Filtering Out Unfair Ratings in Bayesian Reputation Systems *

Andrew Whitby[1], Audun Jøsang[1] and Jadwiga Indulska[2]

[1]Distributed Systems Technology Centre †
Brisbane, Australia
me@andrewwhitby.id.au, ajosang@dstc.edu.au

[2]School of Information Technology and Electrical Engineering
University of Queensland, Australia
jaga@itee.uq.edu.au

Abstract

The quality of a reputation system depends on the integrity of the ratings it receives as input. A fundamental problem is that a rater can rate an agent more positively or more negatively than the real experience with the agent would dictate. When ratings are provided by agents outside the control of the relying party, it is a priori impossible to know when a rater provides such unfair ratings. However, it is often the case that unfair ratings have a different statistical pattern than fair ratings. This paper uses that idea, and describes a statistical filtering technique for excluding unfair ratings, and illustrates its effectiveness through simulations.

1 Introduction

Reputation systems represent a promising method for fostering trust amongst strangers in online environments. The basic idea is to let parties rate each other's performance during interactions. The aggregated ratings about a given party are used to derive a reputation score, which can assist other parties in deciding whether or not to transact with that party in the future.

This is different from trust referral systems, where agents exchange general recommendations about other parties, such as described by Jøsang *et al.* (2005) [7] and Yolum & Singh (2003) [13]. In reputation systems, agents only express their experience from particular transactions in the form of ratings, and it is the aggregation of ratings that is important.

Without such mechanisms, where strangers are interacting, for example, in an e-commerce setting, the temptation to act deceptively for immediate gain could be more appealing than cooperation. In the physical world, capturing and distributing ratings can be costly. In comparison, the Internet is extremely efficient, and reputation scores can be published for everyone to see. The effect of reputation systems is to provide an incentive for honest behaviour, and also to deter dishonest parties from participating.

Finding ways to avoid or reduce the influence of unfairly positive or unfairly negative ratings is a fundamental problem in reputation systems where ratings from others are taken into account. This is because the relying party can not control the sincerity of the ratings when they are provided by agents outside its control. Effective protection against unfair ratings is a basic requirement in order for a reputation system to be robust.

* Appears in the Proceedings of the Workshop on Trust in Agent Societies, at the Autonomous Agents and Multi Agent Systems Conference (AAMAS2004), New York, July 2004.

† The work reported in this paper has been funded in part by the Cooperative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Industry, Science & Resources).

The methods of avoiding bias from unfair ratings can broadly be grouped into two categories, *endogenous* and *exogenous*, as described below.

1. **Endogenous Discounting of Unfair Ratings**
   This category covers methods that exclude or give low weight to presumed unfair ratings based on analysing and comparing the rating values themselves. The assumption is that unfair ratings can be recognised by their statistical properties only. Proposals in this category include Dellarocas (2000) [4] and Chen & Singh (2001) [2].

2. **Exogenous Discounting of Unfair Ratings**
   This category covers methods where external factors, such as the reputation of the rater, are used to determine the weight given to ratings. The assumption is that raters with low reputation are likely to give unfair ratings and vice versa. This method requires that raters' reputation can be determined by including external evidence other than the ratings themselves. Proposals in this category include Buchegger & Le Boudec (2003) [1], Cornelli *et al.* (2002) [3], Ekström & Björnson (2002) [5] and Yu & Singh (2003) [14].

Our proposal falls in the first category, where we only analyse the statistical properties of the ratings. The scheme that most closely resembles ours is the one by Dellarocas (2000) [4] which is targeted at detecting and excluding ratings that are unfairly positive. In the static case where it is assumed that agent behaviour never changes, Dellarocas' scheme first uses collaborative filtering to determine a neighbourhood group of raters whose ratings over many subjects are similar. Then the ratings over a particular subject s are divided into two groups by the Macnaughton-Smith *et al.* (1964) cluster filtering algorithm [10], in order to filter out the ratings that are most likely to be unfairly positive. Under the assumption that the fair and unfair ratings follow the same distribution type (e.g. normal), but with different parameters, this technique typically reduces the unfair bias by more than 80%.

In the dynamic case where it is assumed that reputation scores change due to changing agent behaviour, Dellarocas applied the same technique to neighbourhood ratings from a sliding window in the frequency domain. Simulation results indicate that variations in seller quality over time does not reduce the effectiveness of cluster filtering.

Because Dellarocas' scheme only assumes unfairly positive ratings, the effect of cluster filtering in the absence of positive unfair ratings is to produce small negative bias on the reputation score, because some positive fair ratings are filtered out. The study says nothing about how cluster filtering would work in the presence of both unfairly positive and negative ratings.

In this paper we propose a filtering technique that applies to both unfairly positive and unfairly negative ratings in Bayesian reputation systems. We use the reputation system described in Jøsang & Ismail (2002) [9], and integrate the filtering method to the reputation systems simulator described in Jøsang *et al.* (2003) [8].

The assumption behind our filtering method is that ratings provided by different raters on a given agent will follow more or less the same probability distribution. When an agent changes its behaviour, it is assumed that all honest raters who interact with that agent will change their ratings accordingly. By comparing the overall reputation score of a given agent with the the probability distribution of the ratings on that agent from each rater, our scheme dynamically determines an upper and lower threshold for which raters should be judged unfair and thereby excluded. Instead of using a sliding time window, our scheme uses a longevity factor that gradually reduces the weight of received ratings as a function of their age.

2 The Reputation System

The simulations were carried out using an enhanced version of the Bayesian reputation system described in Jøsang *et al.* [8, 9]. This reputation system provides a general mechanism for incorporating reputation services into e-commerce applications.

