
MASSACHUSETTS
INSTITUTE of TECHNOLOGY
LIBRARIES

Business@MIT

MIT Sloan School of Management

Sloan Working Paper 4180-01
eBusiness@MIT Working Paper 101
October 2001

BUILDING TRUST ON-LINE: THE DESIGN OF RELIABLE REPUTATION REPORTING BUILDING TRUST ON-LINE: THE DESIGN OF RELIABLE REPUTATION REPORTING

Chrysanthos Dellarocas

Building Trust On-Line: The Design of Reliable Reputation Reporting Mechanisms for Online Trading Communities

Chrysanthos Dellarocas

Sloan School of Management
Massachusetts Institute of Technology
Room E53-315
Cambridge, MA 02139
dell@mit.edu

Abstract:

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

1. Introduction

At the heart of any bilateral exchange there is a temptation, for the party who moves second, to defect from the agreed upon terms in ways that result in individual gains for it (and losses for the other party). For example, in transactions where the buyer pays first, the seller is tempted to not provide the agreed upon goods or services or to provide them at a quality which is inferior to what was advertised to the buyer. Unless there are some other guarantees, the buyer would then be tempted to hold back on her side of the exchange as well. In such situations, the trade will never take place and both parties will end up being worse off. Unsecured bilateral exchanges thus have the structure of a Prisoner's Dilemma.

Our society has developed a wide range of informal mechanisms and formal institutions for managing such risks and thus facilitating trade. The simple act of meeting face-to-face to settle a transaction helps reduce

the likelihood that one party will end up empty-handed. Written contracts, commercial law, credit card companies and escrow services are additional examples of institutions with exactly the same goals.

Although mechanism design and institutional support can help reduce transaction risks, they can never eliminate them completely. One example is the risk involving the exchange of goods whose "real" quality can only be assessed by the buyer a relatively long time *after* a trade has been completed (e.g. used cars). Even where society does provide remedial measures to cover risks in such cases (for example, the Massachusetts "lemon law"), these are usually burdensome and costly and most buyers would very much rather not have to resort to them. Generally speaking, the more the two sides of a transaction are separated in time and space, the greater the risks. In those cases, no transaction will take place unless the party who moves first possesses some sufficient degree of *trust* that the party who moves second will indeed honor its commitments. The production of trust, therefore, is a precondition for the existence of any market and civilized society in general (Dunn, 1984; Gambetta, 1990).

In "bricks and mortar" communities, the production of trust is based on several cues, often rational but sometimes purely intuitive. For example, we tend to trust or distrust potential trading partners based on their appearance, the tone of their voice or their body language. We also ask our already trusted partners about their prior experiences with the new prospect, under the assumption that past behavior is a relatively reliable predictor of future behavior. Taken together, these experiences form the *reputation* of our prospective partners.

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 the search and transaction costs (Bakos, 1997; Bakos, 1998). 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. Formal institutions, such as legal guarantees, are less effective in global electronic markets, which span multiple jurisdictions with, often conflicting, legal systems (Johnson and Post, 1996). For example, it is very difficult, and costly, for a buyer who resides in the U.S.A. to resolve a trading dispute with a seller who lives in Indonesia. 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 and Resnick, 1999; Kollock, 1999).

Furthermore, many of the cues based on which we tend to trust or distrust other individuals are absent in electronic markets where face-to-face contact is the exception. Finally, one of the motivating forces behind electronic markets is the desire to open up the universe of potential trading partners and enable transactions among parties who have never worked together in the past. In such a large trading space, most of one's already trusted partners are unlikely to be able to provide much information about the reputation of many of the other prospects that one may be considering.

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 lead to a new breed of systems, which are quickly becoming an indispensable component of every successful online trading community: electronic reputation reporting systems.

We are already seeing the first generation of such systems in the form of online *ratings*, *feedback* or *recommender systems* (Resnick and Varian, 1997; Schafer et.al., 2001). The basic idea is that online community members are given the ability to rate or provide feedback about their experiences with other community members. Feedback systems aim to build trust by aggregating such ratings of past behavior of their users and making them available to other users as predictors of future behavior. eBay (www.ebay.com), for example, encourages both parties of each transaction to rate one another with either a positive (+1), neutral (0) or a negative (-1) rating plus a short comment. eBay makes the cumulative ratings of its members, as well as all individual comments publicly available to every registered user.

