# The Design of Reliable Trust Management Systems for Electronic Trading Communities[†]

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

**Abstract:**

The objective of this paper is to contribute to the development of a rigorous discipline for designing trust management mechanisms in online communities. The importance of such a discipline for management science is without question: trust is a precondition for the continued existence of any market and organization in general. Furthermore, several properties of online interaction are challenging the accumulated wisdom of our communities on how to produce trust and require the development of new mechanisms and systems. The paper introduces a mathematical framework for defining trustworthiness and trust. Based on that framework, it defines the related concept of reputation and argues that reputation reporting systems is one of the most promising approaches for producing trust in online communities. It also provides a critical overview of the current state of the art in that area. Following that, it identifies a number of important ways in which unfair buyer and seller behavior can compromise the reliability of the current generation of reputation reporting systems. It then proposes and analyzes a number of novel "immunization mechanisms" for addressing those risks and explains how various parameters of an online marketplace microstructure, most notably the anonymity regime and the initial reputation policies for new sellers, can influence their effectiveness. Finally, it concludes by discussing the implications of the findings for the design of current and future online trading communities and identifies some important open issues for future research.

---

[†] **Working Paper.**

## 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).

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. Taken together, these experiences form the *reputation* of our prospective partners. Finally, once we start doing business with a partner who proves *trustworthy*, we tend to be reluctant to switch, even if we identify somebody else who claims that she can offer us better deals. The production of trust thus often acts as a switching cost.

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. 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 full 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 others trusted third parties in order to produce the equivalent of trust. This has lead to a new breed of systems, which are quickly becoming an indispensable component of every successful digital community: electronic trust management 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). 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 properties have not yet been systematically researched (Weber 2000). As Web users grow to depend on them, online trust management systems

deserve new scrutiny and the study of trust management in digital communities deserves to become a new addition to the burgeoning field of Management Science.

This paper makes several contributions in this direction: First, it introduces a mathematical framework for defining trustworthiness and trust (Section 2). Based on that framework it defines the related concept of reputation and argues that reputation reporting systems is one of the most promising approaches for producing trust in online communities (Section 3). It also provides a critical overview of the current state of the art in that area (Section 4). Following that, it identifies a number of important ways in which the reliability of the current generation of reputation reporting systems can be compromised by unfair buyers and sellers (Section 5). It then proposes a number of novel "immunization mechanisms" for addressing those risks and explains how various parameters of the marketplace microstructure, most notably the anonymity regime and the initial reputation policies for new sellers, can influence their effectiveness (Section 6). Finally, it concludes by discussing the implications of the findings for the design of current and future online trading communities and identifies some important open issues for future research (Section 7).

## 2. What is Trust

Before we can attempt to design and evaluate reliable systems whose objective is to help produce trust in online communities, it is important to understand the exact meaning of the underlying notions of trustworthiness, trust and reputation. This is especially important because these concepts, although they are so ubiquitous and pervasive in our daily lives, have been notoriously difficult to formally define.

Trust is a basic fact of human life. Despite that (or maybe because of that) there is an evident lack of coherence among researchers in the definition of trust. There is a huge body of literature on trust in fields as diverse as evolutionary biology (Bateson, 1990), sociology (Luhmann, 1979; Luhmann, 1990), social psychology (Deutsch, 1962), economics (Hart et al., 1990; Dasgupta, 1990), history (Gambetta, 1990a; Pagden, 1990), and philosophy (Lagenspetz, 1992; Hertzberg, 1988; Wittgenstein, 1977). For notable attempts to compare and integrate the various viewpoints, the interested reader is referred to (Gambetta, 1990b; Marsh 1994).

Perhaps the most popular and widely accepted definition of trust is that of Deutsch (1962), which states that:

> (a) the individual is confronted with an ambiguous path, a path that can lead to an event perceived to be beneficial ($V_+$) or to an event perceived to be harmful ($V_-$);
>
> (b) he perceives that the occurrence of $V_+$ or $V_-$ is contingent on the behavior of another person; and
>
> (c) he perceives the strength of $V_-$ to be greater than the strength of $V_+$.
>
> If he chooses to take an ambiguous path with such properties, I shall say he makes a trusting choice; if he chooses not to take the path, he makes a distrustful choice.
>
> (Deutsch, 1962, page 303)

The use of the word 'perceives' many times in this definition implies that trust is a subjective, or agent-centered notion, one in which the choices that are made are based on subjective views of the world. This is of importance in the discussions and definitions to follow.

In the rest of this section we will clarify and formalize this definition in the context of transaction-oriented agent communities:

Let *a, b, c, ...* be the universe of autonomous agents. Agents can be humans or machines. By autonomous we mean that no agent has direct control and power over the actions of another agent. For the purposes of this paper, we define a *community* of agents as a subset of the universe of agents grouped together by the fact that they engage in frequent transactions of class *T*. For example, the eBay community is the set of agents that engage in instances of the class of transactions defined as "buying and selling through the eBay website".

In the following discussion we assume, for simplicity, that all transactions are bilateral, that is, they only involve two agents. We will use the symbols *b* (buyer) and *s* (seller) to refer to the two parties of a bilateral transaction. It is important to emphasize, however, that the definitions of this section apply not only to buy-sell transactions, but also to any other type of bilateral transaction.

**Definition 1[1]:** A *critical attribute* of agent *s* from the perspective of agent *b* in the context of a transaction $t_i \in T$ is an attribute whose value affects the utility of agent *b* and is contingent upon the behavior of agent *s* in the course of transaction $t_i$.

Since critical attributes relate to an agent's individual utility, they are purely subjective and may differ even among agents engaged in transactions of the same type. For example, it is reasonable to expect a situation where the critical attribute set of an eBay seller from the perspective of eBay buyer $b_1$ is {days between payment was made and book was delivered, final price}, whereas the critical attribute set of the same seller from the perspective of a different buyer $b_2$ is {final price, book condition}, i.e. the second buyer does not care about delivery time but cares about the book condition. We will discuss the importance of this observation in Section 4.

Note, also, that critical attributes need not necessarily correspond to intrinsic attributes of agent *s*. For example, in a used car trade, the most critical attribute is the quality of the car itself. In all cases they must be contingent upon the behavior of agent *s* in the context of transaction $t_i$.

[1] For brevity, the definitions that follow will only be given from the perspective of agent *b*, with the understanding that the equivalent definitions from the perspective of the other party (agent s) are symmetric.

Depending on the nature of its domain, a critical attribute can be *continuous* or *discrete*. Price is an example of a continuous attribute. Service quality, expressed on an integer scale of 1-10 is an example of a discrete attribute. *Binary attributes* are a special case of discrete attributes, where the domain consists only of two values. The attribute "product delivered by agreed upon deadline", whose domain is the set {yes, no}, is an example of a binary attribute.