2.1 Bayesian Reputation Systems

Bayesian systems are based on computing reputation scores by statistical updating of beta probability density functions (PDF). The *a posteriori* (i.e. the updated) reputation score is computed by combining the *a priori* (i.e. previous) reputation score with the new rating

[6, 9, 11, 12]. The reputation score can be represented in the form of the beta PDF parameter tuple $(\alpha,\ \beta)$ (where α and β represent the amount of positive and negative ratings respectively), or in the form of the probability expectation value of the beta PDF, and optionally accompanied by the variance or a confidence parameter. The advantage of Bayesian systems is that they provide a statistically sound basis for computing reputation scores, and the only disadvantage that it might be too complex for average persons to understand.

The beta-family of distributions is a continuous family of distribution functions indexed by the two parameters α and β. The beta PDF denoted by $\mathrm{beta}(p \mid \alpha, \beta)$ can be expressed using the gamma function Γ as:

$$\mathrm{beta}(p \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (1)$$

where $0 \leq p \leq 1$ and $\alpha, \beta > 0$, with the restriction that the probability variable $p \neq 0$ if $\alpha < 1$, and $p \neq 1$ if $\beta < 1$. The probability expectation value of the beta distribution is given by:

$$\mathrm{E}(p) = \alpha/(\alpha+\beta). \qquad (2)$$

When nothing is known, the *a priori* distribution is the uniform beta PDF with $\alpha = 1$ and $\beta = 1$ illustrated in Figure 1. Then after observing r positive and s negative outcomes, the *a posteriori* distribution is the beta PDF with $\alpha = r + 1$ and $\beta = s + 1$. For example the beta PDF after observing 7 positive and 1 negative outcomes is illustrated in Figure 2.

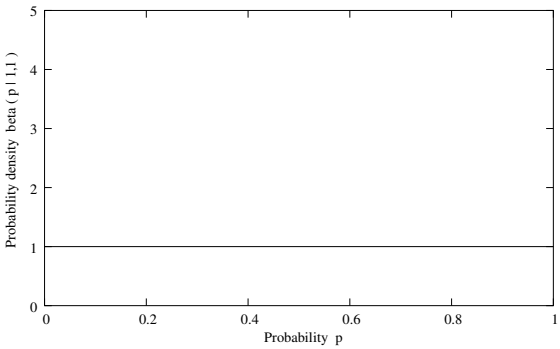

Figure 1: Uniform PDF $\mathrm{beta}(p \mid 1, 1)$

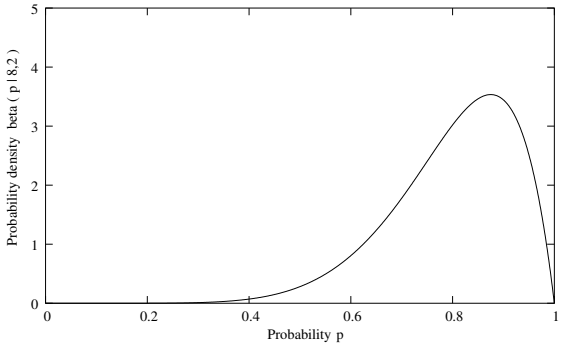

Figure 2: Example PDF $\mathrm{beta}(p \mid 8, 2)$

A PDF of this type expresses the uncertain probability that a process will produce positive outcomes during future observations. The probability expectation value of Figure 2 according to Eq.(2) is $\mathrm{E}(p) = 0.8$. This can be interpreted as saying that the relative frequency of a positive outcome in the future is somewhat uncertain, and that the most likely value is 0.8.

The variable p is a probability variable, so that for a given p the probability density $\mathrm{beta}(p \mid \alpha, \beta)$ represents second order probability. The first-order variable p represents the probability of an event, whereas the density $\mathrm{beta}(p \mid \alpha, \beta)$ represents the probability that the first-order variable has a specific value. Since the first-order variable p is continuous, the second-order probability $\mathrm{beta}(p \mid \alpha, \beta)$ for any given value of $p \in [0, 1]$ is vanishingly small and therefore meaningless as such. It is only meaningful to compute $\int_{p_1}^{p_2} \mathrm{beta}(p \mid \alpha, \beta)$ for a given interval $[p_1, p_2]$, or simply to compute the expectation value of p. The most natural is to define the reputation score as a function of the expectation value. This provides a sound mathematical basis for combining ratings and for expressing reputation scores.

2.2 Collecting Ratings

The general reputation system allows for an agent to rate another agent, both positively and negatively, by arbitrary amounts, for a single transaction. This rating takes the form of a vector:

$$\rho = \left[\begin{array}{c} r \\ s \end{array} \right], \text{ where } r \geq 0 \text{ and } s \geq 0. \quad (3)$$

A simple binary rating system can then be implemented by using $\rho^+ = [1, 0]$ for a satisfactory transaction and $\rho^- = [0, 1]$ for an unsatisfactory transaction [8].

A particular rating can be denoted as:

$$\rho^X_{Z,t_R} \quad (4)$$

which can be read as X's rating of Z at time t_R. Where they are not relevant, these super- and subscripts will be omitted.

2.3 Aging Ratings

Agents (and in particular human agents) may change their behaviour over time, so it is desirable to give greater weight to more recent ratings. This can be achieved by introducing a longevity factor λ, which controls the rate at which old ratings are 'forgotten':

$$\rho^{X,t}_{Z,t_R} = \lambda^{t-t_R} \rho^X_{Z,t_R} \quad (5)$$

where $0 \leq \lambda \leq 1$, t_R is the time at which the rating was collected and t is the current time.

