MECHANISMS FOR COPING WITH UNFAIR RATINGS AND DISCRIMINATORY BEHAVIOR IN ONLINE REPUTATION REPORTING SYSTEMS

Chrysanthos Dellarocas
Sloan School of Management
Massachusetts Institute of Technology
U.S.A.

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 a set of mechanisms that 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 deceitful buyers and sellers.

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 (Kollock 1999). 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, 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 (1999), 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. (1999) 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 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 RATINGS AND DISCRIMINATORY BEHAVIOR IN ONLINE TRADING COMMUNITIES

In most online trading communities, participants can be distinguished into buyers and sellers. We further assume that only buyers can rate sellers. In a future study, we will consider the implications of 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 aspects of quality (for example, Zagat's 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 past ratings using a family of techniques collectively known as *collaborative filtering* (Resnick et al. 1994). 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 (Bresee et al. 1998).

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:

1. 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, allowing that seller to receive more orders from buyers and at a higher price than deserved.

- *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.

2. 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 scenarios presented here 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 the dispersion 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 SOLUTION

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 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 (Kollock 1999; Friedman and Resnick 1999), 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 consumer-to-consumer [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 afterward. 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.

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 of the colluding 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. (1964), in order to separate the set of raters N into two

clusters N_l and N_u based on their average 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. STEADY STATE ANALYSIS OF PROPOSED SOLUTION

As with all filtering approaches, it is important to ascertain that our proposed approach does a good job of removing unwanted noise while not significantly distorting the useful signal content.

The simplest scenario to analyze is one where 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 $N(\mu', \sigma')$[1] and unfair ratings a distribution $N(\mu', \sigma')$.[2] In such a scenario, the *fair* reputation estimate $\hat{R}_b(s)_{fair} \approx \mu'$.

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.[3] 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. In the absence of cluster filtering, the *biased* reputation estimate calculated by the above algorithm will approximate the average of all ratings in N, that is:

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

The strategy which maximizes the above reputation estimate is one where $\mu' = 100$ and $\sigma' = 0$, i.e., where all colluding buyers give the maximum possible rating to the seller. We define the *reputation bias B* to be the maximum difference (expressed in reputation units) between the biased and fair reputation estimates for a given pair of μ and σ 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 σ (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. to a large number of test cases and have calculated the maximum positive reputation bias that colluding 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 bias achieved by any of them, with and without cluster filtering. The results are plotted in Figures 2 and 3.

The most important conclusions of this analysis can be summarized as follows:

- The optimal collusion 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 $0.4\mu'$ (actual values depend on the fair rating mean μ). Intuitively, this means that colluders 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 in the lower cluster N_l.

- The maximum reputation bias rises with the percentage of unfair raters in the nearest neighbor set and is inversely proportional to the fair reputation estimate.

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

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

[3] We assume that due to the use of controlled anonymity, unfair negative ratings are not an issue in this analysis.

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

Figure 1. Test Case Parameters.

