Chapter VII

Building Trust Online: The Design of Robust Reputation Reporting Mechanisms for Online Trading Communities

Chrysanthos Dellarocas
Massachusetts Institute of Technology, USA

ABSTRACT

Several properties of online interaction are challenging the accumulated wisdom of trading communities on how to produce and manage trust. Online reputation systems have emerged as a promising trust management mechanism in such settings. The objective of this chapter is to contribute to the construction of online reputation systems that are robust in the presence of unfair and deceitful raters. The chapter sets the stage in identifying a number of important ways in which the reliability of the current generation of reputation systems can be compromised by unfair buyers and sellers. The central contribution of the chapter is a number of novel "immunization mechanisms" for countering the undesirable effects of such fraudulent behavior. The chapter describes the mechanisms, proves their properties and explains how various parameters of the marketplace, most notably the anonymity and authentication regimes, can influence their effectiveness. Finally, it concludes by discussing the implications of the findings for managers and users of current and future electronic marketplaces and identifies some important open issues for future research.

INTRODUCTION

The emergence of electronic markets and other types of online trading communities are changing the rules on many aspects of doing business. Electronic markets promise substantial gains in productivity and efficiency by bringing together a much larger set of buyers and sellers, and substantially reducing search and transaction costs (Bakos, 1997). In theory, buyers can then look for the best possible deal and end up transacting with a different seller on every single transaction. None of these theoretical gains will be realized, however, unless market makers and online community managers find effective ways to produce trust among their members. The production of trust is thus emerging as an important management challenge in any organization that operates or participates in online trading communities.

Several properties of online communities challenge the accumulated wisdom of our societies on how to produce trust (Kollock, 1999). Formal institutions, such as legal guarantees, are less effective in global electronic markets that span multiple jurisdictions with often conflicting legal systems (Johnson & Post, 1996). The difficulty is compounded by the fact that, in many electronic markets, it is relatively easy for trading partners to suddenly "disappear" and reappear under a different online identity (Friedman & Resnick, 2001).

As a counterbalance to those challenges, electronic communities are capable of storing complete and accurate information about all transactions they mediate. Several researchers and practitioners have, therefore, started to look at ways in which this information can be aggregated and processed by the market makers or other trusted third parties in order to help online buyers and sellers assess each other's trustworthiness. This has led to a new breed of systems, which are quickly becoming an indispensable component of every successful online trading community: *online feedback mechanisms* (Dellarocas, 2003), also known as *reputation systems* (Resnick, Zeckhauser, Friedman, & Kuwabara, 2000), are using the Internet's bi-directional communication capabilities to artificially engineer large-scale word-of-mouth networks in which individuals share opinions and experiences on a wide range of topics, including companies, products, services, and even world events. Figure 1 lists several noteworthy examples of such mechanisms in use today.

The disembodied nature of online environments introduces several challenges related to the interpretation and use of online feedback. Some of these challenges have their roots in the subjective nature of feedback information. Brick-and-mortar settings usually provide a wealth of contextual cues that assist in the proper interpretation of opinions and gossip (such as familiarity with the person who acts as the source of that information, the ability to draw inferences from the source's facial expression or mode of dress, etc.). Most of these cues are absent from online settings. Readers of online feedback are thus faced with the task of evaluating the opinions of complete strangers. Other challenges to feedback interpretation have their root in the ease with which online identities can be changed. This opens the door to various forms of strategic manipulation. For example, community members can use fake online identities to post dishonest feedback and thus try to inflate their reputation or tarnish that of their competitors. An important prerequisite for the widespread acceptance of online feedback mechanisms is, therefore, a better understanding of how such systems can be compromised, as well as the development of adequate defenses.

Figure 1. Examples of Online Feedback Mechanisms (In Use as of April 2003)

Website	Category	Summary of feedback mechanism	Format of solicited feedback	Format of published feedback
Citysearch	Entertainment guide	Users rate restaurants, bars, clubs, hotels and shops	Users rate multiple aspects of reviewed items from one to 10 and answer a number of yes/no questions; readers rate reviews as "useful", "not useful", etc.	Weighted averages of ratings per aspect reflecting both user and editorial ratings; user reviews can be sorted according to "usefulness"
eBay	Online auction house	Buyers and sellers rate one another following transactions	Positive, negative or neutral rating plus short comment; rated party may post a response	Sums of positive, negative and neutral ratings received during past six months
eLance	Professional services marketplace	Contractors rate their satisfaction with subcontractors	Numerical rating from one to five plus comment; rated party may post a response	Average of ratings received during past six months
Epinions	Online opinions forum	Users write reviews about products/services; other members rate the usefulness of reviews	Users rate multiple aspects of reviewed items from one to five; readers rate reviews as "useful", "not useful", etc.	Averages of item ratings; % of readers who found a review "useful"
Google	Search engine	Search results are ordered based on how many sites contain links that point to them	A Web page is rated based on how many links point to it, how many links point to the pointing page, etc.	No explicit feedback scores are published; ordering acts as an implicit indicator of reputation
Slashdot	Online discussion board	Postings are prioritized or filtered according to the ratings they receive from readers	Readers rate posted comments	