The majority of the current generation of online feedback systems have been developed by Internet entrepreneurs and their reliability has not yet been systematically researched. In fact, there is ample anecdotal evidence, as well as one recent legal case[1], related to the ability to effectively manipulate people's actions by using online feedback forums (stock message boards in this case) to spread false opinions. As more and more organizations participate in electronic marketplaces, online reputation reporting systems deserve new scrutiny and the study of trust management systems in digital communities deserves to become a new addition to the burgeoning field of Management Science.

The objective of this paper is to contribute to the construction of online reputation reporting systems that are robust in the presence of unfair and deceitful raters. The paper sets the stage by providing a critical overview of the current state of the art in this area (Section 2). Following that, it identifies a number of important ways in which the predictive value of the current generation of reputation reporting systems can be severely compromised by unfair buyers and sellers (Section 3). The central contribution of the paper is a number of novel "immunization mechanisms" for effectively countering the undesirable effects of such fraudulent behavior. The paper describes the mechanisms, proves their properties and explains how various parameters of the marketplace microstructure, most notably the anonymity and authentication regimes, can

influence their effectiveness (Section 4). Finally, it concludes by discussing the implications of the findings for the managers and users of current and future electronic marketplaces and identifies some open issues for future research (Section 5).

2. Reputation reporting mechanisms in online communities

The relative ease with which computers can capture, store and process huge amounts of information about past transactions, makes past behavior (reputational) information a particularly promising way on which to base the production of trust in online communities. This fact, together with the fact that the other traditional ways of producing trust (institutional guarantees, indirect cues) do not work as well in cyberspace, has prompted researchers and practitioners to focus their attention on developing online trust building mechanisms based on reputational information. This section provides a critical survey of the state-of-the-art in this field.

A *reputation*, as defined by Wilson (Wilson, 1985) is a "characteristic or attribute ascribed to one person by another. Operationally, this is usually represented as a prediction about likely future behavior. It is, however, primarily an empirical statement. Its predictive power depends on the supposition that past behavior is indicative of future behavior". Reputation has been the object of study of the social sciences for a long time (Schmalensee, 1978; Shapiro, 1982; Smallwood and Conlisk, 1979). Several economists and game theorists have demonstrated that, in the presence of imperfect information, the formation of reputations is an important force that helps buyers manage transaction risks, but also provides incentives to sellers to provide good service quality.

Having interacted with someone in the past is, of course, the most reliable source of information about that agent's reputation. But, relying only on direct experiences is both inefficient and dangerous. Inefficient, because an individual will be limited in the number of exchange partners he or she has and dangerous because one will discover untrustworthy partners only through hard experience (Kollock, 1999). These shortcomings are especially severe in the context of online communities where the number of potential partners is huge and the institutional guarantees in case of negative experiences are weaker.

Great gains are possible if information about past interactions is shared and aggregated within a group in the form of *opinions*, *ratings* or *recommendations*. In the "bricks and mortar" communities this can take many forms: informal gossip networks, institutionalized rating agencies, professional critics, etc. In cyberspace, they take the form of online reputation reporting systems, also known as *online recommender systems* (Resnick and Varian, 1997). The following sections provide a brief discussion of the most important design challenges and categories of these systems.

2.1 Design challenges in online reputation reporting systems

Although the effective aggregation of other community members' opinions can be a very effective way to gather information about the reputation of prospective trading partners, it is not without pitfalls. The following paragraphs describe two important issues that need to be addressed by opinion-based reputation reporting mechanisms:

Subjectively measurable attributes.

In the rest of the paper we will use the term "agent" to refer to a participant (buyer or seller, human or software) of an online trading community. We say that an attribute Q of an agent s is *subjectively measurable* if identical behavior of agent s vis-à-vis two different agents b_1 and b_2 may result in two different ratings $R^s_{b_1} \neq R^s_{b_2}$ for attribute Q by the respective raters. The most common example of a subjectively measurable attribute is the notion of product or service "quality". In most transaction types, some of the attributes of interest are subjectively measurable.

In order for agent b to make use of other agents' ratings for subjectively measurable attributes as a basis for calculating agent s's reputation, it must first try to "translate" each of them into its own value system. In traditional communities we address the above issue by primarily accepting recommendations from people whom we know already. In those cases, our prior experience with these people helps us gauge their opinions and "translate" them into our value system. For example, we may know from past experience that Bill is extremely demanding and so a rating of "acceptable" on his scale would correspond to "brilliant" on our scale. As a further example, we may know that Mary and we have similar tastes in movies but not in food, so we follow her opinions on movies while we ignore her recommendations on restaurants.

Due to the much larger number of potential trading partners, in online communities it is, once again, less likely that our immediate "friends" will have had direct experiences with several of the prospects considered. It is, therefore, more likely that we will have to rely on the opinions of strangers so gauging such opinions becomes much more difficult.

