings for seller $s$ by any buyer.
To calculate an unbiased personalized reputation estimate $\hat{R}_b(s,t)$
1. Using some collaborative filtering mechanism, construct the initial nearest neighbor set $N_{initial}$ of $b$
2. Construct the active nearest neighbor set of raters $N_{active} \subseteq N_{initial}$, consisting of raters $u \in N_{initial}$ who have submitted at least one rating for $s$ within the time window $W = [t-k\cdot\lambda_s, t]$
3. For each $u \in N_{active}$, calculate the average frequency of ratings $f_u(s)$. The average frequency is the total number of ratings submitted by $u$ for $s$ within a sufficiently large time window, divided by the width of that time window. We recommend that the time window used to calculate average ratings frequency be at least as large as $10n\lambda_s$, where $n$ is the total number of buyers in the community, in order to ensure that there will be at least a few ratings per rater within that window.
4. For each rater in $u \in N_{active}$ discard all ratings $R_u(s,t)$, $t \in W$ except the most recent one
5. Apply the clustering algorithm of Macnoughton-Smith et. al. [5] to $N_{active}$ basing the clustering on the most recent rating values of members.
6. For each resulting cluster $N_l$, $N_u$ re-apply the clustering algorithm, this time basing the clustering on the average ratings frequency of members.
7. Calculate the average rating value in each of the four clusters produced by Step 6. Keep the cluster $N_{lowest}$ with the lowest average rating value and discard the other three.
8. Return the average rating value of $N_{lowest}$ as $\hat{R}_b(s,t)$
9. Update the estimate of the fair ratings interarrival time for seller $s$ as follows:
       $\lambda_{s,new} = \alpha\cdot\lambda_{s,old} + (1-\alpha)\cdot 1 / (n\cdot\min(f_u(s) \mid u \in N_{lowest}))$
   The best $\alpha \in [0,1]$ is derived by experiment.
```

Figure 4. Enhanced Cluster Filtering Algorithm.

As an epilogue to this section, we would like to graphically demonstrate the effects of cluster filtering in a setting where seller quality oscillates over time between 0 and 40 with a period equal to 15 time windows, reputation estimates are based on a time window intended to contain k=5 fair ratings and 10 unfair buyers attempt to inflate the seller's reputation by flooding the system with ratings following a normal distribution N(μ=100, σ=48) at 20 times the frequency of fair ratings. Figure 6(a) shows the resulting reputation estimates when no cluster filtering is used. It is fairly obvious that the reputation estimates have little relation with the actual seller quality in this case. The unfair buyers have been very effective in completely destroying the reputation system's reliability. Figure 6(b) shows the reputation estimates when value-based-only cluster filtering is used. The reputation bias has been significantly reduced. Finally, Figure 6(c) shows the additional improvements achieved by applying cluster filtering in both the value and the frequency domains. More than any amount of explanation, we believe that these graphs clearly demonstrate, both the potentially disruptive effect of unfair ratings flooding, as well as the effectiveness of cluster filtering in immunizing a reputation reporting system against this risk.

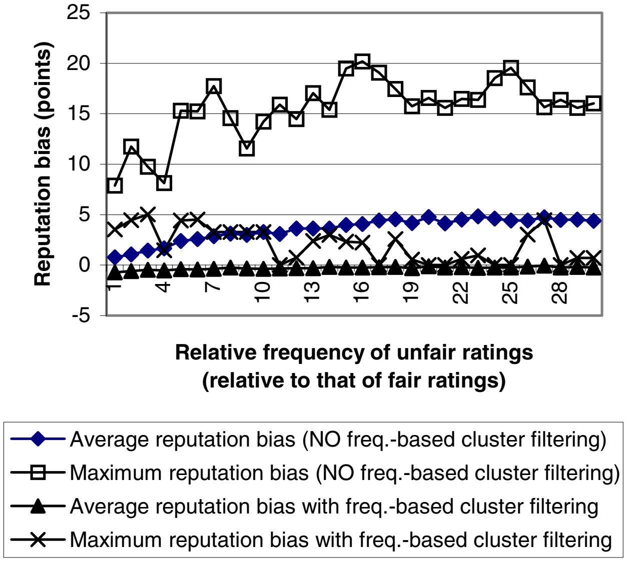

Figure 5. Effectiveness of frequency-based cluster filtering in neutralizing the effects of unfair ratings flooding.