The objective of this chapter is to contribute to the construction of online reputation systems that are robust in the presence of unfair and deceitful raters. The chapter sets the stage by identifying a number of important ways in which the predictive value of online reputation systems can be compromised by unfair buyers and sellers. The central contribution of the chapter is a number of novel "immunization mechanisms" for countering the undesirable effects of such fraudulent behavior. The chapter describes the mechanisms, proves their properties, and explains how various parameters of the marketplace, most notably the anonymity and authentication regimes, can influence their effectiveness. Finally, it concludes by discussing the implications of the findings for managers and users of current and future electronic marketplaces, and identifies some open issues for future research.

UNFAIR RATINGS IN ONLINE REPUTATION SYSTEMS

This section looks at this problem of unfair online ratings in more detail. More specifically, our goal is to study a number of unfair rating scenarios and analyze their effects in compromising the reliability of an online reputation system.

The setting of this chapter is a large-scale B2C marketplace, such as eBay or eLance.com, where consumers transact with a large number of sellers. In a typical transaction t, a buyer b contracts with a seller s for the provision of a service. Upon conclusion of the transaction, b provides a numerical rating $R_b^s(t)$, reflecting some attribute

Q of the service offered by s as perceived by b (ratings can only be submitted in conjunction with a transaction). For the sake of simplicity, I assume that $R_b^s(t)$is a scalar quantity, although in most transactions there are several quality attributes and $R_b^s(t)$ would be a vector.

I further assume the existence of a ratings aggregation mechanism, whose goal is to store and process past ratings in order to calculate reliable personalized "reputation" estimates $\hat{R}_b^s(t)$ for seller s upon request of a prospective buyer b. In settings where the attribute Q for which ratings are provided is subjectively measurable, there exist four scenarios where buyers and/or sellers can intentionally try to "rig the system," resulting in biased reputation estimates that deviate from a "fair" assessment of attribute Q for a given seller:

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.

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. If the rated attribute is not objectively measurable, it will be very difficult or impossible to distinguish ratings dispersion due to genuine taste differences from that which is due to unfair ratings or discriminatory behavior.

In the following analysis, I assume the use of *collaborative filtering* techniques in order to address the issue of subjective ratings (Goldberg, Nichols, Oki, & Terry, 1992; Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994; Shardanand & Maes, 1995; Billsus & Pazzani, 1998). More specifically, I assume that, in order to estimate the *personalized* reputation of s from the perspective of b, some collaborative filtering technique is used 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 the similarity of their ratings (for other commonly rated sellers) with those of b. Sometimes, this step will filter out all 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.

Effects When Seller Behavior is Steady Over Time

The simplest scenario to analyze is one where we can assume that seller behavior, and therefore the attribute *Q* that is being rated by buyers, remains steady over time. That means, collaborative filtering algorithms can take into account all ratings in their database, no matter how old.

To make our analysis more concrete, I will make the assumption that fair ratings can range between $[R_{\min}, R_{\max}]$ and that they follow a distribution of the general form:

$$\tau_b^s(R) = \max(R_{\min}, \min(R_{\max}, z)) \text{ where } z \sim N(\mu, \sigma) \quad (1)$$

which in the rest of the chapter will be approximated by $\tau_b^s(R) \approx N(\mu, \sigma)$. The introduction of minimum and maximum rating bounds corresponds nicely with common practice. For example, Amazon.com allows buyers to rate products on a scale from 1 to 5. The assumption of normally distributed fair ratings requires more discussion. It is based on the previous assumption that those ratings belong to the nearest-neighbor set of a given buyer, and therefore represent a single taste cluster. Within a taste cluster, it is expected that fair ratings will be relatively closely clustered around some value and hence the assumption of normality.

In this chapter I focus on the reliable estimation of the unknown quality attribute *Q*. Suppose that the true value of Q is equal to μ. The goal of a reliable reputation system is the calculation of a fair *mean reputation estimate* (MRE) $\hat{R}_b^s$ that is equal to or very close to the mean of the fair ratings distribution in the nearest-neighbor set (an unbiased estimator of μ). Ideally, therefore:

$$\hat{R}_{b,fair}^s = \mu \quad (2)$$

The goal of unfair raters is to strategically introduce unfair ratings in order to *maximize* the distance between the *actual* MRE $\hat{R}_{b,actual}^s$ calculated by the reputation system and the fair MRE. More specifically the objective of a ballot-stuffing agent is to maximize the MRE while bad-mouthing agents aim to minimize it. Note that, in contrast to the case of fair ratings, it is not safe to make *any* assumptions about the form of the distribution of unfair ratings. Therefore, all analyses in the rest of this chapter will calculate system behavior under the most disruptive possible unfair ratings strategy.