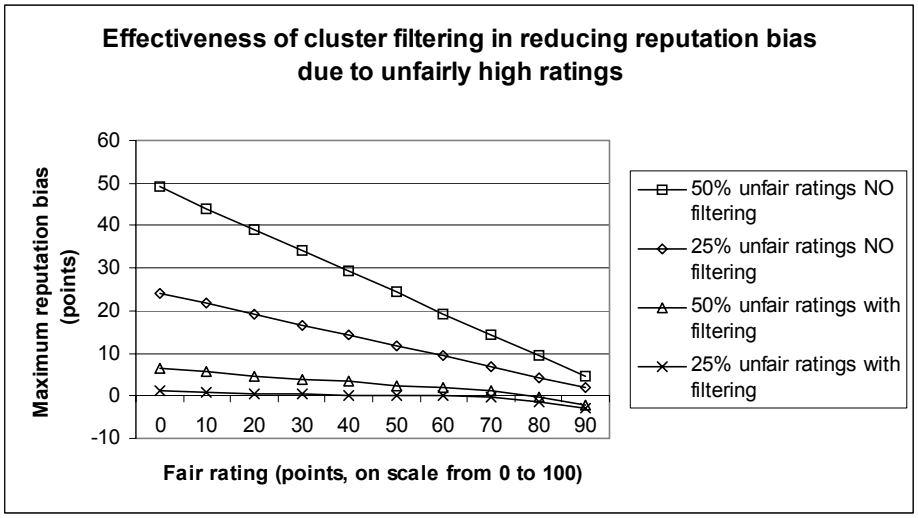

Figure 2. Cluster Filtering Significantly Reduces Reputation Bias

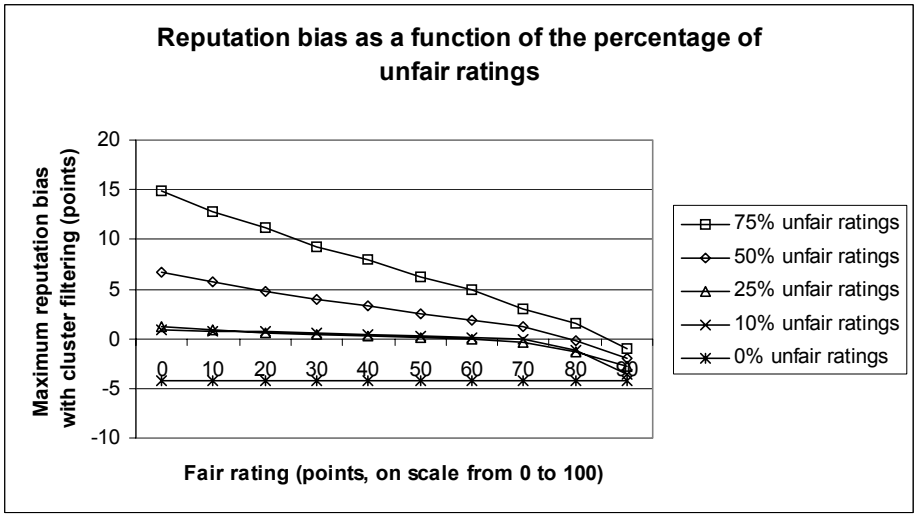

Figure 3. Even with Cluster Filtering, Reputation Bias Increases with the Fraction of Unfair Ratings

- 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.

The effectiveness of cluster filtering in significantly reducing reputation bias is evident from those results.

5. ONGOING AND FUTURE WORK

Our initial results indicate that the combination of controlled anonymity and cluster filtering is a powerful technique for improving the reliability of reputation systems in the presence of deceitful buyers and sellers. Our ongoing work extends our analysis in a number of important directions:

- Proposing concrete ways in which controlled anonymity can be implemented in various important practical domains

- Investigating techniques for reducing the fraction δ of unfair raters that make it into the nearest neighbor set N

- Analyzing the performance of cluster filtering when ratings vary over time

- Experimenting with alternative clustering algorithms for separating N_u and N_l

Some of the results of this ongoing work have been reported in Dellarocas (2000). Some additional results will be reported at the conference.

References

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

Dellarocas, C. "Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior," in *Proceedings of the Second ACM Conference on Electronic Commerce*, Minneapolis, MN, October 17-20, 2000.

Friedman, E. J., and Resnick, P. "The Social Cost of Cheap Pseudonyms," Working Paper, School of Information, University of Michigan, Ann Arbor (available online at http://www.si.umich.edu/~presnick/papers/identifiers/index.html). An earlier version was presented at the *Telecommunications Policy Research Conference*, Washington, DC, October 1998.

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

Macnaughton-Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. "Dissimilarity Analysis: A New Technique of Hierarchical Sub-division," *Nature* (202), 1964, pp. 1034-1035.

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

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

Zacharia, G., Moukas, A., and Maes, P. "Collaborative Reputation Mechanisms in Online Marketplaces," in *Proceedings of Thirty-Second Hawaii International Conference on System Sciences (HICSS-32)*, Los Alamitos, CA: IEEE Computer Society Press, January 1999.