A Detailed Comparison of Probabilistic Approaches for Coping with Unfair Ratings in Trust and Reputation Systems

Jie Zhang[1] Murat Şensoy[2] Robin Cohen[1]
[1]School of Computer Science, Universtiy of Waterloo, Ontario, Canada
[2]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

Abstract

The unfair rating problem exists when a buying agent models the trustworthiness of selling agents by also relying on ratings of the sellers from other buyers. Different probabilistic approaches have been proposed to cope with this issue. In this paper, we first summarize these approaches and provide a detailed categorization of them. This includes our own "personalized" approach for addressing this problem. Based on the implication of such analysis, we then focus on experimental comparison of our approach with two key models in a framework that simulates a dynamic electronic marketplace environment. We specifically examine different scenarios, including ones where the majority of buyers are dishonest, buyers lack personal experience with sellers, sellers may vary their behavior, and buyers may provide a large number of ratings. Our study provides the basis for deciding which approach is most appropriate to employ, in which scenario.

1 Introduction

In electronic marketplaces populated by self-interested agents, buying agents would benefit by modeling the trustworthiness of selling agents, in order to make effective decisions about which agents to trust. How to effectively represent the trustworthiness of sellers then becomes a challenge that must be addressed in order to ensure that users feel secure when engaging in commerce online. One method for representing sellers' trustworthiness is to ask other buying agents in the system (called advisors) to provide ratings of sellers. The problem of unfair ratings may then arise. Advisors may provide unfairly high ratings to promote sellers. This is referred to as "ballot stuffing" [1]. Advisors may also provide unfairly low ratings, in order to cooperate with other sellers to drive a seller out of the marketplace. This is referred to as "bad-mouthing".

A variety of approaches have been proposed to use probabilistic reasoning for addressing the problem of unfair ratings [14, 16, 11, 13, 15]. These probabilistic approaches provide a theoretically sound basis [5]. For example, the beta reputation system (BRS) [14] estimates the trustworthiness of sellers using a probabilistic model based on the beta probability density function. It filters out the ratings that are not in the majority amongst other ones. Teacy et al. [11] propose the TRAVOS model to discount unfair ratings by modeling the trustworthiness of advisors based on buyers' personal experience with the advisors' ratings. The personalized approach proposed by Zhang and Cohen [16] combines buyers' private knowledge and the public knowledge of the advisors held by the system, to model the trustworthiness of the advisors.

We begin by surveying some of these existing probabilistic approaches to the unfair rating problem, characterizing their capabilities and categorizing them in terms of three main dimensions: public versus private, global versus local, and endogenous versus exogenous. Based on the study, we then focus on experimental comparison of the representative approaches, including BRS, TRAVOS and the personalized approach. We propose a framework that simulates a dynamic electronic marketplace environment involving possibly deceptive buying and selling agents. We specifically examine different scenarios, including ones where the majority of buyers are dishonest, buyers lack personal experience with sellers, sellers may vary their behavior, and buyers may provide a large number of ratings.

Our results show that in general the personalized approach obtains the best performance and TRAVOS outperforms BRS. BRS performs much worse when the majority of buyers are dishonest and is affected by the situation where buyers may provide a large number of ratings. The more direct comparison between the personalized approach and TRAVOS in a scenario where buyers do not have much experience with sellers implies that it is better to consider public knowledge of advisors. In this scenario, BRS also performs better than TRAVOS when the majority of buyers are honest. TRAVOS is also heavily affected by the situation where selling agents may vary their behavior widely.

The rest of the paper is organized as follows. Sections 2 and 3 present a detailed categorization of some existing probabilistic approaches for coping with unfair ratings. Section 4 provides the framework used for simulating an e-marketplace and for conducting experiments. Section 5 presents the results of comparing the three approaches. Finally, Section 6 discusses the difference of our work with some related work, and Section 7 concludes the paper and proposes future work.

2 Probabilistic Approaches

In this section, we provide a brief summary of some existing probabilistic approaches for coping with the unfair rating problem. Advantages and shortcomings of the approaches are also pointed out.