I will only analyze the case of ballot stuffing; the case of bad mouthing is symmetric. Assume that the initial collaborative filtering step constructs a nearest-neighbor set *N*, in which the proportion of unfair raters is d and the proportion of fair raters is (1-δ). Furthermore, my analysis assumes that the actual MRE $\hat{R}_{b,actual}^s$ is taken to be the sample mean of the *most recent rating* given to *s* by each qualifying rater in *N*. This simple estimator is consistent with the practice of most current-generation commercial

Figure 2. Maximum MRE Bias when MREs are Based on the Mean of the Ratings Set; Highlighted Cells Indicate Biases Above 5% of the Ratings Range

Percentage of unfair ratings	Fair Mean Reputation Estimate (R_{min} = 0, R_{max} = 9)				
	0	2	4	6	8
	Reputation Bias				
9%	***0.81***	***0.63***	0.45	0.27	0.09
18%	***1.62***	***1.26***	***0.90***	***0.54***	0.18
27%	***2.43***	***1.89***	***1.35***	***0.81***	0.27
36%	***3.24***	***2.52***	***1.80***	***1.08***	0.36
45%	***4.05***	***3.15***	***2.25***	***1.35***	0.45

recommender systems (Schafer, Konstan, & Riedl, 2001). In that case, the *actual* MRE is approximately equal to:

$$\hat{R}^{s}_{b,actual} \cong (1-\delta)\cdot\mu + \delta\cdot\mu_u \tag{3}$$

where μ_u is the mean value of unfair ratings. The strategy that maximizes the above MRE is one where $\mu_u = R_{\max}$, i.e., where all unfair buyers give the maximum possible rating to the seller.

I define the *mean reputation estimate bias* for a contaminated set of ratings to be:

$$B = \hat{R}^{s}_{b,actual} - \hat{R}^{s}_{b,fair} \tag{4}$$

In the above scenario, the maximum MRE bias is given by:

$$B_{\max} = (1-\delta)\cdot\mu + \delta\cdot R_{\max} - \mu = \delta\cdot(R_{\max} - \mu) \tag{5}$$

Figure 2 tabulates some values of $B_{\max}$ for several different values m and d, in the special case where ratings range from [0,9]. For the purpose of comparing this baseline case with the "immunization mechanisms" described in this chapter, I have highlighted biases above 5% of the ratings range (i.e., biases greater than ±0.5 points on ratings which range from 0-9). As can be seen, equation (5) can result in very significant inflation of a seller's MRE, especially for small μ and large δ.

Effects When Seller Behavior Varies Over Time

This section expands our analysis by discussing some additional considerations, which arise in environments where seller behavior may vary over time. I identify some additional unfair rating strategies that can be very disruptive in such environments.

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 analysis of the economic effects of reputation, Shapiro

(1981) 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 proved 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 chapter I base my analysis on an approach that approximates Shapiro's desiderata, but is simpler to implement and analyze. The principal idea is to calculate time-varying personalized MREs $\hat{R}_b^s(t)$ as averages of ratings submitted within the most recent time window $W = [t\text{-}\varepsilon, t]$. 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 a time varying personalized MRE $\hat{R}_b^s(t)$, 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 $t \in W$.

According to equation (5), the maximum reputation bias due to unfair ratings is proportional to the ratio δ of unfair ratings that "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, Moukas, and Maes (1999) 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-neighbor set $N_{initial}$ contains m fair raters and n unfair raters. In most cases n would be much smaller than m. Assume further that the average inter-arrival time of fair ratings for a given seller is λ, and that personalized MREs $\hat{R}_b^s(t)$ are based only on ratings for s submitted by buyers $u \in N_{initial}$ 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 five and 20). For k much smaller than m, I 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. (1999), I 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 $\delta = n/(n+k)$. 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 $k = 5$, $\delta = 10/(10+5) = 0.67$. We therefore see that, for relatively small time windows, even a small (e.g.,

five to 10) number of colluding buyers can successfully use unfair ratings flooding to dominate the set of ratings used to calculate MREs and completely bias the estimate provided by the system.

The results of this section indicate that even a relatively small number of unfair raters can significantly compromise the reliability of online reputation systems. This requires the development of effective measures for addressing the problem. The next section proposes and analyzes several such measures.

MECHANISMS FOR IMMUNIZING ONLINE REPUTATION SYSTEMS AGAINST UNFAIR RATER BEHAVIOR

Having recognized the problem of unfair ratings as a real and important one, this section proposes a number of mechanisms for eliminating or significantly reducing its adverse effects on the reliability of online reputation systems.

Avoiding Negative Unfair Ratings Using Controlled Anonymity

The main argument of this section is that the anonymity regime of an online community can influence the kinds of reputation system attacks that are possible. A slightly surprising result is the realization that a fully transparent marketplace, where everybody knows everybody else's true identity, incurs more dangers of reputation system fraud than a marketplace where the true identities of traders are carefully concealed from each other, but are known to a trusted third entity (usually the market-maker).