Intentionally false opinions

For a number of reasons (see Section 3) agents may deliberately provide false opinions about another agent, that is, opinions, which bear no relationship to their truthful assessment of their experiences with that other agent. In contrast to subjective opinions, for which we have assumed that there can be a possibility of "translation" to somebody else's value system, false opinions are usually deliberately constructed to mislead their recipients and the only sensible way to treat them is to ignore them. In order to be able to ignore them, however, one has to first be able to identify them. Before accepting opinions, raters must, therefore, also assess the trustworthiness of other agents with respect to giving honest opinions. (Yahalom et. al., 1993) correctly pointed out that the so-called "recommender trustworthiness" of an agent is

orthogonal to its trustworthiness as a service provider. In other words, an agent can be a high-quality service provider and a very unreliable recommendation provider or vice versa.

In the rest of the section we will briefly survey the various classes of proposed online reputation reporting systems and will discuss how each of them fares in addressing the above issues.

2.2 Recommendation repositories

Recommendation repositories store and make available recommendations from a large number of community members without attempting to substantially process or qualify them. The Web is obviously very well suited for constructing such repositories. In fact, most current-generation web-based recommendation systems (message boards, opinion forums, etc.) fall into this category. A typical representative of this class of systems is the feedback mechanism of auction site eBay. Other popular auction sites, such as Yahoo and Amazon employ very similar mechanisms.

eBay encourages the buyer and seller of an eBay-mediated transaction to leave feedback for each other. Feedback consists of a numerical rating, which is can be +1 (praise), 0 (neutral) or –1 (complaint) plus a short (80 characters max.) text comment. eBay then makes the list of all submitted feedback ratings and comments accessible to any other registered user of the system. eBay does calculate some rudimentary statistics of the submitted ratings for each user (the sum of positive, neutral and negative ratings in the last 7 days, past month and 6 months) but, otherwise, it does not filter, modify or process the submitted ratings.

Recommendation repositories are a step in the right direction. They make lots of information about other agents available to interested users, but they expect users to "make sense" of those ratings themselves and draw their own conclusions. On the one hand, this is consistent with the viewpoint that the assessment of somebody's trustworthiness is an essentially subjective process (Boon and Holmes, 1991). On the other hand, however, this baseline approach does not scale very well. In situations where there are dozens or hundreds of, possibly conflicting, ratings, users need to spend considerable effort reading "between the lines" of individual ratings in order to "translate" other people's ratings to their own value system or in order to decide whether a particular rating is honest or not. What's more, in communities where most raters are complete strangers to one another, there is no concrete evidence that reliable "reading between the lines" is possible at all. In fact, as we mentioned, there is ample anecdotal evidence of people being misled by following the recommendations of false messages posted on Internet feedback forums.

2.3 Professional (specialist) rating sites

Specialist-based recommendation systems employ trusted and knowledgeable specialists who then engage in first-hand transactions with a number of service providers and then publish their "authoritative" ratings. Other users then use these ratings as a basis for forming their own assessment of someone's trustworthiness. Examples of specialist-based recommendations are movie and restaurant critics, credit-

rating agencies (Moody's) and e-commerce professional rating agencies, such as Gomez Advisors, Inc. (www.gomez.com).

The biggest advantage of specialist-based recommendation systems is that it addresses the problem of false ratings mentioned above. In most cases specialists are professionals and take great pain to build and maintain their trustworthiness as disinterested, fair sources of opinions (otherwise they will quickly find themselves out of business). On the other hand, specialist-based recommendation systems have a number of shortcomings, which become even more severe in online communities:

First, specialists can only test a relatively small number of service providers. There is time and cost involved in performing these tests and, the larger and the more volatile the population of one community, the lower the percentage of certified providers. Second, specialists must be able to successfully conceal their identity or else there is a danger that providers will provide atypically good service to the specialist for the purpose of receiving good ratings. Third, specialists are individuals with their own tastes and internal ratings scale, which do not necessarily match that of any other user of the system. Individual users of specialist ratings still need to be able to gauge a specialist's recommendation, in order to derive their own likely assessment. Last but not least, specialists typically base their ratings on a very small number of sample interactions with the service providers (often just one). This makes specialist ratings a very weak basis from which to estimate the *variability* of someone's service attributes, which is an important aspect of someone's trustworthiness, especially in dynamic, time-varying environments.