2.4 Aggregating Ratings

Ratings may be aggregated by simple addition of the components (vector addition).

For each pair of agents (X, Z), an aggregate rating $\rho^t(X, Z)$ can be calculated that reflects X's overall opinion of Z at time t:

$$\rho^t(X, Z) = \sum \rho^{X,t}_{Z,t_R}, \text{ where } t_R \leq t\ . \quad (6)$$

Also, Z's aggregate rating by all agents in a particular set S can be calculated:

$$\rho^t(Z) = \sum_{X \in S} \rho^t(X, Z). \quad (7)$$

In particular, the aggregate rating for Z, taking into account ratings by the entire agent community C, can be calculated:

$$\rho^t(Z) = \sum_{X \in C} \rho^t(X, Z). \quad (8)$$

2.5 The Reputation Score

Once aggregated ratings for a particular agent are known, it is possible to calculate the reputation probability distribution for that agent. This is expressed as:

$$\text{beta}(\rho^t(Z)) = \text{beta}(p \mid r + 1,\ s + 1), \quad (9)$$

where

$$\rho^t(Z) = \left[\begin{array}{c} r \\ s \end{array} \right].$$

However probability distributions, while informative, cannot be easily interpreted by users. A simpler point estimate of an agent's reputation is provided by $\text{E}[\ \text{beta}(\rho^t(Z))\ ]$, the expected value of the distribution. This provides a score in the range $[0, 1]$, which can be scaled to any range (including, for example, '0% reliable - 100% reliable').

Definition 1 (Reputation Score) *Let $\rho^t(Z) = [r, s]'$ represent target Z's aggregate ratings at time t. Then the function $R^t(Z)$ defined by:*

$$R^t(Z) = \text{E}[\ \text{beta}(\rho^t(Z))\ ] = \frac{r + 1}{r + s + 2} \quad (10)$$

is called Z's reputation score at time t.

The reputation score $R^t(Z)$ can be interpreted like a probability measure as an indication of how a particular agent is expected to behave in future transactions.

3 The Problem of Unfair Ratings

A reputation system aggregates the ratings of many individuals. In most cases these ratings are subjective and unverifiable. This leads to a problem: while the reputation system may keep the seller-ratees honest, what guarantee is there that the buyer-raters will be honest in their assessment of sellers through ratings? In most cases there is no guarantee at all, and the method described below is aimed at finding methods to detect and filter out dishonest ratings.

Dellarocas [4] identifies two categories of unfair ratings:

- unfairly positive ratings (which he calls 'ballot stuffing'), and
- unfairly negative ratings (which he calls 'badmouthing').

The risk of unfair ratings is highest when they can be used to manipulate the reputation system to an agent's advantage. For instance, a buyer may collude with a seller to badmouth the seller's competitors, resulting in gains to the seller.

By careful construction of the reputation system, this risk can be diminished, by either eliminating the incentive to lie, or increasing the cost of doing so. An example of the former technique is to hide the identity of agents from each other [4]. In a sufficiently large market, this eliminates the possibility of badmouthing another agent (although ballot stuffing is still possible, as the colluding seller and buyer may communicate their identities by some other means). An example of the latter technique is to allow buyers to rate sellers only after a transaction is conducted, and charge a small amount for every transaction, thus raising the cost of ballot-stuffing or badmouthing. This approach is taken by eBay.

In many systems, however, these techniques cannot be used. In these cases, it is desirable to be able to automatically detect and exclude (via statistical or machine-learning methods) unfair ratings.

Binary rating systems present a particularly difficult problem. In a system with continuous rating values (e.g. a scale from 1–10), a very low rating (e.g. 1) may be said to be unusual if a ratee's historical ratings have been high (e.g. 9.5 average) and can therefore be rejected. In contrast, in a system with binary ratings a negative rating (i.e. 0) cannot be rejected simply because the ratee has received mostly positive ratings (i.e. 1).

In general, this problem appears insoluble. However if raters and ratees interact repeatedly, it becomes possible to compare long-run average ratings and reject raters who, over some period, have rated significantly differently from the average. This is the basis for our filtering algorithm.

3.1 An Iterated Filtering Algorithm Based on the Beta Distribution

The filtering algorithm is executed whenever an agent Z's reputation score must be recalculated. It assumes the existence of cumulative rating vectors $\rho^t(X, Z)$ for each rater X in the community. The basic principle is to verify that the score $R(Z)$ of an agent Z falls between the q quantile (lower) and $(1-q)$ quantile (upper) of $\rho^t(X, Z)$ for each rater X. Whenever that is not the case for a give rater, that rater is considered unfair, and its ratings are excluded.

By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.01 (or 1%) quantile is the point at which 1% percent of the data fall below and 99% fall above that value. As en example, the uniform PDF of Figure 1 will always have a q quantile of q, whereas the PDF of Figure 2 has a 0.01 quantile of 0.456 and a 0.99 quantile of 0.983 as illustrated in Figure 3.

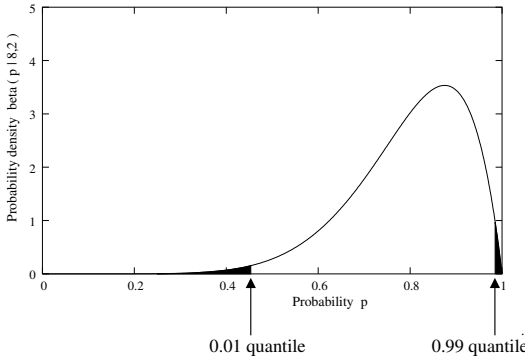

Figure 3: 1% and 99% quantiles of $\mathrm{beta}(p \mid 8, 2)$

To be meaningful, the quantile parameter q must be in the range $[0.0,\ 0.5]$. The smaller the value of q, the less false positives (i.e. the less honest raters are excluded) and the more false negatives (i.e. the more unfair raters are included). The larger the value of q, the less false negatives (i.e. the less unfair raters are included) and the more false positives (i.e. the more fair raters are excluded).