**Definition 2:** Let $X_1, X_2, ..., X_n$ be the critical attributes of agent *s* from the perspective of agent *b* in the context of a bilateral transaction $t_i \in T$ between *b* and *s*. Further, let $D_1, D_2, ..., D_n$ be their respective domain sets. The *critical rating vector* $\mathbf{R}_b^s(t_i) \in D_1 \times D_2 \times ... \times D_n$, specifies agent *b*'s subjective rating of all critical attributes of agent *s* at the end of transaction $t_i$. In a way, $\mathbf{R}_b^s(t_i)$ defines the outcome of transaction $t_i$ from the perspective of agent b.

Before we proceed to the definitions of trustworthiness and trust, it is useful to introduce here a further distinction of critical attributes that will play an important role in our later discussion of trust system reliability.

**Definition 3:** Let *C* be a community of agents where *X* is a critical attribute of agent *s* in the context of transaction class *T* from the perspective of all agents $b_i \in C$. Let $b_i, b_j \in C$ be two agents and $t_i, t_j \in T$ denote transactions of those respective agents with agent *s*. Finally, let $R_{b_i}^s(t_i), R_{b_j}^s(t_i)$ be the respective ratings of attribute *X* from the perspective of agents $b_i$ and $b_j$ at the end of transactions $t_i$ and $t_j$. We say that attribute *X* is *objectively measurable* if and only if, assuming truthful ratings, the following property holds:

$$t_i \equiv t_j \Leftrightarrow R_{b_i}^s(t_i) = R_{b_j}^s(t_i) \text{ for all agents } b_i, b_j \in C \qquad (1)$$

where the symbol "≡" denotes identity of transactions, in the sense that agents $b_i$ and $b_j$ made identical requests and agent *s* behaved in an identical manner in both. We say that a critical attribute is *subjectively measurable* if there exists at least a pair of agents $b_i, b_j \in C$ for which property (1) does not hold.

Intuitively, an attribute of an agent is objectively measurable if, a given agent behavior results in identical ratings from the perspective of all other agents who may have interacted with it. An attribute is subjectively measurable if identical behavior may result in different ratings from the perspective of different transaction partners.

"Final price" and "time of delivery" are two examples of objectively measurable attributes. On the other hand, "quality of service" and "merchandise condition" are two examples of subjectively measurable attributes.

In most agent communities, at least some of the critical attributes are subjectively measurable. As we will see in Sections 4, 5, and 6, this creates some very important complications for the construction of reliable trust management systems.

We are now ready to introduce the important notion of trustworthiness and the related notion of trust.

In a community of autonomous agents, agent *b* cannot control the behavior of agent *s*. Therefore, when considering a transaction which is sequenced in time, agent *b* is confronted with the possibility that agent *s* may behave in ways that will result in a transaction outcome with negative (or positive, but unacceptably low) utility for *b*. In order for agent *b* to be able to decide whether to proceed with the transaction, it is important that *b* has some information that will enable it to assess agent *s*'s likely behavior. We call that prior subjective assessment of *s*'s behavior the trustworthiness of *s* as perceived by agent *b*. More formally:

**Definition 4:** The *trustworthiness* $\tau_b^s(\mathbf{R}_b^s(t_i))$ of agent *s* as perceived by agent *b* in the context of a transaction $t_i \in T$ is the a priori subjective joint probability distribution function of the critical rating vector $\mathbf{R}_b^s(t_i)$ from the perspective of agent *b*. For the sake of notational simplicity, in the next of the paper trustworthiness will be denoted simply as $\tau_b^s(\mathbf{R}, t_i)$.

Armed with an assessment of another agent's trustworthiness, agent *b* is now able to reason about the transaction risks. If we assume that *b* is a rational utility-maximizing agent, *b* will only proceed with the transaction if it is sufficiently confident that its utility at the end of the transaction will be above a, subjectively defined, minimum threshold:

**Definition 5:** The minimum threshold of satisfaction $u_0$ for agent *b* in the context of a transaction $t_i \in T$ is the minimum utility that agent *b* is willing to accept at the end of the transaction in order to consider it satisfactory.

At this point we have all the ingredients necessary to define trust:

**Definition 6:** The *level of trust* $T_b^s(t_i)$ of agent *b* for agent *s* in the context of a transaction $t_i \in T$ is the a priori probability that the utility of agent *b* will meet or exceed its minimum threshold of satisfaction $u_0$ at the end of transaction $t_i$, given *b*'s perceived trustworthiness of agent *s*. Simply stated, trust is the level of confidence of agent *b* that the outcome of a transaction with another agent *s* will be satisfactory for it. More formally:

$$T_b^s(t_i) = \int_{U_b(\mathbf{R}) \geq u_0} \tau_b^s(\mathbf{R}, t_i) \cdot d\mathbf{R} \qquad (2)$$

where $U_b(\mathbf{R})$ is the utility function of agent *b*. Since trust has been defined as a probability, it ranges from [0,1].

The above definitions have a number of interesting properties, which correspond nicely with the intuitive properties of trust in our everyday life.

- Trustworthiness is subjective. Different agents *b* may have different assessments of agent *s*'s likely behavior in the same type of transactions.
- Trustworthiness is defined relative to a particular set of critical attributes. Agents can have very different trustworthiness functions for different sets of attributes. When agent *b* is considering agent *s* as a potential partner in a transaction of type *T*, it is very important that the right trustworthiness function is used. This corresponds to the intuitive notion that the same agent could be considered very trustworthy as a partner in one set of transactions and very untrustworthy in another. Example: You may trust your mechanic to fix your car but you might not trust him to teach your lectures!
- Trustworthiness is defined at a given point in time. In the general case, the trustworthiness function will vary with time, as agent *b* accumulates more information about agent s or as agent *s* genuinely modifies its behavior. In the rest of the paper we will often replace the argument $t_i$ in $\tau_b^s(\mathbf{R}, t_i)$ with *t*, denoting time, and will consider trustworthiness as a function of time.
- Trustworthiness is defined as a probability distribution, not as a single value! In the general case, the calculation of trust in formula (2) requires the knowledge of the entire trustworthiness distribution. This is an extremely important observation, given that many current-generation online trust management systems attempt to calculate a single, scalar cumulative measure of reputation and trust.

We will revisit this last observation in Section 4. In the meantime, we will discuss a number of special cases where the calculation of trust can be simplified.

Monotonic utility functions

In many communities, agent utility functions are monotonically increasing (decreasing) functions of a given critical attribute. For example, in most real-life cases, buyers' utility is a monotonically increasing function of "product quality" and a monotonically decreasing function of "total price". Let us assume, for further simplification that this attribute is the only critical attribute in a given transaction class. Under those assumptions, formula (2) can be rewritten as

$$T_b^s(t_i) = \int_{R_0}^{+\infty} \tau_b^s(R, t_i) \cdot dR \qquad \text{if } U_b(R) \text{ is monotonically increasing} \qquad (3a)$$

and

$$T_b^s(t_i) = \int_{-\infty}^{R_0} \tau_b^s(R, t_i) \cdot dR \qquad \text{if } U_b(R) \text{ is monotonically decreasing} \qquad (3b)$$

where $R_0 = U_b^{-1}(u_0)$

Gaussian trustworthiness functions

Let us assume, as above, that $U_b(R)$ is a monotonically increasing (decreasing) function of R. If, in addition, $\tau_b^s(R,t_i)$ approximates a normal (Gaussian) distribution $N(\mu,\sigma)$, then equations (3) can be further simplified. By applying the well-known properties of normal probability distributions to equations (3) we get:

$$T_b^s(t_i) = 1 - \int_{-\infty}^{R_0} \tau_b^s(R,t_i) \cdot dR = \Phi(\frac{\mu - R_0}{\sigma}) \qquad \text{if } U_b(R) \text{ is monotonically increasing} \quad (4a)$$

$$T_b^s(t_i) = \int_{-\infty}^{R_0} \tau_b^s(R,t_i) \cdot dR = \Phi(\frac{R_0 - \mu}{\sigma}) \qquad \text{if } U_b(R) \text{ is monotonically decreasing} \quad (4b)$$

where $\Phi(x)$ is the standard normal CDF.