2.4 Collaborative filtering systems

Collaborative filtering techniques (Goldberg et. al., 1992; Resnick et. al., 1994; Shardanand and Maes, 1995; Billsus and Pazzani, 1998) attempt to process "raw" ratings contained in a recommendation repository in order to help raters focus their attention only on a subset of those ratings, which are most likely to be useful to them. The basic idea behind collaborative filtering is to use past ratings submitted by an agent b as a basis for locating other agents $b_1, b_2, \ldots$ whose ratings are likely to be most "useful" to agent b in order to accurately predict someone's reputation from its own subjective perspective.

There are several classes of proposed techniques:

Classification or *clustering* approaches rely on the assumption that agent communities form a relatively small set of taste clusters, with the property that ratings of agents of the same cluster for similar things are similar to each other. Therefore, if the taste cluster of an agent b can be identified, then ratings of other members of that cluster for an attribute Q of agent s can be used as statistical samples for calculating the *estimated rating* $\hat{R}_b^s$ for that same attribute from the perspective of b.

The problem of identifying the "right" taste cluster for a given agent reduces to the well-studied problem of classification/data clustering (Kaufman and Rousseeuw, 1990; Jain et, al. 1999; Gordon, 1999). In the context of collaborative filtering, the similarity of two buyers is a function of the distance of their ratings for commonly rated sellers. Collaborative filtering researchers have experimented with a variety of approaches, based on statistical similarity measures (Resnick et. al., 1994; Bresee et. al., 1998) as well as machine learning techniques (Billsus and Pazzani, 1998).

Regression approaches rely on the assumption that the ratings of an agent b_i can often be related to the ratings of another agent b_j through a linear relationship of the form

$$R^s_{b_i} = \alpha_{ij} \cdot R^s_{b_j} + \beta_{ij} \qquad \text{for all agents } s \qquad (1)$$

This assumption is motivated by the belief, widely accepted by economists (Arrow, 1963; Sen, 1986) that, even when agents have "similar" tastes, one user's internal scale is not comparable to another user's scale. According to this belief, in a given community the number of strict nearest neighbors will be very limited while the assumption of (1) opens the possibility of using the recommendations of a much larger number of agents as the basis for calculating an agent's reputation. In that case, if we can estimate the parameters α_{ij}, β_{ij} for each pair of agents, we can use formula (1) to "translate" the ratings of agents b_j to the "internal scale" of agent b_i and then treat the translated ratings as statistical samples for estimating the reputation $\hat{R}^s_{b_i}$ from the perspective of agent b_i. The problem of estimating those parameters reduces to the well-studied problem of linear regression. There is a huge literature on the topic and a lot of efficient techniques, which are applicable to this context (Malinvaud, 1966; Pindyck and Rubinfeld, 1981).

Both classification and regression approaches relate buyers to one another based on their ratings for a common set of sellers. If the universe of sellers is large enough, even active buyers may have rated a very small subset of sellers. Accordingly, classification and regression approaches may be unable to calculate estimated reputations for many seller-buyer pairs. Furthermore, the accuracy of such reputation estimates may be poor because fairly little ratings data can be used to derive them. This problem is known as *reduced coverage* and is due to the sparse nature of ratings.

Such weaknesses are prompting researchers to experiment with the use of techniques from the field of Knowledge Discovery in Databases (Fayyad et. al. 1996), which discover *latent relationships* among elements of sparse databases in the context of online reputation reporting systems. The promising use of one such technique, Singular Value Decomposition (SVD), has been reported in (Billsus and Bazzani 1998; Sarwar et. al. 2000).

3. The effects of unfair ratings in online reputation reporting systems

Of the various classes of systems surveyed in the previous section, we believe that recommendation repositories with collaborative filtering have the best potential for scalability and accuracy. Nevertheless, while these techniques address issues related to the subjective nature of ratings,th ey do not address the problem of unfair ratings. This section looks at this problem 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 a collaborative-filtering-based reputation reporting system.

To simplify the discussion, in the rest of the paper we are making the following assumptions: We assume a trading community whose participants are 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 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). Again, for the sake of simplicity we assume that $R_b^s(t)$ is a scalar quantity, although, in most transactions there are more than one critical attributes and $R_b^s(t)$ would be a vector.

We further assume the existence of an online reputation reporting 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 critical 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, which deviate from a "fair" assessment of attribute Q for a given 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. 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. This creates a *moral hazard*, which requires additional mechanisms in order to be either avoided, or detected and resolved.

In the following analysis, we assume the use of collaborative filtering techniques in order to address the issue of subjective ratings. More specifically, we 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 reputation is steady over time

The simplest scenario to analyze is one where we can assume that agent behavior, and therefore reputation, remains steady over time. That means that, collaborative filtering algorithms can take into account all ratings in their database, no matter how old.