In the simulations, the quantile parameter was set to $q = 0.01$, which provided a good balance. In practice it means that if an agent's overall reputation score is outside the 1% quantile and the 99% quantile of the ratings provided by a rater X, then that rater's ratings will be discarded from the set of ratings that contribute to the agent's score. The pseudocode for the filtering algorithm used in the simulations is given below.

C is the set of all raters
F is the set of all assumed fair raters
Z is the target agent

$F = C$

do loop

$\rho^t(Z) := \sum_{X \in F} \rho^t(X, Z)$

$R^t(Z) := \mathrm{E}(\rho^t(Z))$

for each rater R in F

$f := \mathrm{beta}(\rho^t(R, Z))$
$l := q$ quantile of f
$u := (1-q)$ quantile of f

if $l > R^t(Z)$ or $u < R^t(Z)$
$F := F \setminus \{R\}$

loop

until F does not change

return $R^t(Z)$

This algorithm is flexible and robust, because it is based on variable distributions rather than a fixed tolerance. If the spread of ratings from all raters is wide, then it will tend not to reject individual raters. On the other hand, if a ratee is rated consistently highly by all raters (e.g. 95% positive, 5% negative ratings) except for one (e.g. 50% positive, 50% negative ratings), then the exceptional rater will be rejected. The algorithm's sensitivity can be increased or decreased by modifying the q parameter.

4 Description of Market Simulation

4.1 Market Structure

The market consists of N_s sellers and N_b buyers who trade in a single commodity. The simulation is divided into sessions, and each session is divided into rounds. During a single round, each buyer can trade at most once, with a seller of its choice, while each seller may trade as many times as its productive capacity allows. At the end of each session, the buyers and sellers adapt their behaviour based on their success in the last round (and previous rounds).

4.2 Reputation System

The Bayesian reputation system described above, with aging and filtering of unfair ratings, is used in the simulations. However, unlike in the description above, individual ratings are not stored. Instead, an aggregate rating, $\rho(X, Z)$, is stored for each rater/ratee pair. New ratings are accumulated into this aggregate. This reduces the storage overhead of the system, but means that instead of aging individual ratings when calculating the reputation score, aggregate ratings must be aged discretely at regular intervals:

$$\rho(X, Z) := \lambda \cdot \rho(X, Z). \tag{11}$$

The resulting implementation is an arbitrarily accurate approximation of the system described above. In practice the rating aging takes place as the end of each round.

4.3 Seller Behaviour

Seller behaviour is determined by two parameters, a selling price $\mathrm{Prc}(S)$ and an honesty level $\mathrm{Hst}(S)$. In addition, all sellers share the same unit cost of production, $\mathrm{Cst}_{\mathrm{prod}}$ and per-round productive capacity, $\mathrm{Cap}_{\mathrm{prod}}$.

Once the seller has committed to a transaction, it will either ship the item (with a probability equal to $\mathrm{Hst}(S)$) or it will not ship the item (with a probability equal to $1-\mathrm{Hst}(S)$). If the seller does ship the item, it will receive the price $\mathrm{Prc}(S)$ from the buyer, and will incur the unit cost Cst. If the seller does not ship the item, it will still receive Prc from the buyer, but will incur no costs.

Thus the seller's transaction gain for an honest transaction will be:

$$g^{\mathrm{T}}(S) = \mathrm{Prc}(S) - \mathrm{Cst}\ , \tag{12}$$

while for a dishonest transaction it will be:

$$g^{\mathrm{T}}(S) = \mathrm{Prc}(S)\ . \tag{13}$$

4.3.1 Adaptation

At the end of each session, each seller S compares the gains from the session, $g_t^T(S)$, with the gains from the

previous session, $g^T_{t-1}(S)$. Let t_{best} be the session in which the gains were highest; then the price parameter for the next session is chosen randomly as follows:

$$\text{Prc}_{t+1}(S) = \begin{array}{ll} \text{Prc}_{t_{\text{best}}}(S) + \Delta_{\text{Prc}}\ , & (p = \text{Prob}^{+}_{\text{Prc}}) \\ \text{Prc}_{t_{\text{best}}}(S) - \Delta_{\text{Prc}}\ , & (p = \text{Prob}^{-}_{\text{Prc}}) \\ \text{Prc}_{t_{\text{best}}}(S)\ , & (p = \text{Prob}^{=}_{\text{Prc}}) \end{array} \tag{14}$$

where $\text{Prob}^{+}_{\text{Prc}} + \text{Prob}^{-}_{\text{Prc}} + \text{Prob}^{=}_{\text{Prc}} = 1$, and subject to $\text{Prc}_{t+1}(S) \geq 0$. Similarly, the honesty parameter for the next session is chosen randomly as follows:

$$\text{Hst}_{t+1}(S) = \begin{array}{ll} \text{Hst}_{t_{\text{best}}}(S) + \Delta_{\text{Hst}}\ , & (p = \text{Prob}^{+}_{\text{Hst}}) \\ \text{Hst}_{t_{\text{best}}}(S) - \Delta_{\text{Hst}}\ , & (p = \text{Prob}^{-}_{\text{Hst}}) \\ \text{Hst}_{t_{\text{best}}}(S)\ , & (p = \text{Prob}^{=}_{\text{Hst}}) \end{array} \tag{15}$$

where $\text{Prob}^{+}_{\text{Hst}} + \text{Prob}^{-}_{\text{Hst}} + \text{Prob}^{=}_{\text{Hst}} = 1$, and subject to $0 \leq \text{Hst}_{t+1}(S) \leq 1$. In the simulations, all the 6 probabilities above were set to 0.33. $\Delta_{\text{Prc}} = 1$ and $\Delta_{\text{Hst}} = 0.02$. In addition, a set of heuristics overrule the default adaptation under certain circumstances:

1. If the seller did not conduct any transactions during session t, its selling price $\text{Prc}(S)$ is decreased by Δ_{Prc}, and its honesty $\text{Hst}(S)$ is increased by Δ_{Hst}.