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. Usually, victims are selected based on some real-life attributes of their associated principal entities (for example, because they are our competitors or because of religious or racial prejudices). This adverse selection process can be avoided if the community conceals the true identities of the buyers and sellers from each other.

In such a "controlled anonymity" scheme, the marketplace knows the true identity of all market participants by applying some effective *authentication process* before it allows access to any agent (Hutt, Bosworth, & Hoyt, 1995). 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 that change from one transaction to the next, in order to make identity detection very difficult). 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 (Kollock, 1999; Friedman

& Resnick, 2001), here we have an example of a situation where some controlled degree of 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.

If concealing the identities of both parties from each other is not possible, then it may still be useful to conceal the identity of one party only. More specifically, concealing the identity of buyers but not sellers avoids negative discrimination against buyers but does not avoid bad mouthing of sellers. In an analogous manner, concealing the identity of sellers but not buyers avoids bad-mouthing but not negative discrimination. These results are summarized in Figure 3.

Generally speaking, concealing the identities of buyers is usually easier than concealing the identities of sellers (a similar point is made in Cranor & Resnick, 1999). 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 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 for malevolent buyers to intentionally pick specific distributors for subsequent bad mouthing.

It is important to note at this point that 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. The following two sections propose some filtering mechanisms, which are applicable in the cases of ballot stuffing as well.

Reducing the Effect of Unfair Ratings Using Median Filtering

In the second section of this chapter, I based the calculation of reputation bias on the assumption that MREs are based on the sample mean of the nearest-neighbor set. In this section I will demonstrate that the effect of unfair ratings can be significantly reduced if, instead of the sample mean, the calculation of MREs is based on the sample median.

The field of robust statistics has devoted considerable attention to the problem of finding estimators of "location" (mean value), which are robust in the presence of contaminated samples (Huber, 1981). Nevertheless, most of that literature treats contamination as "innocent" noise and does not address the problem of malicious raters who, based on their knowledge of the estimator used, strategically distribute unfair ratings in

Figure 3. Effectiveness of Controlled Anonymity in Preventing Certain Classes of Unfair Behavior

Anonymity Regime		Classes of possible unfair behavior			
Buyer's identity known to seller	***Seller's identity known to buyer***	***Bad-mouthing possible***	***Negative discrimination possible***	***Ballot-stuffing possible***	***Positive discrimination possible***
Yes	Yes	✔	✔	✔	✔
Yes	No		✔	✔	✔
No	Yes	✔		✔	✔
No	No			✔	✔

order to maximize the achievable bias. To the knowledge of the author, the analysis presented in this section is novel.

Definition: The *sample median* $\tilde{Y}$ of n ordered observations $Y_1 \le Y_2 \le ... \le Y_n$ is the middle observation Y_k where $k = (n+1)/2$ if n is odd. When n is even, then $\tilde{Y}$ is considered to be any value between the two middle observations Y_k and Y_{k+1} where $k = n/2$, although it is most often taken to be their average.

In the absence of unfair ratings (i.e., when $\delta = 0$), I previously assumed that $\tau_b^s(R) \approx N(\mu, \sigma)$. It is well known (Hojo, 1931) that as the size n of the sample increases, the median of a sample drawn from a normal distribution converges rapidly to a normal distribution with a mean equal to the median of the parent distribution. In normal distributions, the median is equal to the mean. Therefore, in situations where there are no unfair raters, the use of the sample median results in unbiased fair MREs:

$$\hat{R}^s_{b,fair} \cong \mu \qquad (6)$$

Let us now assume that unfair raters know that MREs are based on the sample median. They will strategically try to introduce unfair ratings whose values will maximize the absolute bias between the sample median of the fair set and the sample median of the contaminated set. More specifically, "ballot stuffers" will try to maximize that bias while "bad mouthers" will try to minimize it. In the following analysis I consider the case of ballot stuffing. The case of bad mouthing is symmetric, with the signs reversed.

Proposition 1: Assume that the nearest neighbor set consists of $n_f = (1-\delta) \cdot n$ fair ratings and $n_u = \delta \cdot n$ unfair ratings, where $0 \le \delta < 0.5$ and n are sufficiently large. If MREs are based on the sample median and fair ratings are drawn from a normal distribution with standard deviation s, then the maximum MRE bias achievable by a strategic "ballot-stuffer" is asymptotically equal to:

$$E[B_{\max}] = E[\hat{R}^s_{b,actual} - \hat{R}^s_{b,fair}] = \sigma \cdot \Phi^{-1}\left[\frac{1}{2 \cdot (1-\delta)}\right] \quad (7)$$

where $\Phi^{-1}(q)$ is the inverse standard normal CDF.

Proof: See Appendix.

Figure 4 shows some of the values of $E[B_{\max}]$for various values of δ and σ in the special case where ratings range from 0 to 9. The maximum bias increases with the percentage of unfair ratings and is directly proportional to the standard deviation of the fair ratings. As before, I have highlighted maximum average biases of 5% of the rating range or more. Figure 4 shows that the use of the sample median as a basis of calculating MREs manages to reduce the maximum average bias to below 5% of the rating range for unfair rater ratios of up to 30% to 40% and a wide range of fair rating standard deviations.