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

$$\tau_b^s(R) = \max(R_{\min}, \min(R_{\max}, z)) \text{ where } z \sim N(\mu, \sigma) \qquad (2)$$

which in the rest of the paper will be approximated to $\tau_b^s(R) \approx N(\mu, \sigma)$. The introduction of minimum and maximum rating bounds corresponds nicely with common practice. 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 paper we will focus on the reliable estimation of the reputation mean. Given all the above assumptions, the goal of a reliable reputation reporting system should be the calculation of a fair *mean reputation estimate* (MRE) $\hat{R}_b^s$ which is equal to or very close to μ, the mean of the fair ratings distribution in the nearest neighbor set. Ideally, therefore:

$$\hat{R}^{s}_{b,fair} = \mu \tag{3}$$

On the other hand, the goal of unfair raters is to strategically introduce unfair ratings in order to *maximize* the distance between the *actual* MRE $\hat{R}^{s}_{b,actual}$ calculated by the reputation system and the fair MRE. More specifically the objective of 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 paper will calculate system behavior under the most disruptive possible unfair ratings strategy.

We will only analyze the case of ballot-stuffing since the case of bad-mouthing is symmetrical. Assume that the initial collaborative filtering step constructs a nearest neighbor set N, in which the proportion of unfair raters is δ and the proportion of fair raters is $1-\delta$. Finally, our baseline analysis in this section assumes that the actual MRE $\hat{R}^{s}_{b,actual}$ 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 recommender systems (Schafer et. al. 2001). In that case, the *actual* MRE will approximate:

$$\hat{R}^{s}_{b,actual} \cong (1-\delta)\cdot\mu + \delta\cdot\mu_{u} \tag{4}$$

where μ_{u} is the mean value of unfair ratings. The strategy, which 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.

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

$$B = \hat{R}^{s}_{b,actual} - \hat{R}^{s}_{b,fair} \tag{5}$$

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

$$B_{\max} = (1-\delta)\cdot\mu + \delta\cdot R_{\max} - \mu = \delta\cdot(R_{\max} - \mu) \tag{6}$$

Figure 1 tabulates some values of $B_{\max}$ for several different values μ and δ, in the special case where ratings range from [0,9]. For the purpose of comparing this baseline case with the "immunization mechanisms" described in Section 4, we 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, formula (6) can result in very significant inflation of a seller's MRE, especially for small μ and large δ.

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

Figure 1. Some values of maximum MRE bias when MREs are based on the mean of the ratings set. Shaded cells indicate biases above 5% of the ratings range.

Effects when reputation varies over time

This section expands our analysis by discussing some additional considerations, which arise in environments where seller behavior, and therefore reputation, may vary over time. We 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 seminal 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 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, which 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 $W=[t-\varepsilon, t]$ 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 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 $t \in W$.

Formula (6) makes 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. 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 $n << m$. Assume further that the average interarrival 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 5 and 20). 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 $\delta = n/(n+k)$. 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. 5-10) 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 collaborative-filtering-based reputation reporting systems. This requires the development of effective measures for addressing the problem. Next section proposes and analyzes several such measures.

4. Mechanisms for immunizing online reputation reporting 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 reporting systems.

4.1 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 et. al. 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 which 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 and Resnick 1999), 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 hand picked buyers but does not avoid bad-mouthing of hand picked 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 2.

Generally speaking, concealing the identities of buyers is usually easier than concealing the identities of sellers (a similar point is made in Cranor and 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, 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.

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			✔	✔

Figure 2. Effects of controlled anonymity in preventing certain classes of unfair behavior.

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.

4.2 Reducing the effect of unfair ratings using median filtering

In Section 3 we have based our calculation of reputation bias on the assumption that MREs are based on the sample mean of the nearest neighbor set. In this section we 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 order to maximize the achievable bias. To the knowledge of the author, the analysis presented in this section is novel.

The sample median $\widetilde{Y}$ 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 $\widetilde{Y}$ 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 δ=0) we have 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 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:

$$\hat{R}^s_{b,fair} \cong \mu \tag{7}$$

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 we consider the case of ballot stuffing. The case of bad-mouthing is symmetric, with the signs reversed.