2. If the seller made a loss during session t, its selling price is increased by Δ_{Prc}.

4.4 Buyer Behaviour

Normal buyer behaviour is determined by a single parameter, risk aversion $\text{Rsk}(B)$ where $\text{Rsk}(B) = 1$ is neutral. In addition, all buyers receive the same utility from an item, Val and have the same propensity to search Srch.

A buyer attempts to purchase one item per round. For a transaction with a given seller S, the buyer expects to receive:

$$\begin{aligned} g^{\text{T}}_{\text{Exp}}(B,S) &= (\text{Val} - \text{Prc}(S)) \cdot R(S)^{\text{Rsk}(B)} \\ &\quad -\text{Prc}(S) \cdot (1 - R(S)^{\text{Rsk}(B)}) \\ &= \text{Val} \cdot R(S)^{\text{Rsk}(B)} - \text{Prc}(S)\ . \end{aligned} \tag{16}$$

This expected gain reflects two possible outcomes. If the seller ships the item (expected probability $R(S)$, weighted to the power of $\text{Rsk}(B)$ to reflect the buyer's attitude to risk) then the buyer will make a profit of $g^{\text{T}}_{\text{ships}} = \text{Val} - \text{Prc}(S)$. If the seller does not ship the item then the buyer will make a loss of $g^{\text{T}}_{\text{ships}} = -\text{Prc}(S)$.

The effort to which a buyer goes in order to maximise this gain is determined by the propensity to search parameter. The buyer will search – that is, find the seller that maximises $g^{\text{T}}_{\text{Exp}}(B,S)$ – with a probability equal to Srch. The rest of the time the buyer will transact with the first seller whose price, $\text{Prc}(S)$ does not exceed the buyer's utility, Val.

Once the transaction is complete, the buyer will rate the seller. If the seller failed to ship the item, the buyer will report a rating of $(0,1)$. If the seller did ship the item, the buyer will report a rating $(1,0)$.

4.4.1 Adaptation

At the end of each session, each buyer B compares the gains from the session, $g^S_t(B)$ with the gains from the previous session, $g^S_{t-1}(B)$. Let t_{best} be the session in which the gains were highest; then the risk attitude parameter for the next session is chosen randomly as follows:

$$\text{Rsk}_{t+1}(B) = \begin{array}{ll} \text{Rsk}_{t_{\text{best}}}(B) + \Delta_{\text{Rsk}}\ , & (p = \text{Prob}^{+}_{\text{Rsk}}) \\ \text{Rsk}_{t_{\text{best}}}(B) - \Delta_{\text{Rsk}}\ , & (p = \text{Prob}^{-}_{\text{Rsk}}) \\ \text{Rsk}_{t_{\text{best}}}(B)\ , & (p = \text{Prob}^{=}_{\text{Rsk}}) \end{array} \tag{17}$$

where $\text{Prob}^{+}_{\text{Rsk}} + \text{Prob}^{-}_{\text{Rsk}} + \text{Prob}^{=}_{\text{Rsk}} = 1$, and subject to $0 \leq \text{Rsk}_{t+1}(B) \leq 1$. The probability values were set to $\text{Prob}^{+}_{\text{Rsk}} = 0.50$ (to make the buyers risk averse), $\text{Prob}^{-}_{\text{Rsk}} = 0.25$ and $\text{Prob}^{=}_{\text{Rsk}} = 0.25$. $\Delta_{\text{Rsk}} = 0.01$.

In addition, a set of heuristics overrule the default adaptation under certain circumstances:

1. If a buyer made a loss in session t, then its risk aversion $\text{Rsk}(B)$ is increased by Δ_{Rsk}.

2. Otherwise, if, in any round of session t, a buyer did not purchase an item, its risk aversion $\text{Rsk}(B)$ is decreased by Δ_{Rsk}.

4.4.2 Unfair Buyer Behaviour

Most of the buyers in a market will be normal buyers, however some may be 'unfair' in that, when trading with chosen sellers, they misreport their ratings.

Unfair buyers exist in three states. In the fair state, an unfair buyer behaves exactly like a fair buyer. In the ballot-stuffing state, a buyer will rate positively regardless of whether the seller shipped the item, with a probability of $\mathrm{Prob}_{\mathrm{unfair}}$, and rate fairly the rest of the time. In the badmouthing state, a buyer will rate negatively regardless of whether the seller shipped the item, with a probability of $\mathrm{Prob}_{\mathrm{unfair}}$, and rate fairly the rest of the time.

All unfair buyers follow the same sequence of states. The state sequence used is given for each for of the simulations below.

5 Simulation Results

The simulations were conducted to assess the effectiveness of iterated filtering in a number of different scenarios.

All simulations were conducted over 1200 sessions. The first 200 sessions of each simulation have not been included in graphs, since the market usually takes some time to stabilise as the reputations change to reflect the seller's behaviour.