Using Frequency Filtering to Eliminate Unfair Ratings Flooding

Equations (5) and (7) confirm the intuitive fact that the reputation bias due to unfair ratings increases with the ratio δ of unfair raters in a given sample. In settings where a seller's quality attributes may vary over time, calculation of reputation should be based on recent ratings only using time discounting or a time-window approach. In those cases, as demonstrated earlier, by "flooding" the system with ratings, a relatively small number of unfair raters can manage to increase the ratio δ of unfair ratings in any given time window above 50% and completely compromise the reliability of the system.

This section proposes an approach for immunizing a reputation system against unfair ratings flooding. The main idea is to filter raters in the nearest-neighbor set based on their ratings submission frequency.

Description of Frequency Filtering

Step 1: Frequency filtering depends on estimating the average frequency of ratings submitted by *each* buyer for a given seller. Since this frequency is a time-varying quantity (sellers can become more or less popular with the passage of time), it too needs to be estimated using a time-window approach. More specifically:

1. Calculate the set $F^s(t)$ of *buyer-specific* average rating submission frequencies $\bar{f}^s_b(t)$ for each buyer b that has submitted ratings for seller s during the rating submission frequency calculation time window $W_f = [t\text{-}E, t]$. More precisely:

   $\bar{f}^s_b(t)$ = (number of ratings submitted by b for s during W_f)/E.

2. Set the cutoff frequency $\bar{f}^s_{cutoff}(t)$ to be equal to the k^{th} order statistic of the set $F^s(t)$ where $k = (1-D)n$, n is the number of elements of $F^s(t)$, and D is a conservative estimate of the fraction of unfair raters in the total buyer population for seller s. For

Figure 4: Asymptotic Upper Bounds of Average Reputation Bias when MREs are Based on the Median of the Ratings Set (Ratings Range from 0-9); Highlighted Cells Indicate Biases Above 5% of the Ratings Range

Percentage of unfair ratings	Standard Deviation of Fair Ratings			
	0.25	**0.50**	**0.75**	**1.00**
	Reputation Bias			
9%	0.03	0.06	0.09	0.13
18%	0.07	0.14	0.21	0.28
27%	0.12	0.24	0.37	0.49
36%	0.20	0.40	***0.59***	***0.79***
45%	0.35	***0.69***	***1.04***	***1.38***

example, if we assume that there are no more than 10% unfair raters among all the buyers for seller s, then $D = 0.1$. Assuming further that $n = 100$, i.e., that the set $F^s(t)$ contains average rating submission frequencies from 100 buyers, then the cutoff frequency would be equal to the 90th smallest frequency (the 10th largest frequency) present in the set $F^s(t)$.

The width E of the ratings submission frequency calculation time window W_f should be large enough to contain at least a few ratings from all buyers for a given seller.

Step 2: During the calculation of an MRE for seller s, eliminate all raters b in the nearest-neighbor set for whom $\bar{f}_b^s > \bar{f}_{cutoff}^s$. In other words, eliminate all buyers whose average ratings submission frequency for seller s is above the cutoff frequency.

Analysis of Frequency Filtering

Frequency filtering provides effective protection against unfair ratings flooding by guaranteeing that the ratio of unfair raters in the MRE calculation set cannot be more than twice as large as the ratio of unfair raters in the total buyer population.

As before, I will assume that the entire buyer population is n, unfair raters are $\delta \cdot n << n$ and the width of the reputation estimation time window is a relatively small W (so that, each rating within W typically comes from a different rater). The following proposition then holds:

Proposition 2: Assume that the frequency of fair ratings is uniformly distributed. Then, after applying frequency filtering to the nearest-neighbor set of raters, the ratio of unfair raters d in the total population of buyers and the ratio δ' of unfair ratings remaining in the nearest-neighbor set satisfy the inequality:

$$\delta / (1 - \delta) \leq \delta' \leq 2\delta \qquad (8)$$

Proof: See Appendix.

Equation (8) shows that, no matter how hard unfair raters may try to "flood" the system with ratings, the presence of frequency filtering guarantees that they cannot inflate their presence in the final MRE calculation set by more than a factor of two.

In most online communities, the exact ratio δ of unfair raters will not be known exactly. In such cases, if we have a belief that $\delta < 0.1$, then setting $D = 0.1$ has been experimentally proven to result in inflation ratios, which also fall within the bounds of equation (8).