4.4 Performance of cluster filtering in the absence of unfair raters

The previous sections have demonstrated that cluster filtering is a very effective technique for significantly reducing reputation bias in the presence of unfair ratings. Before we can declare victory, however, it is important to also investigate the effects of this technique in environments where no unfair ratings exist (one would hope, most environments!).

In the absence of unfair ratings, cluster filtering will assign some of the highest fair ratings given to a seller into the upper cluster N_u, thus eliminating them from the calculation of the reputation estimate. This is expected to result in a small negative bias. When ratings are steady over time, we found this negative bias to be roughly equal to the standard deviation σ of fair ratings (see Section 4.1). If collaborative filtering is effective in including fair buyers of similar tastes in the nearest neighbor set N then σ should be relatively small and therefore the negative bias caused by cluster filtering should be acceptable.

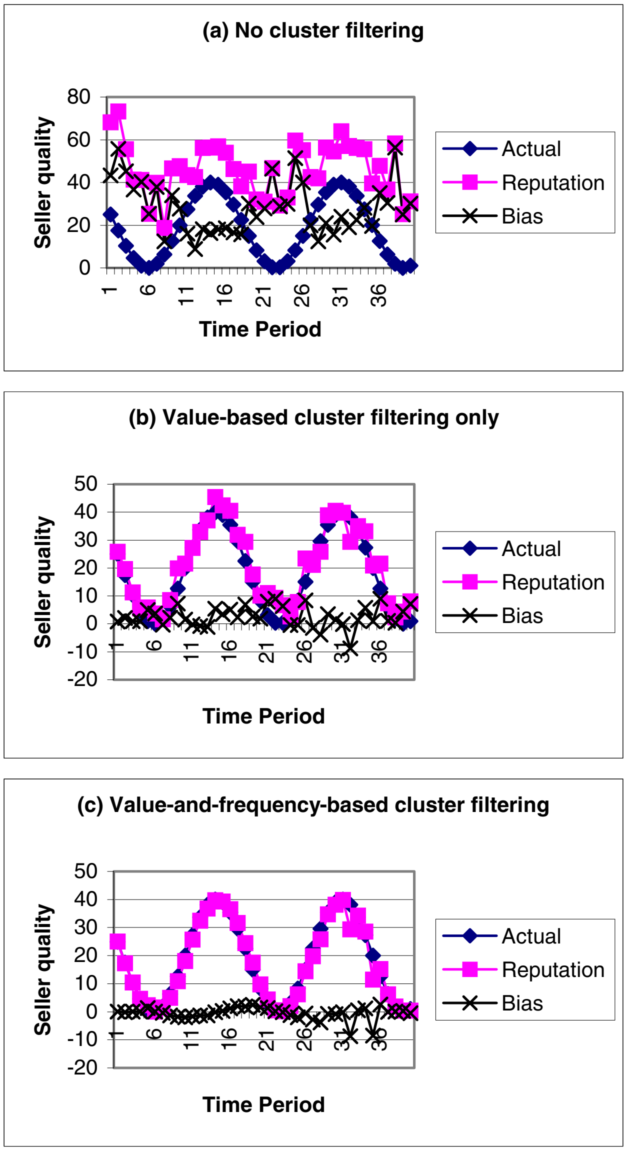

Figure 6. Effectiveness of cluster filtering in immunizing a reputation reporting system against unfair ratings.

In environments where ratings vary over time it is expected that reputation estimates calculated as a function of past ratings will result in a bias due to estimation errors, whether cluster filtering is used or not. Our goal in this section is to compare the bias in the two cases.