The results obtained by simulating markets for many combinations of these parameters demonstrate the strengths and weaknesses of the filtering technique.

5.1 The Effect of Unfair Ratings

Before considering the effect of filtering, it is important to consider the effect of unfair ratings on an otherwise stable market. A relatively honest market was created with the following key parameters:

Parameter	Value
number of buyers	50
fair	50
unfair	0
number of sellers	6
initial seller honesty	0.9

Figure 4 shows the reputation score and true honesty of a single seller over 1000 sessions. Without unfair buyers, the reputation system is very effective, as reputation tracks honesty with only a slight occasional lag.

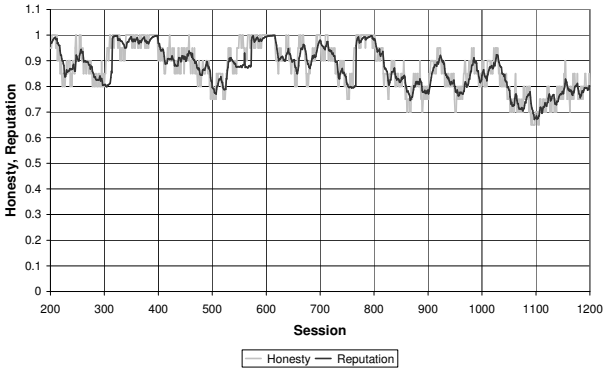

Figure 4: Honesty and reputation score for a single seller with no unfair buyers.

The instability of the seller behaviour is due to two opposing forces. One the one hand, the sellers try to increase their profit by reducing honesty. On the other hand the sellers are forced to stay relatively honest, otherwise they will be denounced by the reputation system, and lose profit as a result of reduced business. Section 4.3 describes the sellers' intelligence for balancing these two forces.

For comparison, a similar market was simulated with the replacement of 5 fair buyers with unfair buyers. The unfair buyers varied their behaviour over time, according to the state sequence in Figure 5.

The parameter $\mathrm{Prob}_{\mathrm{unfair}}$ was set to 1.0 (i.e. unfair buyers, when in a badmouthing or ballot stuffing state, acted unfairly at all times). Figure 6 shows the effect on a single seller in the market. As the graph shows, the unfair raters cause the seller's reputation to diverge from its true honesty by a significant degree (as much as 0.15), especially when the raters are badmouthing in sessions 500–600 and 900–1000. Since the seller is generally quite honest, the effect is most pronounced when the unfair buyers badmouth the seller.

Finally, the market with the same parameters was simulated, but this time with filtering of ratings enabled. As Figure 7 shows, the reputation score follows the honesty closely, which means that the unfair buyers are no longer able to distort the seller's reputation through badmouthing.

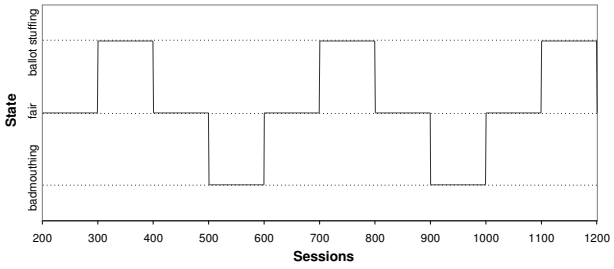

Figure 5: Sequence of states of unfair raters (1)

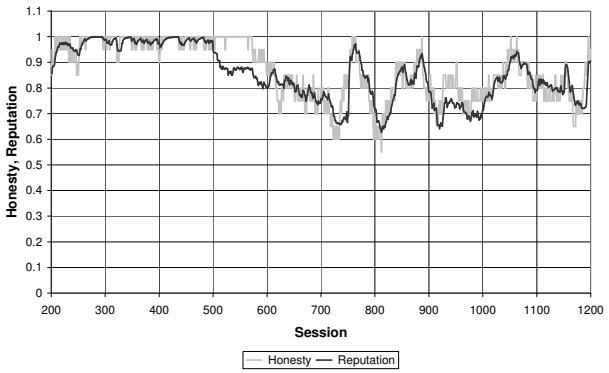

Figure 6: Honesty and reputation score for a single seller with 10% unfair raters and filtering disabled.

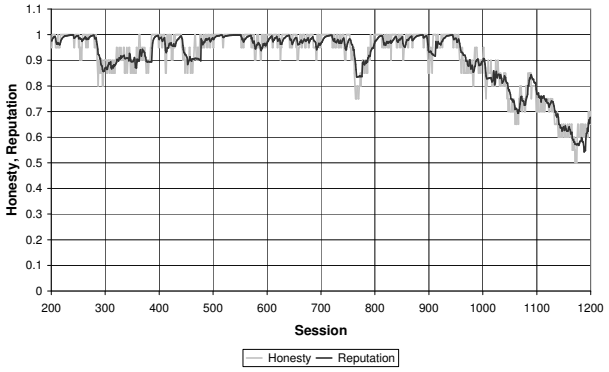

Figure 7: Honesty and reputation score for a single seller with 10% unfair raters when filtering is enabled

5.2 The Effectiveness of Filtering

The effectiveness of filtering was examined by simulating scenarios while varying two parameters:

- The proportion of unfair buyers (out of all buyers)
- The probability that unfair buyers would rate unfairly

5.2.1 Proportion of Unfair Raters

It would be expected that the filtering would be most effective with a low proportion of unfair raters. As the proportion of unfair raters increases, it becomes more difficult to determine which raters are truthful and which raters are lying.