2.1 Beta Reputation System

The beta reputation system (BRS) proposed by Jøsang and Ismail [4] estimates trustworthiness of selling agents using a probabilistic model. This model is based on the beta probability density function, which can be used to represent probability distributions of binary events. This model is able to estimate the trustworthiness of a seller by propagating ratings provided by multiple advisors. Ratings are binary in this model (“1” or “0”, for trustworthy or not trustworthy). Ratings are combined by simply accumulating the amount of (m) ratings supporting good trustworthiness and the amount of (n) ratings supporting bad trustworthiness. An example of the beta probability density function when $m = 7$ and $n = 1$ is shown in Figure 1. The trustworthiness of the seller S is then represented by the expected value of the beta function, which is the most likely probability value that the seller will act honestly in the future. The formalization of this is given as follows:

$$\alpha = m + 1, \qquad \beta = n + 1$$

$$Tr(S) = \frac{\alpha}{\alpha + \beta} \tag{1}$$

To handle unfair ratings provided by advisors, Whitby et al. [14] extend BRS to filter out those ratings that are not in the majority amongst other ones. More specifically, feedback provided by each advisor is represented by a beta distribution. If the cumulated trustworthiness of the seller falls between the lower and upper boundaries of feedback, this feedback will be considered as fair feedback. Figure 1 shows a demonstration of this process when the lower and upper boundaries are 0.1 and 0.99 respectively. When the cumulated trustworthiness of the seller is within the black area, the advisor’s ratings will be considered as unfair ratings. However, this approach is only effective when the significant majority of ratings are fair. This approach also does not consider separately buyers’ personal experience with advisors’ ratings.

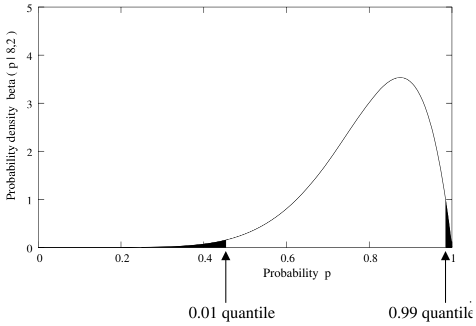

Figure 1. PDF when $m = 7$ and $n = 1$ [14]

2.2 TRAVOS

Teacy et al. [11] propose the TRAVOS model, which is a trust and reputation model for agent-based virtual organizations. This approach is also based on the beta probability density function. It copes with inaccurate reputation advice by accomplishing two tasks. The first task is to estimate the accuracy of the current reputation advice (ratings of “1” or “0”) provided by the advisor based on the buyer’s personal experience with the advisor’s previous advice. More specifically, the TRAVOS model divides the interval of [0,1] into N_{bin} number of equal bins. It then finds out all the previous advice provided by the advisor that is similar to the advice being currently given by the advisor. The two pieces of advice are similar if they are within the same bin. The accuracy of the current advice will be the expected value of the beta probability density function representing the amount of the successful and unsuccessful interactions between the buyer and the seller when the buyer follows the previous advice.

The second task is to adjust reputation advice according to its accuracy. The aim of this task is to reduce the effect of inaccurate advice. This task is necessary because it can deal with the situation where an advisor untruthfully rates a seller a large number of times, also known as the problem of advisors “flooding” the system [1]. Experimental results show that TRAVOS outperforms the BRS system [11]. However, this model also has some weaknesses. It assumes that selling agents act consistently. This assumption might not be true in many cases. The second problem is that this model relies only on the buyer’s personal experience with the advisor’s advice. It will be problematic when the buyer does not have much experience with selling agents.

2.3 The Personalized Approach

The personalized approach proposed by Zhang and Cohen [16] models the trustworthiness of advisors by taking into account both buying agents' private experience with the advisors' advice and the public knowledge about the advisors held by the system. This approach introduces the concept of a time window to discount older ratings and to avoid the situation where some advisors may flood the system. This approach also offers flexibility for buyers to weight their value in the private experience and the public knowledge. More specifically, the personalized approach allows a buying agent to estimate the reputation (referred to as private reputation $R_{pri}(A)$) of an advisor A based on their ratings for commonly rated selling agents. The private reputation value of the advisor is not shared with the public and may vary for different buyers.