Assuming that the nearest neighbor set consists of $n_f = (1-\delta)\cdot n$ fair ratings and $n_u = \delta \cdot n$ unfair ratings, where $0 \le \delta < 0.5$, 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, which 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:

$$\Pr[R_b^s \le x_q] = q \Rightarrow x_q = \sigma \cdot \Phi^{-1}(q) + \mu \tag{8}$$

where $q = \frac{k}{n_f} = \frac{n+1}{2 \cdot n_f} = \left(\frac{n+1}{n}\right) \cdot \left(\frac{1}{2 \cdot (1-\delta)}\right) \xrightarrow{n \to \infty} \frac{1}{2 \cdot (1-\delta)}$ (9)

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:

$$E[B_{\max}] = E[\hat{R}^s_{b,actual} - \hat{R}^s_{b,fair}] = \sigma \cdot \Phi^{-1}\left(\frac{1}{2 \cdot (1-\delta)}\right) \quad (10)$$

Figure 3 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. Given that we have assumed that all ratings in the nearest neighbor set correspond to users in the same taste cluster, it is expected that the standard deviation of the fair ratings will be relatively small. Therefore, we did not consider standard deviations higher than 10% of the ratings range. It is obvious that the maximum bias increases with the percentage of unfair ratings and is directly proportional to the standard deviation of the fair ratings. As before, we have highlighted maximum average biases of 5% of the rating range or more. Figure 3 clearly shows that the use of the sample median as a the 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 30-40% and a wide range of fair rating standard deviations.

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

Figure 3. Asymptotic upper bounds of average reputation bias when MREs are based on the median of the ratings set (ratings range from 0-9).

In most collaborative filtering contexts, nearest neighbor reputation estimates are based on samples with relatively small size, typically 5-15 ratings. Given that the above theoretical results are asymptotic, or "large sample" results, it is important to investigate how well they hold in the case of small sample sizes. To find that out, we have performed simulation experiments for sample sizes n=5 and n=11. The experiments resulted in remarkable correspondence between theory and practice. Details of the experimental results are reported in (Dellarocas 2000).

4.3 Using frequency filtering to eliminate unfair ratings flooding

Formulas (6) and (10) 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 can vary over time (most realistic settings), calculation of reputation should be based on recent ratings only using time discounting or a time-window approach. In those cases, Section 3 demonstrated that 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 effectively immunizing a reputation reporting 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 ratings submission frequencies $\bar{f}_b^s(t)$ for seller s, for each buyer *b* that has submitted ratings for *s* during the ratings submission frequency calculation time window W_f=[*t*-*E*, *t*]. More precisely,

   $\bar{f}_b^s(t)$ = (number of ratings submitted for *s* by *b* during W_f)/*E* (11)

2. Set the cutoff frequency $\bar{f}_{cutoff}^s(t)$ to be equal to the *k*-th order statistic of the set $F^s(t)$ where $k = (1\text{-}D) \cdot n$, *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 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 ratings submission frequencies from 100 buyers, then the cutoff frequency would be equal to the 90-th smallest frequency (the 10-th 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 in order to contain at least a few ratings from all buyers for a given seller.

Step 2: During the calculation of a 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

We will show that 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, we 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). Then, after applying frequency filtering to the nearest neighbor set of raters, in a typical time window we expect to find

- $W \cdot (1-\delta) \cdot n \cdot \int_{-\infty}^{f_{cutoff}} u \cdot \varphi(u) \cdot du$ 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:

$$\frac{\text{unfair ratings}}{\text{fair ratings}} = \frac{\delta'}{1-\delta'} = \frac{\delta}{1-\delta} \cdot \frac{\alpha \cdot f_{cutoff}}{\int_{-\infty}^{f_{cutoff}} u \cdot \varphi(u) \cdot du} = \frac{\delta}{1-\delta} \cdot I \qquad (12)$$

where $I = \frac{\alpha \cdot f_{cutoff}}{\int_{-\infty}^{f_{cutoff}} u \cdot \varphi(u) \cdot du}$ 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 ratings 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 $(1-D)\cdot n$-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:

$$(1-D)\cdot n = q\cdot(1-\delta)\cdot n + \alpha\cdot\delta\cdot n \Rightarrow \mathrm{q} = 1 - D + (\alpha-1)\cdot\frac{\delta}{1-\delta} \tag{13}$$

From this point on, the exact analysis requires some assumptions about the probability density function of fair ratings frequencies. We start by assuming 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 (12), we get the following expression of the inflation I of unfair ratings:

$$I = \frac{2\cdot S\cdot\alpha\cdot f_{cutoff}}{f_{cutoff}^2 - F_{\min}^2} \qquad \text{where } f_{cutoff} = F_{\max} - \frac{D+(\alpha-1)\cdot\delta}{1-\delta}\cdot S \tag{14}$$

After some algebraic manipulation we 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:

$$I = \frac{2\cdot(F_{\max} - \varepsilon\cdot S)}{(1-\varepsilon)\cdot(F_{\min} + F_{\max} - \varepsilon\cdot S)} \text{ where } \varepsilon = \frac{\delta}{1-\delta} \tag{15}$$

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 (12), 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 in the nearest neighbor set using time windowing and frequency filtering:

$$\delta/(1-\delta) \le \delta' \le 2\delta \tag{16}$$

Equation (16) 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 2. This concludes the proof.