We have tested a scenario where a seller's quality oscillates over time between 0 and 40 with a period equal to 15 time windows. As before, reputation estimates are based on the 5 most recent fair ratings. We have calculated the reputation estimates with and without cluster filtering and plotted the results in Figure 7. Figure 7(b) compares the resulting reputation biases. By observing that figure it becomes clear that cluster filtering shifts the reputation bias down by an amount roughly equal to the maximum reputation bias without cluster filtering. More specifically, in periods where seller reputation is *increasing* over time, cluster filtering roughly doubles the *negative* reputation bias (relative to the case where no cluster filtering is used). On the other hand, on periods where seller reputation is *decreasing*, then the negative reputation bias introduced by cluster filtering cancels the positive reputation estimation error and results in pretty accurate predictions.

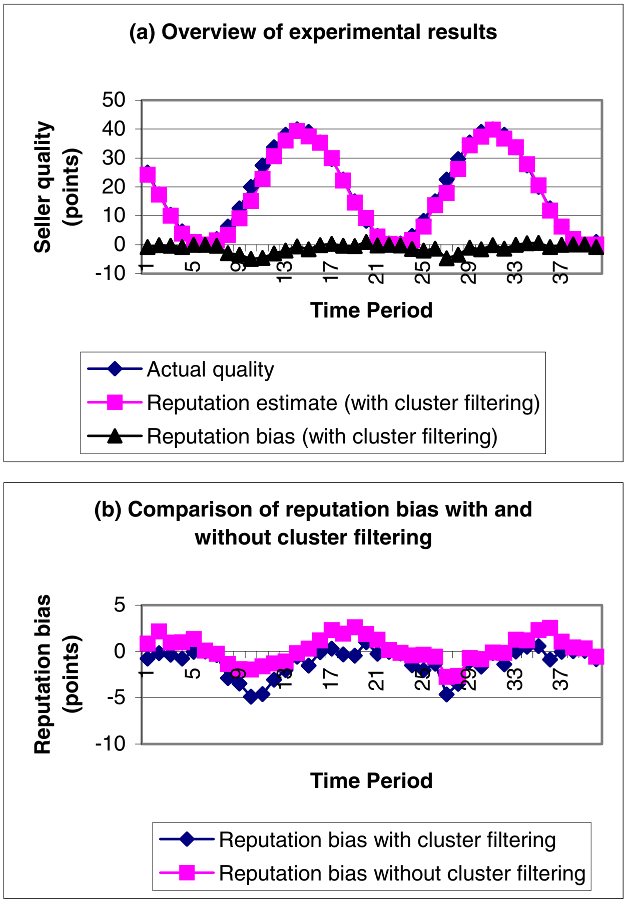

Figure 7. Reputation bias due to cluster filtering in the absence of unfair ratings.

To conclude, in common with most filtering approaches, cluster filtering does have a cost. This cost is manifested in the form of a small negative reputation bias in the absence of unfair ratings. More specifically:

- In periods where seller quality is steady, the *negative* reputation bias due to cluster filtering is roughly equal to the variance of ratings in the nearest neighbor set of a buyer.
- In periods where seller quality is increasing, the *negative* reputation bias due to cluster filtering is equal to about two times the corresponding estimation bias without cluster filtering
- In periods where seller quality is decreasing the reputation bias due to cluster filtering is negligible.

In environments where collaborative filtering has been relatively successful in constructing a coherent nearest neighbor set and where seller quality is not very volatile with time, the above reputation bias will be relatively small (for example, in Figures 3 and 7 it was less than –5 points on a scale from 0-100). We believe that the impressive effectiveness of cluster filtering in practically eliminating the effects of unfair ratings more than compensates for this small negative bias in the absence of unfair ratings.