One important observation is that in the special case of Gaussian trustworthiness functions, the calculation of trust levels only requires assessments of the mean and standard deviation of the trustworthiness function, i.e. two scalar values, as opposed to the entire distribution.

Relative trust

In several cases, agent *b* has already decided to engage in a transaction of type *T* and is confronted with the problem of selecting the "best" trading partner from between a pair of eligible prospects $s_1$ and $s_2$[2]. Let us assume that agent *b* always selects the agent it trusts more. In other words, it calculates its level of trust for each agent and selects prospect $s_1$ if $T_b^{s_1}(t_i) > T_b^{s_2}(t_i)$ and prospect $s_2$ otherwise. In those cases, what matters most are not the *absolute* trust levels but rather, their *relative* magnitudes.

In the special case of monotonically increasing utilities and Gaussian trustworthiness functions $\tau_b^{s_1}(R,t_i) \sim N(\mu_1,\sigma_1)$ and $\tau_b^{s_2}(R,t_i) \sim N(\mu_2,\sigma_2)$ from equation (4a) we get:

$$T_b^{s_1}(t_i) > T_b^{s_2}(t_i) \Leftrightarrow \Phi(\frac{\mu_1 - R_0}{\sigma_1}) > \Phi(\frac{\mu_2 - R_0}{\sigma_2}) \Leftrightarrow \frac{\mu_1 - R_0}{\sigma_1} > \frac{\mu_2 - R_0}{\sigma_2} \qquad (5)$$

[2] The analysis can easily be generalized in the case of *n* prospects.

If, in addition to all the above assumptions we further assume that $\sigma_1 \approx \sigma_2$ then the above formula can be further simplified and gives:

$$T_b^{s_1}(t_i) > T_b^{s_2}(t_i) \Leftrightarrow \mu_1 > \mu_2 \quad (6)$$

In this very special case, relative trust can be based on the knowledge of the mean of the trustworthiness distribution only.

If $U_b(R)$ is a monotonically decreasing function of R then the results are similar with the direction of inequalities reversed.

Binary attributes

A last notable special case is the case where the critical attribute set consists of a single binary critical attribute *X*, which can take one of two values $V_+$ (beneficial outcome) and $V_-$ (harmful outcome) with probabilities *p* and *(1-p)* respectively. Then:

$$T_b^s(t_i) = p \quad (7)$$

This is another special case where a single scalar value (*p*) is sufficient in order to estimate trust levels. It is also the case that corresponds to the Deutsch's definition of trust mentioned at the beginning of the chapter.

By connecting back to the definition of trust that we started with, we have come full circle. Based on this section's definitions, the next section discusses the role of communities in helping agents assess the various quantities needed in order to estimate trust levels.

## 3. Mechanisms of trust production

From formula (2) we can infer that the production of trust has three prerequisites:

- an agent should know its utility function
- an agent should set a minimum threshold of satisfaction relative to a transaction
- an agent should estimate the trustworthiness of its prospective trading partners

Of the three elements of trust computation the first is usually internal and private to an agent. The second is either internal or the explicit result of a negotiation process that precedes a transaction. The last one, trustworthiness, is the trickiest one to assess. According to the preceding discussion, it, too, is the result of a subjective process, which combines external information with an agent's general trusting disposition (Boon and Holmes, 1991). The role of external information is very important in this case however. A community's success in producing trust among its members depends on its ability to

help agents construct reliable assessments of the trustworthiness of other community members.

There are three basic ways that communities go about doing this:

- norms backed up by institutional guarantees
- indirect cues
- reputational information

*Norms and institutional guarantees* attempt to reduce the uncertainty on the behavior of other agents by prescribing specific allowed behavioral ranges (which, usually correspond to satisfactory outcomes $V_+$ for the majority of transaction types and society members) and by providing institutions, which prevent deviations or make them highly unlikely because of quick detection and effective sanctions (Parsons, 1964). Institutional guarantees reduce the problem of trusting individual agents to that of trusting the institutions: if one trusts that institutions will do their job, there is less need to assess the trustworthiness of every single individual agent. In the case of binary attributes, the situation can be described mathematically

$$\tau_b^s(R) = p(R = V_+|I) \cdot p(I) + p(R = V_+ \mid \neg I) \cdot (1 - p(I)) \qquad (8)$$

where *I* denotes the assumption that institutions function effectively. In the context of the above equation, institutions promise that $p(R = V_+|I) = 1$, which gives:

$$\tau_b^s(R) = p(I) + p(R = V_+ \mid \neg I) \cdot (1 - p(I)) \cong p(I) \qquad \text{when } p(I) \text{ gets close to 1} \qquad (9)$$

The use of institutional guarantees has a number of important shortcomings when applied to digital communities. Assessing the effectiveness of institutions is not always trivial, especially for newcomers to a given digital community. Even more important, however, institutions are less effective in online communities than they are in more traditional "brick and mortar" communities. There are two main reasons for this: first, the most successful online communities span the boundaries of several territorially-based jurisdictions and their members are governed by different, and often conflicting, legal systems (Johnson and Post, 1996). Second, in many online communities it is relatively easy to change identities (Friedman and Resnick, 1999). Although the evolution of Internet law may may change this in the future, the overall effect is that institutional guarantees are generally weaker in online environments and thus, there is more need to accurately assess the trustworthiness of other potential trading partners before engaging in a transaction with them.

*Indirect cues* are attributes of an agent, which we have associated with certain likely behaviors based on our experience, intuition and training. For example, most people tend to perceive a well-dressed, well-mannered businessperson as being trustworthier than an unkempt, unruly one. Formally, the translation of cues into trustworthiness assessments involves conditional subjective probability distributions of the form *p(behavior|cue)* that we or our community has accumulated over long time and passed on to us through

tradition and formal education. Many people give very high value to those cues and consider them important factors of their decision-making. However, it is exactly this kind of cues that are usually absent in online communities. Also, these cues are useless in the emerging class of multi-agent markets where the traders are software programs (Maes et. al., 1999; Dellarocas and Klein, 2000b).

*Reputational information* is information about or observations of an agent's past behavior on similar situations, aggregated and distributed by means of word-of-mouth or through trusted third parties, such as credit rating agencies, consumer reports, etc. Reputational information can help agents construct estimates on another agent's trustworthiness under the assumption that agents have an underlying distribution of behavior, which is relatively stable over time[3]. Then, information about past behavior can be used as statistical samples from which to construct an estimate of the trustworthiness distribution for the purpose of predicting future behavior.