To test the effectiveness of filtering as the proportion of unfair raters increases, the following simulation parameters were used, once with filtering enabled and once with it disabled:

Parameter	Value
number of buyers	50
fair	$45, 40, 35, 30, 25$
unfair	$5, 10, 15, 20, 25$
number of sellers	6
initial seller honesty	0.5

Figure 8 shows the average reputation error of all buyers when 10 buyers (20%) rate unfairly, according to the pattern shown in Figure 5. With no filtering, the sellers' reputations deviate substantially from their true honesties whenever the unfair buyers rate unfairly. The pattern of bad-mouthing and ballot-stuffing is quite evident in its effect on the sellers' reputations. With filtering enabled, most of this error is eliminated.

Filtering is also effective with a higher proportion of unfair raters. With 30% of buyers rating unfairly (Figure 9), the sellers' reputations begin to be affected, despite the filtering. However is still an improvement on no filtering, reducing the session error from around 0.3 to 0.1.

With 40% of raters rating unfairly, iterating filtering finally breaks down. As Figure 10 shows, the filtered scores become quite erratic, sometimes reflecting the sellers' average honesty (as for sessions 1100-1200) but sometimes deviating quite dramatically (as for sessions 700-800). With such a high proportion of unfair raters, there is no clear majority rating, thus while the filter will sometimes

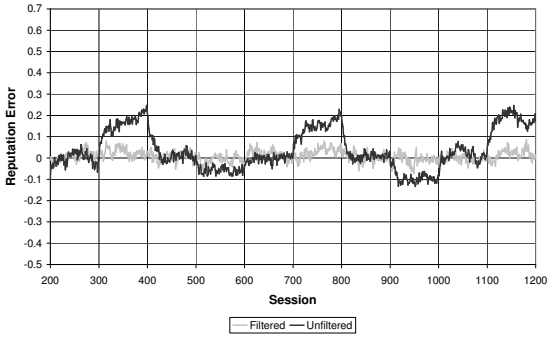

Figure 8: Average reputation error for all sellers with 10 (20%) unfair raters

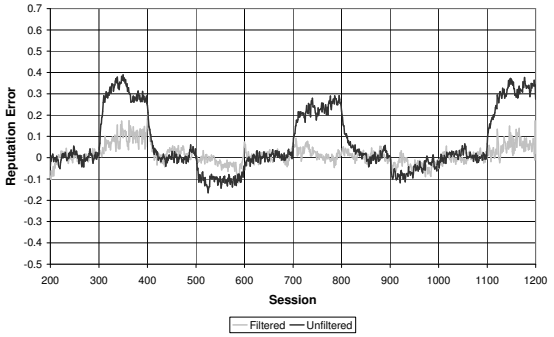

Figure 9: Average reputation error for all sellers with 15 (30%) unfair raters

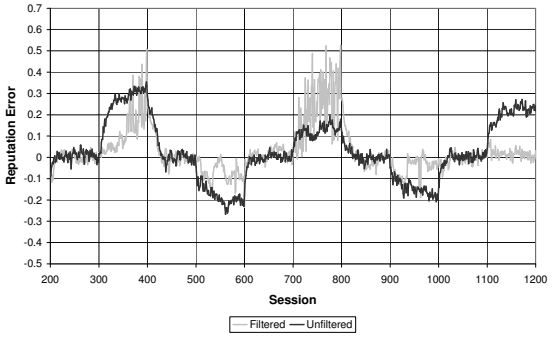

Figure 10: Average reputation error for all sellers with 20 (40%) unfair raters

correctly reject the unfair raters, it may incorrectly reject the fair raters, actually worsening the problem.

5.2.2 Degree of Unfairness of Unfair Raters

While a naive unfair rater might rate unfairly on every transaction, a more intelligent attacker could attempt to avoid detection by interspersing unfair ratings with fair ratings. This strategy can be simulated by reducing the probability that an unfair rater will actually rate unfairly on a transaction ($\mathrm{Prob}_{\mathrm{unfair}}$) below 1.0.

The effectiveness of filtering against this type of behaviour was assessed by running a simulation, over 1200 sessions, with the probability of unfairness gradually increasing from 0 to 1. For maximum effect, the proportion of unfair raters was set to 30% – the highest level at which filtering was effective in previous simulations. The sequence of states followed by unfair raters was also changed (see Figure 11), to provide a more detailed trace.

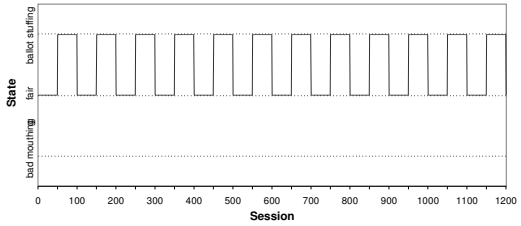

Figure 11: Sequence of states of unfair raters (2)

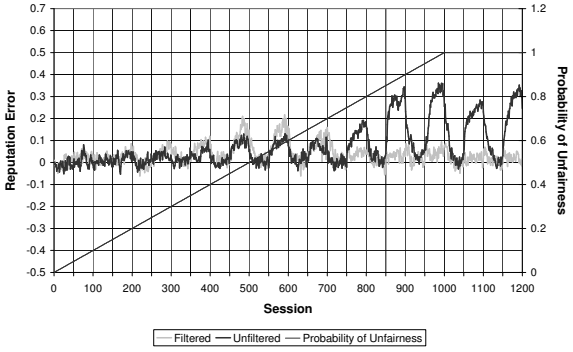

Figure 12: Average reputation error for all sellers with 30% unfair raters whose unfairness increases over time

As Figure 12 shows, the low unfairness probability strategy is partially effective in defeating filtering. When the probability of unfairness is around 0.5 (e.g. sessions 450–500 and 550–600), it is high enough to have a significant effect on seller's reputations (up to 0.2), but still low enough to escape detection by the filter. Below 0.5, the unfair rater has very little effect on the reputations anyway; above 0.5, the filter correctly detects and eliminates the unfair raters.