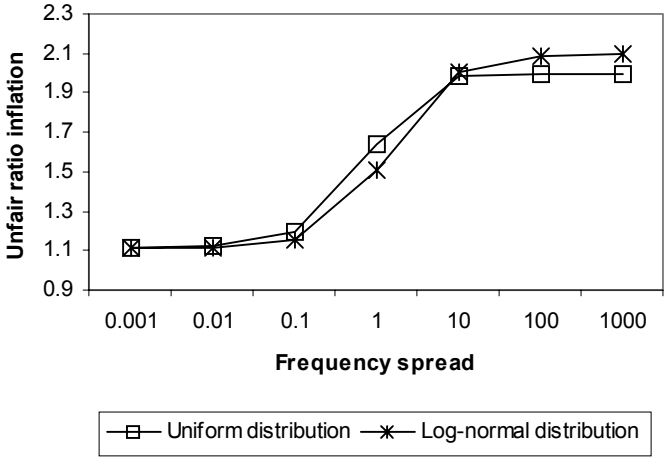
A more realistic assumption about fair ratings frequencies is that they follow a lognormal distribution. This assumption is consistent with the findings of researchers in marketing (Lawrence, 1980). In this case, the expression for the final ratio δ' cannot be given in closed form. However, a numerical solution yields results, which approximate very closely those obtained analytically for uniformly distributed fair rating frequencies (Figure 5).

One possible criticism of the frequency filtering approach is that it potentially eliminates those fair buyers who transact most frequently with a given seller. In fact, in the absence of unfair raters, all raters who would be filtered out based on their high ratings submission frequency would be fair raters. Nevertheless, I do not believe that this property constitutes a weakness of the approach. I argue that the "best customers" of a given seller often receive preferential treatment, which is in a way a form of positive discrimination on behalf of the seller. Therefore, I believe that the potential elimination of such raters from the final reputation estimate in fact benefits the construction of more unbiased estimates for the benefit of first-time prospective buyers.

Issues in Communities Where Buyer Identity is Not Authenticated

The effectiveness of frequency filtering relies on the assumption that a given principal can only have one buyer agent acting on its behalf in a given marketplace. The

Figure 5: Maximum Unfair Ratings Inflation Factors Achievable Through Flooding when Frequency Filtering is Used ($\delta = D = 0.1$); Frequency Spread Indicates the Difference Between the Maximum and Minimum Rating Submission Frequencies



technique is also valid in situations where principals have multiple buyer agents with authenticated identifiers. In that case, frequency filtering works if we consider all agents of a given principal as a single buyer for frequency filtering purposes.

In non-authenticated online communities (communities where "pseudonyms" are "cheap," to use the term of Friedman and Resnick) with time-windowed reputation estimation, unfair buyers can still manage to "flood" the system with unfair ratings by creating a large number of pseudonymously known buyer agents acting on their behalf. In that case the total ratio d of unfair agents relative to the entire buyer population can be made arbitrarily high. If each of the unfair agents takes care of submitting unfair ratings for seller *s* with frequency $f_b^s \leq f_{cutoff}$, because δ will be high, even in the presence of frequency filtering, unfair raters can still manage to severely contaminate a seller's fair reputation.

Further research is needed to develop immunization techniques that are effective in communities where the "true" identity of buyer agents cannot be authenticated. In the meantime, the observations of this section make a strong argument for using some reasonably effective authentication regime *for buyers* in online communities where trust is based on online reputation. For example, the community can require that all newly registering buyers supply a valid credit card for authentication purposes, or it can insert cookies into buyer computers so that attempts to assume different "false" identities from the same computer fail.

CONCLUSIONS AND MANAGERIAL IMPLICATIONS

I began this chapter by arguing that managers of online marketplaces should pay special attention to the design of effective trust management mechanisms that will help guarantee the stability, longevity, and growth of their respective communities. This chapter has contributed in this direction by presenting a number of novel techniques for "immunizing" online reputation systems against unfair ratings. The proposed techniques are summarized in Figure 6. The combination of frequency filtering and median filtering is capable of guaranteeing reputation biases of less than 5% (e.g., less than ±0.5 points when ratings range from one to 10) when the ratio of unfair raters is up to 15% to 20% of the total buyer population for a given seller.

The conclusions of this chapter are directly applicable to the design of current and future electronic marketplaces. More specifically, the analysis of the proposed techniques has resulted in a number of important guidelines that managers of online marketplaces should take into account in order to embed effective reputation systems into their respective communities:

- It is important to be able to authenticate the identity of rating providers. Unauthenticated communities are vulnerable to unfair rating "flooding" attacks.
- Concealing the (authenticated) identity of buyers and sellers from one another can prevent negative unfair ratings and discriminatory behavior. Managers of electronic marketplaces and B2B hubs can consider adding this function into the set of services they provide to their members.
- Numerical reputation estimates should be based on the median (and *not* the mean) of the relevant rating set. Also, frequency filtering should be applied in order to

Figure 6. Summary of Proposed Immunization Techniques

Technique	Description	Effect	Prerequisites
Controlled anonymity	Market-maker conceals the true identities of buyers and sellers from one another and only reveals their respective reputation estimates	Prevents bad-mouthing and/or negative discrimination	Ability to practically implement with reasonable cost
Median filtering	Calculation of reputation estimate using the median of the ratings set	Results in robust estimations in the presence of up to 30% to 40% of unfair ratings	Ratio of unfair ratings less than 50%
Frequency filtering	Ignores raters whose ratings submission frequency for a given seller is significantly above average	Eliminates raters who attempt to flood the system with unfair ratings; maintains the final ratio of unfair raters at low levels	Ability to authenticate the true identity of online raters

eliminate raters who might be attempting to flood ("spam") the system with potentially unfair ratings.