We have deliberately used the term reputational information in order to distinguish it from the notion of reputation itself. 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". Wilson's definition of reputation is very close to our definition of trustworthiness in the special case where trustworthiness is primarily assessed on the basis on past behavior data (as opposed to institutional guarantees or indirect cues). This leads to the following definition:

**Definition 7:** The *reputation* of an agent *s* as perceived by agent *b* in the context of transaction $t_i \in T$ with critical attribute set **R** is its trustworthiness distribution $\tau_b^s(\mathbf{R}, t_i)$ in the special case where the estimation of $\tau_b^s(\mathbf{R}, t_i)$ is based on information about the past behavior of *s* in transactions of class *T*.

In the rest of the paper we will often use the terms reputation and trustworthiness interchangeably. Reputational information, as distinct from reputation, is the past behavior data used by an agent in order to derive another agent's trustworthiness/reputation. This information can come in the form of isolated observations ("last time I transacted with X, I wasn't very happy with the service I got") or in the form of cumulative trustworthiness/reputation assessments from the perspective of other agents ("agent Y can provide very good service, but its quality has not been consistent in the past few months"). In fact, one of the most interesting design dimensions in online reputation reporting systems is the decision about whether reputational information should be provided in the form of "raw" ratings or cumulative measures (see Section 4).

[3] By relatively stable we mean that, even when this distribution is changing over time, its rate of change is slow relative to the rate of observations.

Reputation has been the object of study of the social sciences for a long time (Rogerson, 1983; 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. Reputation is most effective when buyers and sellers persist in a trading community for a long time (Wilson, 1985). This persistence is often tricky to guarantee in online communities.

The relative ease with which computers can capture, store and process huge amounts of information about past transactions, makes 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. The next section will survey the current state of the art in online reputation reporting mechanisms.

## 4. Reputation reporting mechanisms in online communities

Having interacted with someone in the past is, of course, the most reliable source of information about that agent's reputation because then the observations used to estimate someone's reputation are direct samples of the subjective variable *R* whose distribution we seek to estimate. 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 focus of this section is to provide a brief, critical survey of the most important issues and categories of these systems.

### 4.1 Design issues in online reputation reporting systems

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

- *Consensus on critical attributes.* Reputation is defined relative to a specific set of critical attributes. The same agent may have very different reputation for different attributes. When accumulating other agents opinions, it is, therefore, extremely important to ascertain that all opinions refer to the same critical attributes. This requires careful research from the part of the rating mechanism designers, in order to identify the complete set of critical attributes for a given community, as well as careful communication of those attributes to community members.

- *Subjectively measurable attributes.* For subjectively measurable critical attributes (see Section 2, Definition 3) the same behavior of agent *s* vis-à-vis two different agents $b_1$ and $b_2$ may result in two different ratings $R^s_{b_1} \neq R^s_{b_2}$. In order for agent *b* to make use of these conflicting ratings 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.

- *False opinions.* For a number of reasons 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 our framework, this fact is a simple corollary of the definition of trustworthiness relative to a specific set of critical attributes.

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 addresses the above issues.

## 4.2 Recommendation repositories

Recommendation repositories store and make available recommendations from a large number of community members without attempting to substantially process or qualify them. This reduces the search costs of interested agents, who can then find a large number of recommendations in a single place.

The Web is obviously very well suited for constructing such repositories. In fact, most current-generation web-based recommendation systems 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 viewpoint is consistent with the fact that the assessment of trustworthiness and trust is a subjective process. 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. Finally, rating repositories rely at this stage probably more on textual comments than they do on numerical ratings. This makes them unsuitable for use in software agent communities where the buying and selling is performed by automated software programs.

A lot of these shortcomings do not exist in cases where ratings are based on objectively measurable attributes (e.g. on-time records of airlines, number of lost baggage incidents per month etc.). In those cases, simple rating repositories can be very effective.

4.3 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 restaurant critics (Zagat's), 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. 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 *probability distribution* of someone's service attributes which is what we have defined as trustworthiness/reputation.

4.4 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 a user $b_0$ as a basis for locating other users $b_1, b_2, \ldots$ whose ratings are likely to be most "useful" to user $b_0$ in order to accurately predict someone's reputation from its own subjective perspective.

There are several related 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 very similar to each other. Each taste cluster $C_k$ then has the property that:

$$t_i \equiv t_j \Leftrightarrow R^s_{b_i}(t_i) \approx R^s_{b_j}(t_i) \text{ for all agents } b_i, b_j \in C_k \quad (10)$$

Therefore, if the taste cluster of a user $b_0$ can be identified, then ratings of other members of that cluster can be readily used as statistical samples for estimating the subjective probability distribution of $R^s_{b_0}(t_i)$ from the perspective of $b_0$.

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). 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

$$R^s_{b_i}(t_i) = \alpha_{ij} \cdot R^s_{b_j}(t_i) + \beta_{ij} \qquad \text{for all agents } s \quad (11)$$

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 (11) opens the possibility of using the recommendations of a much larger number of agents as the basis for calculating an agent's trustworthiness. In that case, if we can estimate the parameters $\alpha_{ij}, \beta_{ij}$ for each pair of agents, we can use formula (11) 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 of the distribution of $R^s_{b_i}(t_i)$ 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).

### 4.5 The pitfalls of calculating cumulative measures of reputation

Most collaborative filtering systems do not simply compute similarities or regression coefficients between user ratings. They go further and compute cumulative measures, which are intended to be interpreted as "estimates of reputation of user *s*".

The most commonly encountered cumulative measures have the form of a weighted average of individual ratings. Different proposed approaches are using different ways to calculate the weights. For example, Resnick et. al (1994) propose the use of the Pearson correlation coefficient, while Bresee et. al. (1998) proposed the use of vector similarity measures as well as several heuristically derived adjustments to weights.

The computation of cumulative measures of reputation is useful because it reduces the computational burden on the side of the agents. However, we believe that, in the current generation of systems, it is often a misleading and dangerous input for building trust.

First of all, as pointed out by Billsus and Pazzani (1998), most of the currently proposed cumulative measures are not supported by a sound theory of reputation and trust. For example, a weighted average of individual ratings where the weights are correlation coefficients does not have a direct correspondence to any of the trust-related concepts introduced in this paper.

Furthermore, in Section 2 we believe that we have made a strong case for the fact that the calculation of trust levels, whether absolute or relative, requires the knowledge of the entire trustworthiness/reputation *distribution*. A single scalar cumulative measure is usually not sufficient for describing a distribution except in very special cases, such as the distribution of binary attributes, or normal distributions where the variance is considered to be roughly the same throughout the agent population.