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 δ<0.1, then setting D=0.1 has been experimentally proven to result in inflation ratios, which also fall within the bounds of equation (16).

A more realistic assumption about fair ratings frequencies is that they follow a lognormal distribution with mean f_0 and variance related to the frequency spread S. 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 4).

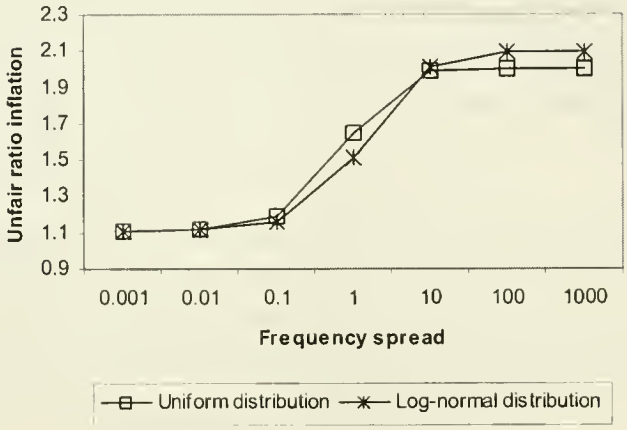

Figure 4. Maximum unfair ratings inflation factors when frequency filtering is used ($\delta = D = 0.1$).

Given that median filtering guarantees reputation biases of less than 5% of the ratings scale (e.g. less than ±0.5 points when ratings range from 1-10) for contamination ratios of up to 30-40% and frequency filtering guarantees that unfair raters cannot use flooding to inflate their presence by more than a factor of two, the combination of frequency filtering and median filtering of guarantees reputation biases of less than 5% when the ratio of unfair raters is up to 15-20% of the total buyer population for a given seller.

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, we do not believe that this property constitutes a weakness of the approach. We 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, we 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.

4.4 Issues in communities where buyer identity is not authenticated

The effectiveness of frequency filtering relies on the assumption that a given principal entity can only have one buyer agent acting on its behalf in a given marketplace. The technique is also valid in situations where principal entities have multiple buyer agents with authenticated identifiers. In that case, frequency filtering works if we consider all agents of a given principal entity 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 δ 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 $f_b^{'} \leq f_{cutoff}$, 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 in order 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* (for example, requiring that all newly registering buyers supply a valid credit card for authentication purposes) in all online communities where trust is based on reputational information.

5. Conclusions and Management Implications

We began this paper 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. We pointed out some of the challenges of producing trust in online environments and argued that online reputation reporting systems, an emerging class of information systems, hold the potential of becoming an effective, scalable, and relatively low-cost approach for achieving this goal, especially when the set of buyers and sellers is large and volatile. Understanding the proper implementation, usage and limitations of such systems (in order to eventually overcome them) is therefore important, both for the managers as well as for the participants of online communities.

This paper has contributed in this direction, first, by analyzing the reliability of current-generation reputation reporting systems in the presence of buyers who intentionally give unfair ratings to sellers and, second, by proposing and evaluating a set of "immunization mechanisms", which eliminate or significantly reduce the undesirable effects of such fraudulent behavior.

In Section 3, the motivations for submitting unfair ratings were discussed and the effects of such ratings on biasing a reputation reporting system's mean reputation estimate of a seller were analyzed. We have concluded that reputation estimation methods based on the ratings mean, which are commonly used in commercial recommender systems, are particularly vulnerable to unfair rating attacks, especially in contexts where a seller's reputation may vary over time.

Technique	Description	Effect	Prerequisites
Controlled anonymity	Market-maker conceals the true identities of buyers and sellers from one another and	Prevents bad-mouthing and/or negative discrimination	Ability to practically implement with reasonable cost

	sellers from one another and only reveals their respective reputation estimates	discrimination	reasonable cost
Median filtering	Calculation of mean reputation estimate using the median of the ratings set	Results in very robust estimations in the presence of up to 30-40% 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

Figure 5. Summary of proposed immunization techniques.