This chapter suggests several topics for further research. The calculation of robust estimates of reputation *variance*, the development of "immunization" techniques that avoid unfair ratings "flooding" in *non-authenticated* online communities, and the analysis of unfair ratings in environments where *bi-directional ratings* are possible (that is, both parties can rate one another) are just some of the issues left open by this work. It is our hope that the analysis and techniques proposed by this work will provide a useful basis that will stimulate further research in the important and promising field of online reputation systems.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (CAREER Grant No. 9984147) and an MIT Center for eBusiness Vision Fund Award.

REFERENCES

Bakos, Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management Science*, *43*(12).

Billsus, D., & Pazzani, M.J. (1998). Learning collaborative information filters. *Proceedings of the 15th International Conference on Machine Learning* (pp. 46-54), July.

Bresee, J.S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)* San Francisco (July 24-26, pp. 43-52).

Cadwell, J.H. (1952). The distribution of quantiles of small samples. *Biometrika, 39,* 207-211.

Cranor, L.F., & Resnick, P. (2000). Protocols for automated negotiations with buyer anonymity and seller reputations. *Netnomics, 2*(1), 1-23.

Dellarocas, C. (2003). The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Science*, forthcoming.

Friedman, E., & Resnick, P. (2001). The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy, 10*(1).

Goldberg, D., Nichols, D., Oki, B.M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61-70.

Hojo, T. (1931). Distribution of the median, quartiles and interquartile distance in samples from a normal population. *Biometrika, 23,* 315-360.

Huber, P. (1981). *Robust Statistics*. New York: John Wiley & Sons.

Hutt, A.E., Bosworth, S., & Hoyt, D.B. (eds.). (1995). *Computer Security Handbook* (3rd edition). New York: John Wiley & Sons.

Johnson, D.R., & Post, D.G. (1996). Law and borders—the rise of law in cyberspace. *Stanford Law Review*, *48*.

Kollock, P. (1999). The production of trust in online markets. In E.J. Lawler, M. Macy, S. Thyne, & H.A. Walker (Eds.), *Advances in Group Processes* (Vol. 16). Greenwich, CT: JAI Press.

Lawrence, R.J. (1980). The lognormal distribution of buying frequency rates. *Journal of Marketing Research, XVII*(May), 212-226.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work* (pp. 175-186). New York: ACM Press.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45-48.

Schafer, J.B., Konstan, J., & Riedl, J. (2001). Electronic commerce recommender applications. *Journal of Data Mining and Knowledge Discovery*, *5*(1/2), 115-152.

Shapiro, C. (1982). Consumer information, product quality, and seller reputation. *Bell Journal of Economics, 13*(1), 20-35.

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating "word of mouth." *Proceedings of the Conference on Human Factors in Computing Systems* (CHI95) Denver, Colorado (pp. 210-217).

Zacharia, G., Moukas, A., & Maes, P. (1999). Collaborative reputation mechanisms in online marketplaces. *Proceedings of 32nd Hawaii International Conference on System Sciences (HICSS-32)*, Maui, Hawaii (January).

APPENDIX

Proof of Proposition 1

Assuming that the nearest neighbor set consists of $n_f = (1 - \delta) \cdot n$ fair ratings and $n_u = \delta \cdot n$ unfair ratings ($0 \le \delta < 0.5$), the most disruptive unfair ratings strategy—in terms of influencing the sample median—is one where all unfair ratings are higher than the sample median of the contaminated set. In that case and for $\delta < 0.5$, all the ratings that are lower than or equal to the sample median will have to be fair ratings. Then, the sample median of the contaminated set will be identical to the k^{th} order statistic of the set of n_f fair ratings, where $k = (n+1)/2$.

It has been shown (Cadwell, 1952) that, as the size n of the sample increases, the k^{th} order statistic of a sample drawn from a normal distribution $N(\mu, \sigma)$ converges rapidly to a normal distribution with mean equal to the q^{th} quantile of the parent distribution where $q = k/n$. Therefore, for large rating samples n, under the worst possible unfair ratings strategy, the sample median of the contaminated set will converge to x_q where x_q is defined by:

$$\Pr[R_b^s \le x_q] = q \Rightarrow x_q = \sigma \cdot \Phi^{-1}(q) + \mu$$

where $$q = \frac{k}{n_f} = \frac{n+1}{2 \cdot n_f} = \left(\frac{n+1}{n}\right) \cdot \left(\frac{1}{2 \cdot (1-\delta)}\right) \xrightarrow{n \to \infty} \frac{1}{2 \cdot (1-\delta)}$$

and $\Phi^{-1}(q)$ is the inverse standard normal CDF.