### 4.6 Summary

This section has surveyed a number of different classes of current-generation reputation reporting mechanisms in online communities. Of the various classes of systems surveyed, our conclusion is that collaborative filtering approaches have the best potential for scalability and accuracy. Nevertheless, further research is required in order for such systems to become reliable and trustworthy enough. We have identified a number of problems that still need to be addressed:

- achieving consensus on the critical attributes for which ratings are stored
- deriving theoretically sound cumulative measures of reputation
- coping with the possibility of intentionally false ratings

The rest of the paper focuses on the last problem.

## 5. The effects of unfair ratings in online reputation reporting systems

The preceding discussion on trust building in online communities has identified two important challenges for the effective use of reputational information as a basis for trust production: First the subjective nature of ratings on many commonly used critical attributes and the need to translate somebody else's ratings to our own "value system". Second, the possibility that some of the raters may provide unfair (intentionally false) ratings. Although collaborative filtering researchers have looked at the first problem, to

date the second problem has received very little attention. Our goal in this section 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, a buyer *b* contracts a seller *s* for the provision of a service. Upon conclusion of the transaction, *b* provides a numerical rating $R_b^s(t_i)$, 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 $R_b^s(t_i)$ is a scalar quantity, although, as we noted in the previous sections, in most transactions there are more than one critical attributes and $R_b^s(t_i)$ 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 for sellers *s* upon request of a prospective buyer *b*.

In settings where the critical attribute *Q* for which ratings are provided is not objectively measurable, there exist four scenarios where buyers and/or sellers can intentionally try to "rig the system", resulting in biased reputation estimates, which do not reflect the true expected distribution 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 with those of *b* on other commonly rated sellers[4]. 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:

$$\tau_b^s(R) = \max(R_{\min}, \min(R_{\max}, z)) \text{ where } z \sim N(\mu, \sigma) \qquad (12)$$

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

[4] In the case of regression-based systems the nearest neighbor set of buyer $b_i$ would be computed on the basis of the "translated" ratings $R'^s_{b_j} = \alpha_{ij} \cdot R^s_{b_j} + \beta_{ij}$

assumption of normality. In the near future we intend to empirically verify this assumption by analyzing some existing ratings database.

In Section 2 we have shown that, in the special case where $\tau_b^s(R) \approx N(\mu,\sigma)$, the calculation of trust levels only requires the estimation of the two scalar parameters $\mu,\sigma$ of the reputation distribution. In this paper we will focus on the reliable estimation of the reputation mean. The reliable estimation of the reputation standard deviation is the topic of a forthcoming paper.

Given all the above assumptions, the goal of a reliable reputation reporting system should be the calculation of a fair *mean reputation estimate* (MRE) which is equal to or very close to $\mu$, the mean of the fair ratings distribution in the nearest neighbor set. Ideally, therefore:

$$\hat{R}^s_{b,fair} = \mu \tag{13}$$

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*, which includes $(1-\delta)\cdot 100\%$ fair raters and $\delta\cdot 100\%$ unfair raters. Finally, assume 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*. In that case, the *actual* MRE will approximate:

$$\hat{R}^s_{b,actual} \cong (1-\delta)\cdot\mu + \delta\cdot\mu_u \tag{14}$$

where $\mu_u$ 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:

$$B = \hat{R}^s_{b,actual} - \hat{R}^s_{b,fair} \tag{15}$$

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

$$B_{\max} = (1-\delta)\cdot\mu + \delta\cdot R_{\max} - \mu = \delta\cdot(R_{\max} - \mu) \quad (16)$$

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]. We have generally considered biases above 5% of the ratings range (i.e. biases greater than 0.5 points on ratings which range from 0-10) to be unacceptable. As can be seen, formula (16) can result in very significant inflation of a seller's MRE, especially for small μ and large δ.

*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 (16) 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 $\lambda$ 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)[5]. 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.

## 6. 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.

The handling of any kind of harmful *exceptions,* that is, deviations from desirable or normal behavior, fundamentally involves two classes of mechanisms: *avoidance* mechanisms, which proactively try to prevent this behavior from occurring at all and *recovery* mechanisms, which detect occurrences of this behavior and attempt to reduce its harmful consequences for the interested parties and the community at large (Dellarocas

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

and Klein 2000a). Based on this distinction, we are classifying our proposed mechanisms into avoidance mechanisms and recovery mechanisms.

6.1 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 the market-maker.

We start by introducing some concepts that are needed in order to characterize the anonymity regime of a marketplace. First, we assume that agents, whether human or machine, are exactly that. That is, they participate in communities and engage in transactions on behalf of some real-life *principal entity P*. *P* can be an individual or an organization. What is important here is that *P* has a fixed and persistent real-world existence and identity, which we assume is impossible to change.

An *identifier I* is a piece of information which is publicly known within an online community and which is used in order to refer to an agent in the context that community. At the minimum, an identifier should provide a way for information to reach an agent, as well as for an agent to send information to other agents. IP addresses and email addresses are examples of identifiers with this property.

The *authentication regime* of an online community specifies the degree of certainty with which community activity performed using identifier *I* can be linked to a unique principal entity *P* by some participant of the community. Perfectly authenticated communities guarantee that if anybody uses identifier *I* to send or receive information within a community, that that somebody can only be principal *P*. The design of effective authentication regimes and processes is an important research topic within the field of Computer Security (Hutt et. al., 1995). Online communities form a spectrum with regards to their authentication regimes, ranging from very well authenticated to non-authenticated. In non-authenticated communities, principals are basically free to create multiple identifiers or to discontinue using them, effectively disappearing and reappearing at will.

The *transparency regime* of an online community specifies which members of the community have the right to apply or access the results of a community authentication process. Otherwise stated, it specifies who is allowed to know the true identity of the principal *P* related to an identifier *I*. At one end of the spectrum, every member of the community is given that right. In that case, we have fully transparent communities. At the other extreme, there are communities where the only entity who has access to the true identity of community members is the party who controls the infrastructure resources of the community.

Below we argue that, under the assumption that the market maker can be trusted, full transparency incurs more dangers than a scheme where identities are authenticated but carefully concealed.

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. 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.

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.

### 6.2 Reducing the effect of unfair ratings using median filtering

In Section 5 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[6].

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 $\tilde{Y}$ 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 $\delta=0$) 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

[6] The sample median turned out to be the best out of several different candidate "robust" estimators of MRE tested by the author. The detailed comparisons among the various measures are outside the scope of this work and are described in a forthcoming paper.

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:

$$\hat{R}^{s}_{b,fair} \cong \mu \tag{17}$$

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:

$$\Pr[R^{s}_{b} \le x_q] = q \Rightarrow x_q = \sigma \cdot \Phi^{-1}(q) + \mu \tag{18}$$

where

$$q = \frac{k}{n_f} = \frac{n+1}{2\cdot n_f} = \left(\frac{n+1}{n}\right)\cdot\left(\frac{1}{2\cdot(1-\delta)}\right) \xrightarrow[n\to\infty]{} \frac{1}{2\cdot(1-\delta)} \tag{19}$$

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[7] reputation bias achievable by $\delta \cdot 100\%$ unfair ratings when fair ratings are drawn from a normal distribution

[7] We are assuming here that unfair raters have knowledge of $\mu$ and $\sigma$ but do *not* have knowledge of the *exact* individual values of fair ratings, which, in a time-windowed system, are rapidly changing anyway. Therefore their objective is to maximize the *expected value* of the MRE bias.