When the buyer has limited private knowledge of the advisor, the reputation (referred to as public reputation $R_{pub}(A)$) of the advisor will also be considered. The public reputation is based on the public's opinions about the advisor's advice. It is estimated based on all ratings for the sellers ever rated by the advisor. The advisor will have a high public reputation value if its ratings are consistent with the majority of other ones. The public reputation of the advisor is shared by all of the public. It is the same for every buyer.

Finally, the trustworthiness of the advisor A will be modeled by combining the weighted private and public reputations, as follows:

$$Tr(A) = wR_{pri}(A) + (1 - w)R_{pub}(A) \qquad (2)$$

The weight w above is determined based on the estimated reliability of the private reputation using the Chernoff Bound theorem [8].

2.4 Bayesian Network Approach

Wang and Vassileva [13] propose a Bayesian network-based trust model in a peer-to-peer file sharing system. In this system, file providers' capabilities are evaluated by different aspects, including download speed, file quality, and file type. A naïve Bayesian network is constructed to represent conditional dependencies between the trustworthiness of file providers and the aspects. Each user holds a naïve Bayesian network for each file provider. If a user has no personal experience with a file provider, he may ask other users (advisors) for recommendations. A recommendation provided by an advisor will be considered by the user according to the trust value he has of the advisor. The trust value is updated by a reinforcement learning formula. More specifically, it will be increased/decreased after each comparison between the naïve Bayesian networks held by the user and the advisor for the file provider. The Bayesian network-based trust model takes into account preference similarity between users and advisors. However, this approach assumes that the aspects of file providers' capabilities are conditionally independent. This assumption is unrealistic in many systems. For instance, users may prefer high quality video and picture files, but not care much about the quality of text files.

2.5 Weighted Majority Algorithm

Yu and Singh propose an algorithm that handles unfair ratings using a version of the weighted majority algorithm (WMA) [15]. In their algorithm, weights are assigned to the advisors. These weights are initialized to be 1 and can be considered as the trustworthiness of the corresponding advisors. The algorithm predicts the trustworthiness of sellers based on the weighted sum of the ratings provided by those advisors.

Yu and Singh propose to tune the weights of the advisors after an unsuccessful prediction so that the weights assigned to the advisors are decreased. They assume that the ratings of dishonest advisors may conflict with the observations of the buyers receiving these ratings. By decreasing the weights of these advisors over time, unfair ratings are filtered.

3 Characteristics of Approaches

We have summarized different approaches proposed to handle unfair ratings, including BRS, TRAVOS, the personalized approach, the Bayesian network approach, and the weighted majority algorithm. In this section, we characterize these approaches by presenting a categorization of them and an analysis of their capabilities.

3.1 Categories

These approaches can be categorized in terms of three dimensions, an "endogenous-exogenous" dimension, a "public-private" dimension, and a "global-local" dimension.

Endogenous versus Exogenous: Jøsang et al. [5] divide the approaches for handling unfair ratings into two categories, endogenous and exogenous. Methods in the category of "endogenous" assume that unfair ratings can be recognized by their statistical properties. Therefore, these approaches are based on analyzing and comparing the rating values themselves. For example, BRS falls into this category. It relies on the majority ratings of a seller to judge whether a rating is fair/unfair. The public reputation of the personalized approach also falls into this category and assigns low weights to ratings of advisors whose ratings are inconsistent with the majority of others' ratings. Methods in the category of "exogenous" assume that the advisors with low trustworthiness are likely to give unfair ratings and vice versa. Therefore, they use the trustworthiness of advisors to decide which ratings are unfair. The TRAVOS model, the Bayesian network-based trust model, the weighted majority algorithm and the private reputation of the personalized approach all fall into this category. They all update the trustworthiness of an advisor based on the consistency determined from buyers' experience with the advisor.