5. CONCLUSIONS

This paper has made a number of contributions in the construction of more reliable online reputation reporting systems. First, it has identified several scenarios ("ballot stuffing", "bad-mouthing", positive seller discrimination, negative seller discrimination, unfair ratings "flooding") in which buyers and sellers can attempt to "rig" an online reputation reporting system to their advantage, resulting in biased reputation estimates, which do not accurately reflect the expected quality of service of a given seller. Second, it has proposed two mechanisms (controlled anonymity and cluster filtering) for coping with the above scenarios. Third, it has performed an analysis of the effectiveness of cluster filtering in a variety of settings (steady seller quality, time-varying seller quality, with and without unfair ratings).

The results presented in this paper indicate that the combination of controlled anonymity and cluster filtering is a powerful technique for "immunizing" online reputation reporting systems in the presence of unfair ratings and discriminating seller behavior. Given the increasing importance of online reputation mechanisms in building trust and managing risks in online trading communities, further research is needed in order to discover additional ways in which such systems may be compromised, as well as to propose mechanisms for coping with them.

APPENDIX: Dividing a set into two clusters using the iterative method of Macnaughton-Smith et. al. [5]

The following algorithm divides a set *N* into two clusters *A* and *B*:

Step 1. Initially, we set *A*=*N* and *B*=∅. In a first stage, we have to move one object from *A* to *B* (it is assumed that *A* contains more than one object). For each object $i \in A$, we compute the average dissimilarity to all other objects of *A*:

$$D(i, A-\{i\}) = \frac{1}{|A|-1} \sum_{\substack{j \in A \\ j \neq i}} d(i,j) \qquad (1)$$

where *d* is some distance function. We move the object i' for which (1) attains its maximal value from *A* to *B*.

Step 2. In each subsequent stage, we look for another object to move from *A* to *B*. As long as *A* still contains more than one object, we compute:

$$D(i, A-\{i\}) - D(i,B) = \frac{1}{|A|-1} \sum_{\substack{j \in A \\ j \neq i}} d(i,j) - \frac{1}{|B|} \sum_{h \in B} d(i,h) \qquad (2)$$

for each object $i \in A$ and we consider the object i'' that maximizes this quantity. When the maximal value of (2) is strictly positive, we move i'' from *A* to *B* and we repeat Step 2. On the other hand, when the maximal value of (2) is negative or zero we stop the process and the division of *N* into *A* and *B* is completed.

6. ACKNOWLEDGMENTS

This work was supported by DARPA grant F30602-98-2-0099 (Control of Agent Based Systems Program).

7. REFERENCES

[1] Bresee, J.S., Heckerman, D., and Kadie, C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pp. 43-52, San Francisco, July 24-26,

[2] Friedman, E.J. and Resnick, P. (1999) The Social Cost of Cheap Pseudonyms. Working paper[5]. An earlier version was presented at the *Telecommunications Policy Research Conference*, Washington, DC, October 1998.

[3] Goldberg, D., Nichols, D., Oki, B.M., and Terry, D. (1992) Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM* 35 (12), pp. 61-70, December 1992.

[4] Kollock, P. (1999) The Production of Trust in Online Markets. In *Advances in Group Processes* (Vol. 16), eds. E.J. Lawler, M. Macy, S. Thyne, and H.A. Walker, Greenwich, CT: JAI Press.

[5] Macnaughton-Smith, P., Williams, W.T., Dale, M.B., and Mockett, L.G. (1964), Dissimilarity analysis: A New Technique of Hierarchical Sub-division. *Nature* (202), pp. 1034-35.

[6] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994) Grouplens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work*, pp. 175-186, New York, NY: ACM Press.

[7] Shapiro, C. (1982) Consumer Information, Product Quality, and Seller Reputation. *Bell Journal of Economics* 13 (1), pp 20-35, Spring 1982.

[8] Zacharia, G., Moukas, A., and Maes, P. (1999) Collaborative Reputation Mechanisms in Online Marketplaces. In *Proceedings of 32nd Hawaii International Conference on System Sciences (HICSS-32)*, Maui, Hawaii, January 1999.

[5] Available from http://www.si.umich.edu/~presnick/papers/identifiers/index.html.