$N(\mu,\sigma)$ and unfair ratings follow the most disruptive possible unfair ratings distribution, is given by:

$$E[B_{\max}] = E[\hat{R}^{s}_{b,actual} - \hat{R}^{s}_{b,fair}] = \sigma \cdot \Phi^{-1}(\frac{1}{2 \cdot (1-\delta)}) \quad (20)$$

Figure 3 shows some of the values of $E[B_{\max}]$ for various values of $\delta$ and $\sigma$ in the special case where ratings range from 0 to 9[8]. 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 assuming that a maximum average bias of 5% or less of the rating range is acceptable. Given these assumptions, the use of the sample median as a the basis of calculating MREs proves to be an acceptable and robust estimate for high levels of contamination and a wide range of standard deviations.

In most real-life 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. Our experiments simulated a community where fair ratings are integers from 0-9 drawn from a distribution given by:

$$\tau^{s}_{b}(R) = \max(0, \min(9, \lfloor z + 0.5 \rfloor)) \text{ where } z \sim N(\mu,\sigma) \quad (21)$$

The pseudocode of the experiments is listed in Figure 4. The results, for sample sizes *n*=5 and *n*=11 and for several values of n and $\sigma$, are tabulated in Figure 5 and constitute a small sample reality-check of the asymptotic values of Figure 3. The correspondence between theory and practice is remarkable for both tested small sample sizes.

6.3 Using frequency filtering to eliminate unfair ratings flooding

Formulas (16) and (20) confirm the intuitive fact that the reputation bias due to unfair ratings increases with the ratio $\delta$ of unfair raters in a given sample. In settings where a seller's critical 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 5 demonstrated that by "flooding" the system with ratings, a relatively small number of unfair raters can manage to increase the ratio $\delta$ of unfair ratings in any given time window above 50% and completely compromise the reliability of the system.

[8] 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.

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* (22)

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[9].

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 frequency filtering cutoff frequency.

[9] One suggestion is to set $E(t) = 3/\min(\bar{f}_b^s(t-1), \text{for all } b \in F^s(t-1))$, i.e. set the width of the current time window equal to three times the largest buyer-specific ratings inter-arrival period in the last time window.

*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 $\alpha$ 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:

$$\frac{\text{unfair ratings}}{\text{fair ratings}} = \frac{\delta'}{1-\delta'} = \frac{\delta}{1-\delta} \cdot \frac{\alpha \cdot f_{cutoff}}{\int_{-\infty}^{f_{cutoff}} u \cdot \varphi(u) \cdot du} = \frac{\delta}{1-\delta} \cdot I \qquad (23)$$

where *I* 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 design frequency filtering so as to minimize *I*.

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

$$(1-D) \cdot n = q \cdot (1-\delta) \cdot n + \alpha \cdot \delta \cdot n \Rightarrow \mathrm{q} = 1 - (D + \alpha - 1) \cdot \frac{\delta}{1-\delta} \qquad (24)$$

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 (23), we get the following expression of the inflation *I* of unfair ratings:

$$I = \frac{2 \cdot S \cdot \alpha \cdot f_{cutoff}}{f_{cutoff}^2 - F_{\min}^2} \qquad (25a)$$

where $f_{cutoff} = F_{\max} - \frac{D + (\alpha - 1) \cdot \delta}{1 - \delta} \cdot S$ (25b)

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 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:

$$I = \frac{2 \cdot (F_{\max} - \varepsilon \cdot S)}{(1 - \varepsilon) \cdot (F_{\min} + F_{\max} - \varepsilon \cdot S)} \text{ where } \varepsilon = \frac{\delta}{1 - \delta} \quad (26)$$

The above expression for the unfair/fair ratings inflation depends on the spread S of fair ratings frequencies. At the limiting cases we get:

$$\lim_{S \to 0} I = \frac{1}{1 - \varepsilon} \text{ and } \lim_{S \to \infty} I = \frac{2}{1 - \varepsilon} \quad (27)$$

By substituting the above limiting values of *I* in equation (23), 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:

$$\delta / (1 - \delta) \le \delta' \le 2\delta \quad (28)$$

Equation (28) 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.

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.

6.4 The effects of initial reputation policies in the presence of unfair raters

Both the reputation attacks, as well as the immunization techniques described in the previous section, were analyzed in a steady-state scenario, where it was assumed that, at the time of attack, agent *s* had already accumulated a fair reputation from a number of fair buyers who have had the opportunity to transact with it. This section will consider what happens if we assume that attacks commence *immediately* upon the appearance of a new seller in the marketplace. Our analysis has some important implications for the optimal initial reputation policy of a community in the presence of possible unfair raters.

Friedman and Resnick (1999) have proposed two alternative initial reputation policies: (1) assign *minimum* reputation to all newcomers and let them gradually earn their "real" reputation by offering good service, or (2) require newcomers to pay entry fees as a way of *purchasing* units of reputation from the market-maker. Purchased reputation units are lost if a seller decides to disappear or change its identity.

Friedman and Resnick were mostly concerned with the problem of sellers who can easily "disappear" from a marketplace after offering poor service and then "reappear" under a new identity. They have shown that, in the absence of unfair raters, both policies are effective in incurring "reappearance costs" which discourage such seller behavior. In this section, our concern is to analyze the effects of each initial reputation policy in the presence of unfair raters. More specifically, we will analyze how each policy affects the effectiveness of median and frequency filtering if we assume that reputation attacks commence immediately upon the appearance of a new seller in the marketplace. As we will show, the assignment of minimum initial reputation is less robust than the assignment of average initial reputation (with payment of entry fees if identities cannot be perfectly authenticated).

Policy 1: Newcomers are assigned minimum initial reputation

*Ballot stuffing*

When the goal of unfair raters is to inflate a seller's reputation, upon appearance of a new seller *s,* colluding buyers will immediately begin engaging in (possibly fake) transactions with it in order to submit very positive ratings. Since, at the very beginning, all ratings will be unfair ratings ($\delta = 1$), the application of median filtering and frequency filtering will have no effect and the MRE will be highly inflated. This will induce fair buyers to start transacting with *s* as well. The first few buyers who will transact with *s* will receive inferior service quality to that implied by the seller's reputation and will, therefore, be very unhappy with the community's reputation reporting accuracy. When enough fair sellers have interacted with *s* at least once, then the situation converges to the steady-state case and the filtering approaches discussed above begin to be effective.

The above scenario is equivalent to an initial reputation policy, which assigns the *maximum* possible reputation to newcomers for free. From (Friedman and Resnick, 1999) we know that this is not an optimal policy. Furthermore, in settings where sellers can

easily change identity, they can "disappear" from the community before their reputation converges to its "fair" value, reappear with a new identity and start the above process all over again ad infinitum.

*Bad-mouthing*

If bad-mouthing agents immediately attack a newcomer with minimum initial reputation, its reputation will remain at minimum levels. This is likely to discourage fair agents from engaging in transactions with the new seller. Therefore, the ratio of unfair raters transacting with this seller is likely to remain high, its reputation will remain unfairly low and the seller will most likely then soon go out of business. Again, because $\delta$ is large, the filtering techniques described in the previous sections cannot help in this scenario.