Public versus Private: An approach for handling unfair ratings is "private" if the buyer estimates the trustworthiness of an advisor based on only its personal experience with previous ratings provided by the advisor. The current rating provided by the advisor is likely to be fair if its past ratings are also fair. For example, the TRAVOS model [11] estimates the accuracy of the advisor's current rating based on the amount of fair and unfair previous ratings provided by it that are similar to its current rating. These private approaches also belong to the "exogenous" category. An approach for handling unfair ratings is "public" if the buyer estimates trustworthiness of the advisor based on all the ratings it has supplied for any of the sellers in the system. A rating is likely to be reliable if it is the same as/similar to most of the other ratings for the same sellers. For example, the BRS approach [14] filters out unfair ratings that are not in the majority amongst others. These pubic approaches also belong to the "endogenous" category.[1]

Global versus Local: An approach is "local" if it filters out unfair ratings based on only the ratings for the seller currently being evaluated as a possible partner (referred to as the current seller). For example, the BRS approach judges whether a rating of a seller is an unfair rating based on whether it is consistent with the majority of other ratings of the same seller. An approach for handling unfair ratings is considered as "global" if it estimates the trustworthiness of an advisor based on ratings for all the sellers that the advisor has ever rated. The Baysian network-based and WMA approaches are "global" approaches.

Table 1. Categorization of Approaches

Categories	Public/Endogenous	Private/Exogenous
Global	Personalized	TRAVOS Personalized Bayesian WMA
Local	BRS	

The categorization of approaches for handling unfair ratings is summarized in Table 1. Note that there is no approach falling in the category of "private and local". This is simply because there is a conflict in this category. A buying agent asks advice about a selling agent from an advisor only when it lacks personal experience with the seller. An approach belonging to the "private and local" category will evaluate the trustworthiness of the advisor based only on the seller's ratings and the advisor's ratings for the seller currently being evaluated as a possible partner (referred to as the current seller). The buyer's limited experience with the current seller is certainly not sufficient for determining the trustworthiness of the advisor. Also note that the personalized approach falls into both the categories of "public/endogenous" and "private/exogenous" because it has the combination of the private and public reputation components.

3.2 Capabilities

To compare the above approaches, we analyze the capabilities they have. We list the following four capabilities that an effective approach should have.

- **Majority:** An effective approach should be able to cope with unfair ratings even when the majority of the ratings of a seller is unfair. Endogenous/public approaches assume that unfair ratings can be recognized by their statistical properties, and therefore may suffer in this situation. For example, the performance of BRS largely decreases when the majority of ratings are unfair, which will be demonstrated in Sections 5.1 and 5.2.1. Approaches that belong to the category of "private" rely on buyer's personal experience with advisors' advice and will not be affected by this situation;

- **Flooding:** An approach should also be able to deal with the situation where advisors may provide a large number of ratings within a short period of time. The approach of BRS is affected by this situation and the reason for this will be further explained in Section 5.2.4. The Bayesian network-based model is also affected because one advisor may be able to quickly build up its reputation by providing a large number of truthful ratings within the short period. One possible way to cope with this is to consider only a limited number of ratings from each advisor within the same period of time, as used by the personalized approach [16]. In the WMA approach, truthful ratings do not increase advisors' trustworthiness, and therefore WMA is not affected by this situation;

- **Lack** (of Experience)**:** An approach should still be effective even when buyers do not have much experience with sellers. Private approaches (TRAVOS, Bayesian, and WMA) suffer from this type of situation. Both BRS and the personalized approach are able to deal

[1] Although the "endogenous-exogenous" and "public-private" dimensions are similar, they categorize approaches based on different aspects.

with this situation because they can rely on the public knowledge of the ratings provided for sellers;

- **Varying:** An approach should be able to deal with changes of selling agents' behavior. Because of changes of selling agents' behavior, buying agents may provide different ratings for the same seller. Even though two ratings provided within different periods of time are different, it does not necessarily mean that one of them must be unfair. Different ways are proposed to deal with this situation. BRS [14] and the personalized approach use a forgetting factor λ ($0 \leq \lambda \leq 1$) to dampen ratings according to the time when they are provided. Older ratings are dampened more heavily than more recent ones.

Table 2. Capabilities of Approaches

Approaches	Majority	Flooding	Lack	Varying
BRS			√	√
TRAVOS	√	√		
Personalized	√	√	√	√
Bayesian	√			
WMA	√	√		

Table 2 lists capabilities of the approaches summarized in the previous section. In this table, the mark "√" indicates that an approach has the capability. For example, the BRS approach is capable of dealing with changes of sellers' behavior and is still effective when buyers do not have much experience. Note that only the personalized approach has all these capabilities.