Following that analysis, a number of novel techniques for "immunizing" online reputation reporting systems against unfair ratings were proposed and analyzed in Section 4. The proposed mechanisms are summarized in Figure 5. Together, 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 1-10) when the ratio of unfair raters is up to 15-20% of the total buyer population for a given seller.

The conclusions of this paper 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 reporting 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 eliminate raters who might be attempting to flood ("spam") the system with potentially unfair ratings.

This paper 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 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 reporting systems.

References

Arrow, Kenneth (1963). *Social Choice and Individual Values*. Yale University Press.

Bakos, Y. (1997). Reducing Buyer Search Costs: Implications for Electronic Marketplaces. *Management Science*, Volume 43, 12, December 1997.

Bakos, Y. (1998). Towards Friction-Free Markets: The Emerging Role of Electronic Marketplaces on the Internet. *Communications of the ACM*, Volume 41, 8 (August 1998), pp. 35-42.

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

Boon, Susan D., & Holmes, John G. (1991). The dynamics of interpersonal trust: resolving uncertainty in the face of risk. Pages 190–211 of: Hinde, Robert A., & Groebel, Jo (eds), *Cooperation and Prosocial Behaviour*. Cambridge University Press.

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, 1998.

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

Cranor, L.F. and Resnick, P. (2000). Protocols for Automated Negotiations with Buyer Anonymity and Seller Reputations. To appear in *Netnomics.*

Dellarocas, C. (2000). The Design of Reliable Trust Management Systems for Online Trading Communities. Working Paper, available from http://ccs.mit.edu/dell/trustmgt.pdf

Dunn, John. (1984) The concept of 'trust' in the politics of John Locke. Chap. 12, pages 279–301 of: Rorty, Richard, Schneewind, J. B., & Skinner, Quentin (eds), *Philosophy in History*. Cambridge University Press.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds. (1996) *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Mass.

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

Gambetta, Diego (ed). (1990). *Trust*. Oxford: Basil Blackwell.

Gordon, A.D. (1999) *Classification*. Boca Raton: Chapman & Hall/CRC.

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.

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

Huber, Peter (1981). *Robust Statistics*. Wiley, New York.

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

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data clustering: a review. *ACM Computing Surveys*, Vol. 31, 3 (Sep. 1999), pages 264 – 323.

Johnson, D. R. and Post D. G. (1996). Law And Borders--The Rise of Law in Cyberspace. *Stanford Law Review*, Vol. 48.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

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.

Lawrence, R.J. (1980) The Lognormal Distribution of Buying Frequency Rates. *Journal of Marketing Research*. Vol. XVII, May 1980, pp. 212-226

Malinvaud, E. (1966). *Statistical Methods of Econometrics*. Paris: North Holland.

Pindyck, R. and Rubinfeld, D.L. (1981). *Econometric Models and Economic Forecasts* (2nd Edition). McGraw-Hill, New York.

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.

Resnick, P. and Varian, H.R. (1997). Recommender Systems. *Communications of the ACM,* Vol. 40 (3), pp. 56-58.

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. (2000) Application of Dimensionality Reduction in Recommender System - A Case Study. In *ACM WebKDD 2000 Web Mining for E-Commerce Workshop.*

Schmalensee, R. (1978). Advertising and Product Quality. *Journal of Political Economy,* Vol. 86, pp. 485-503.

Sen, A. (1986). Social choice theory. In *Handbook of Mathematical Economics, Volume 3.* Elsevier Science Publishers.

Schafer, J.B., Konstan, J., and Riedl, J., (2001) Electronic Commerce Recommender Applications. *Journal of Data Mining and Knowledge Discovery.* January, 2001 (expected).

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

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

Smallwood, D. and Conlisk, J. (1979). Product Quality in Markets Where Consumers Are Imperfectly Informed. *Quarterly Journal of Economics*. Vol. 93, pp. 1-23.

Wilson, Robert (1985). Reputations in Games and Markets. In *Game-Theoretic Models of Bargaining*, edited by Alvin Roth, Cambridge University Press, pp. 27-62.

Yahalom, R., Klein, B., and Beth, T. (1993). Trust Relationships in Secure Systems – A Distributed Authentication Perspective. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, 1993.

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.

Footnotes

[i] Jonathan Lebed, a 15-year-old boy was sued by the SEC for buying large blocks of inexpensive, thinly traded stocks, posting false messages promoting the stocks on Internet message boards and then dumping the stocks after prices rose, partly as a result of his messages. The boy allegedly earned more than a quarter-million dollars in less than six months and settled the lawsuit on September 20, 2000 for $285,000 (Source: Associated Press).