Given that $\hat{R}^s_{b,fair} \cong \mu$, the asymptotic formula for the average reputation bias, achievable by $\delta \cdot 100\%$ unfair ratings when fair ratings are drawn from a normal distribution $N(\mu, \sigma)$ and unfair ratings follow the most disruptive possible unfair ratings distribution, is given by:

$$E[B_{\max}] = E[\hat{R}^s_{b,actual} - \hat{R}^s_{b,fair}] = \sigma \cdot \Phi^{-1}\left[\frac{1}{2 \cdot (1-\delta)}\right] \qquad \text{QED}$$

Proof of Proposition 2

Assume that the entire buyer population is n, unfair raters are $\delta \cdot n$, and the width of the reputation estimation time window is a relatively small W (so that, each rating within W typically comes from a different rater). Then, after applying frequency filtering to the nearest-neighbor set of raters, in a typical time window we expect to find:

- $W \cdot (1-\delta) \cdot n \cdot \int_0^{f_{cutoff}} u \cdot \varphi(u) \cdot du$ fair ratings, where $\varphi(u)$ is the probability density function of fair ratings frequencies, and at most

- $W \cdot \delta \cdot n \cdot \alpha \cdot f_{cutoff}$ unfair ratings, where $0 \le \alpha \le 1$ is the fraction of unfair raters with submission frequencies below f_{cutoff}.

Therefore, the unfair/fair ratings ratio in the final set would be equal to:

$$\frac{\text{unfair ratings}}{\text{fair ratings}} = \frac{\delta'}{1-\delta'} = \frac{\delta}{1-\delta} \cdot \frac{\alpha \cdot f_{cutoff}}{\int_0^{f_{cutoff}} u \cdot \varphi(u) \cdot du} = \frac{\delta}{1-\delta} \cdot I \qquad (9)$$

where $I = \dfrac{\alpha \cdot f_{cutoff}}{\int_0^{f_{cutoff}} u \cdot \varphi(u) \cdot du}$ denotes the *inflation* of the unfair/fair ratings ratio in the final set relative to its value in the original set. The goal of unfair raters is to strategically distribute their rating frequencies above and below the cutoff frequency in order to maximize *I*. In contrast, the goal of the market designer is to pick the cutoff frequency f_{cutoff} so as to minimize *I*.

The cutoff frequency has been defined as the $(1-D)\cdot n^{\text{th}}$ order statistic of the sample of buyer frequencies, where $D \ge \delta$. For relatively large samples, this converges to the *q*-th quantile of the fair rating frequencies distribution, where *q* satisfies the equation:

$$(1-D)\cdot n = q\cdot(1-\delta)\cdot n + \alpha\cdot\delta\cdot n \Rightarrow \mathrm{q} = 1 - D + (\alpha-1)\cdot\frac{\delta}{1-\delta} \qquad (10)$$

From this point on, the exact analysis requires some assumptions about the probability density function of fair ratings frequencies. I assume a uniform distribution between $F_{\min} = f_0/(1+s)$ and $F_{\max} = f_0 \cdot (1+s)$. Let $S = F_{\max} - F_{\min}$. Then, by applying the properties of uniform probability distributions to equation (9), I get the following expression of the inflation *I* of unfair ratings:

$$I = \frac{2 \cdot S \cdot \alpha \cdot f_{cutoff}}{f_{cutoff}^2 - F_{\min}^2} \qquad \text{where } f_{cutoff} = F_{\max} - \frac{D + (\alpha-1)\cdot\delta}{1-\delta} \cdot S \qquad (11)$$

After some algebraic manipulation I find that $\frac{\partial I}{\partial \alpha} > 0$ and $\frac{\partial I}{\partial D} > 0$. This means that unfair raters will want to maximize *a*, the fraction of ratings that are less than or equal to f_{cutoff}, while market makers will want to minimize *D*, i.e., set *D* as close as possible to an accurate estimate of the ratio of unfair raters in the total population. Therefore, at equilibrium, $\alpha = 1, D = \delta$ and:

$$I = \frac{2 \cdot (F_{\max} - \varepsilon \cdot S)}{(1-\varepsilon) \cdot (F_{\min} + F_{\max} - \varepsilon \cdot S)} \text{ where } \varepsilon = \frac{\delta}{1-\delta} \quad (12)$$

The above expression for the unfair/fair ratings inflation depends on the spread S of fair ratings frequencies. At the limiting cases we get $\lim_{S \to 0} I = \frac{1}{1-\varepsilon}$ and $\lim_{S \to \infty} I = \frac{2}{1-\varepsilon}$.

By substituting the above limiting values of *I* in equation (9), we get the final formula for the equilibrium relationship between δ, the ratio of unfair raters in the total population of buyers, and δ′, the final ratio of unfair ratings remaining in the nearest-neighbor set using time windowing and frequency filtering:

$$\delta/(1-\delta) \le \delta' \le 2\delta \quad \text{QED}$$