Clearly, this initial reputation policy is not satisfactory in the presence of unfair raters.

Policy 2: Entry fees and average initial reputation using artificial ratings

We denote the entrance time of a new seller *s* by $t_0$. Let us propose the following concrete initial reputation strategy:

1. All new sellers are required to pay an entry fee[10].
2. All new sellers are assigned an initial reputation $\overline{R}$ equal to the average reputation of all sellers in the community at time $t_0$. This is done as follows: the system generates *k* artificial ratings of value $\overline{R}$ and places them in uniformly distributed random points within the time window [$t_0$-ε, $t_0$]. The number of artificial ratings is given by $k = \bar{f}(t_0) \cdot \varepsilon$, where $\bar{f}(t_0)$ is the average ratings submission frequency for *any* seller by *any* buyer: $\bar{f}(t_0) = Average(\bar{f}^{s_i}(t_0), \text{all } s_i)$ where $\bar{f}^{s_i}(t_0)$ is calculated using (23).

The above initial reputation policy essentially sets $\bar{f}^{s}(t_0) = \bar{f}(t_0)$. This makes frequency filtering immediately effective and therefore limits the fraction of unfair ratings that "make it" into the final calculation of the MRE. This, in turn, also immediately makes median filtering effective. The net result is that the MRE of *s* will quickly converge to its fair levels, relatively unaffected by the presence of unfair raters. Furthermore, the existence of an entry fee incurs a cost, which prevents sellers whose fair reputation is below $\overline{R}$ from attempting to quickly disappear and reappear into the marketplace before their MRE reaches its fair level.

[10] To alleviate the concerns raised by Friedman and Resnick about the adverse effects of an entry fee, the fee could be considered as a bond, or security deposit, to be refunded to a seller if, upon exit from the community, its reputation is equal to or greater to its initially assigned levels.

The result of this section is that, in the presence of unfair raters, an initial reputation policy which charges entry fees and assigns average initial reputation to all newcomers via the generation of artificial ratings is preferable to a minimum initial reputation entry strategy.

6.5 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 $\delta$ 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 $\delta$ will be high, even in the presence of frequency filtering, unfair raters can still manage to severely contaminate a seller's fair reputation.

Evidently, 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.

## 7. Summary and Conclusions

The objective of this paper is to contribute to the development of a rigorous discipline for designing trust management mechanisms in online communities. The importance of such a discipline is without question: trust is a precondition for the continued existence of any market and civilized community in general. Furthermore, several properties of online interaction are challenging the accumulated wisdom of our communities on how to produce trust and require the development of new mechanisms and systems.

In order to study the production of trust, we thought it necessary to first precisely define what trust means. For that reason, in Section 2, we have introduced a mathematical framework for defining trust in the context of a transaction-oriented community. We have found that the most central notion in trust production is that of *trustworthiness,* which we have defined as an agent's subjective assessment of the probability distribution of another agent's future behavior in the context of a class of transactions.

From a community perspective, the production of trust, therefore, requires the existence of mechanisms that help agents accurately assess the trustworthiness of other agents. "Brick and mortar" communities employ a variety of mechanisms for this purpose, including the establishment of behavioral norms backed up by institutional guarantees, the use of indirect cues and the dissemination of past behavior data as a way of predicting an agent's future behavior.

Of those mechanism classes, institutional guarantees and reliance on indirect cues are less appropriate at this stage of evolution of online communities. On the other hand, the ability of online communities to store and process complete and accurate information about all transactions mediated by them, makes them ideally suited for using of past behavior data (reputational information) as the basis for building trust.

We have defined reputation to be someone's trustworthiness, in the special case where it is assessed on the basis of past behavior data. A number of researchers and practitioners have already built the first generation of online reputation reporting systems. However, most of these systems have not been built on the basis of a rigorous framework of trust and trust building. In Section 4 we have surveyed the current state-of-the-art in reputation reporting systems from the perspective of the framework introduced in this paper. We have concluded that, in order to build reliable online reputation reporting issues, a number of issues need to be satisfactorily addressed. These issues include:

- the need to build consensus among the community on the attributes about which reputational information is being collected and reported
- the need to help users of reputational information draw accurate conclusions in the case of attributes, which are not objectively measurable
- the need to develop cumulative measures of reputation which are backed up by theory
- the need to address the possibility of unfair ratings about other agents

Of the four issues, the first requires careful system design and communication with community members. The second is being addressed by the set of techniques commonly known as collaborative filtering. This paper has focused on the third, and particularly on the fourth issue.

We have remarked that a lot of the cumulative measures of reputation proposed by other researchers are not based on rigorous definitions of trust and trustworthiness. In our model, the production of trust requires the assessment of someone's entire trustworthiness probability distribution. In the important special case where we can assume normally distributed trustworthiness, we have shown that the production of trust requires estimates of the mean and standard deviation of that distribution only.

In Section 5, we have discussed the motivations for submitting unfair ratings and have analyzed their effects on biasing a reputation reporting system's estimate of the mean of someone's trustworthiness. We have concluded that severe distortions are possible, especially in situations where estimation of reputation is based on recently submitted ratings only.

One of the central contributions of this paper is the proposal and analysis of a number of novel techniques for "immunizing" online reputation reporting systems against unfair ratings. The proposed mechanisms are summarized in Figure 6.

The analysis of the proposed techniques has resulted in a number of important guidelines for the design of current and future electronic marketplaces:

- 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. Electronic marketplaces and B2B hubs can consider adding this function into the set of services they provide to their members.
- The initial reputation policy for new sellers is crucial in the presence of unfair raters. A minimum initial reputation policy makes newcomers vulnerable to bad-mouthing attacks. On the other hand a policy, which involves entry fees (or security deposits) and an average initial reputation works well in conjunction with the proposed immunization techniques.

This paper has simply scratched the surface of an important set of problems. The calculation of robust estimates of reputation standard deviation and the development of "immunization" techniques that avoid unfair ratings "flooding" in non-authenticated communities are just two of the issues left open by this paper. It is our hope that the framework and techniques proposed by this work will provide a useful basis that will stimulate further research in the important and exciting area of online trust management 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.

Bateson, Patrick. (1990) The Biological Evolution of Cooperation and Trust. Chap. 2, pages 14–30 of: Gambetta, Diego (ed), *Trust*. Blackwell.

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.*

Dasgupta, Partha. (1990). Trust as a Commodity. Chap. 4, pages 49–72 of: Gambetta, Diego (ed), *Trust*. Blackwell.

Dellarocas, C. and Klein, M. (2000a). A Knowledge-Based Approach for Handling Exceptions in Business Processes. *Information Technology and Management*, Vol.1, 3, pp. 155-169

Dellarocas, C., Klein, M. and Rodriguez-Aguilar, J.A. (2000b). An exception-handling architecture for open electronic marketplaces of contract net software agents. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, Minneapolis, MN, October 17-20, 2000.