Figure 12 illustrates particularly clearly the dynamics of the filtering method described in this paper. The filter sensitivity with respect to the unfairness probability that triggers ratings to be filtered out, can be adjusted with the quantile parameter described in Sec.3.1.

6 Discussion and Conclusion

The simulation results presented above demonstrate the power of filtering in overcoming the problem of unfair raters in Bayesian reputation systems. The filtering algorithm is most effective when a moderate (less than 30%) number of raters behave consistently unfairly. If either the proportion of unfair raters exceeds 30%, or the unfair raters are only sometimes unfair, the filter is less effective. When the proportion of unfair raters reaches 40% the filter is ineffective, and is in fact counterproductive.

In the context of real applications, these limitations are not unreasonable. It would be difficult to ensure the accuracy of reputations in any system once the proportion of unfair raters is almost one half. Real systems could expect a much lower rate of unfair raters. The requirement that unfair raters be consistently unfair is somewhat more difficult, since this would enable a clever unfair rater to manipulate the reputation system to his or her advantage, simply by disguising an attack amongst fair ratings. However by forcing an unfair rater to adopt this 'go slow' strategy, the filter has still reduced the degree to which the unfair rater can influence the system. Also, different choices of q in the filtering algorithm may improve the effectiveness of the filter (the correct choice will depend on the application).

A number of general observations can be made. Sellers' honesty parameters tend to move freely between low and high levels, which does not seem totally realistic. This is because buyers only consider the expected value of a transaction, so a buyer is indifferent between a seller with a high honesty level and high price, and a seller with a low honesty level and a low price. This is, of course, subject to the buyer's risk attitude, but because the risk attitude is adapted to maximise profit, buyers tend to become superrational and thus risk-neutral. Presumably, real buyers are more risk averse than this, although this may in part be due to the absence of accurate reputation mechanisms in the real world.

More interestingly, when unfair raters succeed in causing sellers' reputations to diverge from their honesty levels, buyers respond by becoming more risk averse or more risk seeking (e.g. if the reputations are, in general, artificially inflated, then buyers react by becoming more risk averse). This (completely rational) behaviour agrees with common experience and helps validate the simulation, as well as providing further justification for effective filtering.

It is clear that statistical and machine-learning methods such as iterated filtering are – when used in isolation – insufficient to fully combat the problem of unfair ratings. However in combination with other filtering methods, such as discounting raters as a function of how well/poorly their ratings correspond with own experience, discounting raters based on trust referral systems, or blacklisting raters in response to overwhelming complaints, they can be an effective tool to improve the robustness of reputation systems. In particular, iterated filtering shows great promise as a technique for improving the accuracy, and hence the practical viability, of Bayesian reputation systems.

References

[1] S. Buchegger and J.-Y. Le Boudec. A Robust Reputation System for Mobile Ad-hoc Networks. Technical Report IC/2003/50, EPFL-IC-LCA, 2003.

[2] M. Chen and J. Singh. Computing and Using Reputations for Internet Ratings. In *Proceedings of the Third ACM Conference on Electronic Commerce (EC'01)*. ACM, October 2001.

[3] F. Cornelli et al. Choosing Reputable Servents in a P2P Network. In *Proceedings of the eleventh inter-*

national conference on World Wide Web (WWW'02). ACM, May 2002.

[4] C. Dellarocas. Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior. In *ACM Conference on Electronic Commerce*, pages 150–157, 2000.

[5] M. Ekstrom and H. Bjornsson. A rating system for AEC e-bidding that accounts for rater credibility. In *Proceedings of the CIB W65 Symposium*, pages 753–766, September 2002.

[6] A. Jøsang. Trust-Based Decision Making for Electronic Transactions. In L. Yngström and T. Svensson, editors, *Proceedings of the 4th Nordic Workshop on Secure Computer Systems (NORDSEC'99)*. Stockholm University, Sweden, 1999.

[7] A. Jøsang, E. Gray, and M. Kinateder. Analysing Topologies of Transitive Trust. In *Proceedings of the Workshop of Formal Aspects of Security and Trust (FAST 2003)*, Pisa, September 2003.

[8] A. Jøsang, S. Hird, and E. Faccer. Simulating the Effect of Reputation Systems on e-Markets. In N. C., editor, *The proceedings of the First International Conference on Trust Management*, Crete, May 2003.

[9] A. Jøsang and R. Ismail. The Beta Reputation System. In *Proceedings of the 15th Bled Electronic Commerce Conference*, Bled, Slovenia, June 2002.

[10] P. Macnaughton-Smith, W. Williams, M. Dale, and L. Mockett. Dissimilarity analysis: A New Technique of Hierarchical Sub-division. *Nature*, 202:1034–35, 1964.

[11] L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, and A. Halberstadt. Ratings in Distributed Systems: A Bayesian Approach. In *Proceedings of the Workshop on Information Technologies and Systems (WITS)*, 2001.

[12] L. Mui, M. Mohtashemi, and A. Halberstadt. A Computational Model of Trust and Reputation. In *Proceedings of the 35th Hawaii International Conference on System Science (HICSS)*, 2002.

[13] P. Yolum and M. Singh. Dynamic Communities in Referral Networks. *Web Intelligence and Agent Systems (WIAS)*, 1(2):105–116., 2003.

[14] B. Yu and M. Singh. Detecting Deception in Reputation Management. In *Proceedings of the Second Int. Joint Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 73–80. ACM, 2003.