Deutsch, Morton. (1962). Cooperation and Trust: Some Theoretical Notes. In: Jones, M. R. (ed), *Nebraska Symposium on Motivation*. Nebraska University Press.

Deutsch, Morton. (1973) *The Resolution of Conflict.* New Haven and London: Yale University Press.

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.

Friedman, E.J. and Resnick, P. (1999) The Social Cost of Cheap Pseudonyms. Working paper[11]. An earlier version was presented at the *Telecommunications Policy Research Conference*, Washington, DC, October 1998.

Gambetta, Diego. (1990a). Mafia: The Price of Distrust. Chap.10, pages 158–176 of: Gambetta, Diego (ed), *Trust.* Blackwell.

Gambetta, Diego (ed). (1990b). *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.

Hart, David M., Anderson, Scott D., & Cohen, Paul R. (1990). Envelopes as a Vehicle for Improving the Efficiency of Plan Execution. Tech. Rept. COINS 90-21. University of Massachusetts at Amherst, Department of Computing and Information Science.

Hertzberg, Lars. (1988). On the Attitude of Trust. *Inquiry*, 31(3), 307–322.

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.

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

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.

Lagenspetz, Olli. (1992). Legitimacy and Trust. *Philosophical Investigations*, 15(1), 1–21.

Luhmann, Niklas. (1979). *Trust and Power*. Chichester: Wiley.

Luhmann, Niklas. (1990). Familiarity, Confidence, Trust: Problems and Alternatives. Chap. 6, pages 94–107 of: Gambetta, Diego (ed), *Trust.* Blackwell.

Maes, P., Guttman, R.H. and Moukas A. (1999) Agents that Buy and Sell. *Communications of the ACM,* Vol. 42 (3), March 1999, pp. 81-91.

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

Marsh, Steven. (1994). *Formalizing Trust as a Computational Concept*. Ph.D. Thesis, University of Stirling, United Kingdom.

Pagden, Anthony. (1990). The Destruction of Trust and its Consequences in the Case of Eighteenth Century Naples. Chap. 8, pages 127–142 of: Gambetta, Diego (ed), *Trust.* Blackwell.

Parsons, T. (1964). *The Social System*. The Free Press.

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.

Rogerson, William P. (1983). Reputation and Product Quality. *The Bell Journal of Economics*, Vol. 14, 2, pp. 508-516.

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.

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.

Weber, Thomas E. (2000) To Build Virtual Trust, Web Sites Develop "Reputation Managers". *The Wall Street Journal*. Monday, July 17, 2000, page B1.

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

Wittgenstein, Ludwig. (1977). *On Certainty — Uber Gewissheit.* Basil Blackwell, Oxford.

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.

| Percentage of unfair ratings | Fair Mean Reputation Estimate ($R_{min}$=0, $R_{max}$=9) | | | | |
|---|---|---|---|---|---|
| | **0** | **2** | **4** | **6** | **8** |
| **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 unacceptably high biases.

| **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.

| *Percentage of unfair ratings* | *Standard Deviation of Fair Ratings* | | | |
|---|---|---|---|---|
| | **0.25** | **0.50** | **0.75** | **1.00** |
| **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). Shaded cells indicate unacceptably high biases.

To calculate the maximum average reputation bias achievable by $n_u$ unfair raters in a sample of size $n = n_f + n_u$, where fair ratings have standard deviation $\sigma$:

1. Calculate the set $U$ of all possible unfair rating strategies. $U$ is the set of all different ways in which $n_u$ integer values can be distributed between 0 and 9.
2. For each unfair rating strategy $U_i \in U$
   a. For *each* possible $\mu$ =0,1,…,9 generate 100,000 random sets $F_j$ of $n_f$ fair ratings drawn from (21)
   b. Calculate the reputation bias of the total ratings set $U_i \cup F_j$ based on the sample median approach:
      $$B_{ij} = Median(U_i \cup F_j) - Median(F_j)$$
   c. Calculate the average reputation bias $\overline{B}_i = Average(B_{ij})$ achievable by unfair ratings distribution $U_i$ over all 100,000 random sets $F_j$ of fair ratings.
3. Calculate the maximum average reputation bias $B_{\max} = Max(\overline{B}_i)$ over all possible unfair rating strategies $U_i \in U$

Figure 4. Pseudocode of the experimental procedure used to test the small sample median-based maximum average MRE bias behavior.

***n=5***

**Asymptotic**

| *Number and percentage of unfair ratings* | | *Standard Deviation of Fair Ratings* | | | |
|---|---|---|---|---|---|
| | | **0.25** | **0.50** | **0.75** | **1.00** |
| 1 | **20%** | 0.08 | 0.16 | 0.24 | 0.32 |
| 2 | **40%** | 0.24 | 0.48 | 0.73 | 0.97 |

**Experimental**

| *Number and percentage of unfair ratings* | | *Standard Deviation of Fair Ratings* | | | |
|---|---|---|---|---|---|
| | | **0.25** | **0.50** | **0.75** | **1.00** |
| 1 | **20%** | 0.00 | 0.11 | 0.21 | 0.30 |
| 2 | **40%** | 0.07 | 0.41 | 0.66 | 0.85 |

***n=11***

**Asymptotic**

| *Number and percentage of unfair ratings* | | *Standard Deviation of Fair Ratings* | | | |
|---|---|---|---|---|---|
| | | **0.25** | **0.50** | **0.75** | **1.00** |
| 1 | **9%** | 0.03 | 0.06 | 0.09 | 0.13 |
| 2 | **18%** | 0.07 | 0.14 | 0.21 | 0.28 |
| 3 | **27%** | 0.12 | 0.24 | 0.37 | 0.49 |
| 4 | **36%** | 0.20 | 0.40 | 0.59 | 0.79 |
| 5 | **45%** | 0.35 | 0.69 | 1.04 | 1.38 |

**Experimental**

| *Number and percentage of unfair ratings* | | *Standard Deviation of Fair Ratings* | | | |
|---|---|---|---|---|---|
| | | **0.25** | **0.50** | **0.75** | **1.00** |
| 1 | **9%** | 0.00 | 0.01 | 0.07 | 0.11 |
| 2 | **18%** | 0.00 | 0.04 | 0.19 | 0.27 |
| 3 | **27%** | 0.00 | 0.12 | 0.34 | 0.47 |
| 4 | **36%** | 0.01 | 0.31 | 0.53 | 0.76 |
| 5 | **45%** | 0.13 | 0.66 | 0.96 | 1.27 |

Figure 5. Comparison between asymptotic and experimentally derived maximum average median-based MRE bias for rating sample sizes $n$=5 and $n$=11. Shaded cells indicate unacceptably high biases.

| Technique | Description | Effect | Prerequisites |
|---|---|---|---|
| **Controlled anonymity** | Market-maker conceals the true identities of buyers and sellers from one another | Prevents bad-mouthing and negative discrimination | Ability to practically implement with reasonable cost |
| **Median filtering** | Calculation of mean reputation estimate using the median of the ratings set | Results in robust estimations in the presence of high percentages 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 6. Summary of proposed immunization techniques.