4 Experimental Framework

In this section, we introduce a framework for conducting experiments to compare different approaches for handling unfair ratings. The marketplace environment used for experiments is populated with self-interested buying and selling agents. The buyers and sellers are brought together by a procurement (reverse) auction, where the auctioneer is a buyer and bidders are sellers. There is a central server that runs the auction. In the marketplace, a buyer B that wants to purchase a product p sends a request to the central server. Sellers interested in selling the product to the buyer will register to participate in the auction. The buyer will first limit the sellers it will consider for the auction, by modeling their trustworthiness. To directly compare the performance of the approaches for coping with unfair ratings, we use an algorithm for the buyer to model the trustworthiness of the sellers, only making use of ratings from advisors.

Assume that a buyer B considers ratings provided by advisors that are trustworthy. It sends a request to the central server to ask for all the ratings provided by the trustworthy advisors $\{A_1, A_2, ..., A_m\}$ ($m \geq 1$) for the seller S. Suppose that the advisor A_i ($1 \leq i \leq m$) provided $N_{pos}^{A_i}$ positive ratings and $N_{neg}^{A_i}$ negative ratings.[2] These ratings will be discounted based on the trustworthiness of the advisor, so that the ratings from less trustworthy advisors will carry less weight than ratings from more trustworthy ones.

Jøsang [3] provides a mapping from beliefs defined by the Dempster-Shafer theory to the beta function as follows:

$$\begin{cases} b = \frac{N_{pos}^{A_i}}{N_{pos}^{A_i}+N_{neg}^{A_i}+2} \\ d = \frac{N_{neg}^{A_i}}{N_{pos}^{A_i}+N_{neg}^{A_i}+2} \\ u = \frac{2}{N_{pos}^{A_i}+N_{neg}^{A_i}+2} \end{cases} \tag{3}$$

where b, d and u represent belief, disbelief and uncertainty parameters, respectively. For our setting of trust modeling, b represents the probability that the proposition that the seller is trustworthy is true, and d represents the probability that the proposition is false. Note that $b + d + u = 1$ and $b, d, u \in [0, 1]$. As also pointed out in [4] and [15], beliefs and disbeliefs can be directly discounted by the trustworthiness of the advisor A_i as follows:

$$\begin{cases} b' = Tr(A_i)b \\ d' = Tr(A_i)d \end{cases} \tag{4}$$

where $Tr(A_i)$ is the trustworthiness of A_i. From Equations 3 and 4, we then can derive a discounting function for the amount of ratings provided by A_i as follows:

$$\begin{cases} D_{pos}^{A_i} = \frac{2Tr(A_i)N_{pos}^{A_i}}{(1-Tr(A_i))(N_{pos}^{A_i}+N_{neg}^{A_i})+2} \\ D_{neg}^{A_i} = \frac{2Tr(A_i)N_{neg}^{A_i}}{(1-Tr(A_i))(N_{pos}^{A_i}+N_{neg}^{A_i})+2} \end{cases} \tag{5}$$

The trustworthiness of seller S can be calculated as follows:

$$Tr(S) = \frac{[\sum_{i=1}^{m} D_{pos}^{A_i}] + 1}{[\sum_{i=1}^{m}(D_{pos}^{A_i} + D_{neg}^{A_i})] + 2} \tag{6}$$

A seller is considered trustworthy if its trust value is greater than a threshold γ. It will be considered untrustworthy if the trust value is less than δ. The buyer in our framework will allow only a limited number of the most trustworthy sellers to join the auction. This can be achieved by using the trust thresholds. If there are no trustworthy sellers, the sellers with trust values between γ and δ may also be allowed to join the auction.

[2] In this work, we consider only ratings that are binary because the approaches we compare are all using the beta density function for representing binary ratings.

The buyer will then convey to the central server which sellers it is willing to consider, and the pool of possible sellers is thus reduced. Sellers $\{S_1, S_2, ..., S_n\}$ ($n \geq 1$) allowed to join the auction submit their bids by setting the prices and values for the non-price features of the product p. The buyer will select the winner of the auction as the seller whose product (described in its bid) gives the buyer the largest profit, based on the buyer's valuation of the product V_B, formalized as follows:

$$S_{win} = \arg \max_{j=1}^{n} (V_B - P_{S_j}) \tag{7}$$

where P_{S_j} is the price of product offered by seller S_j.

Once the buyer has selected the winning seller, it pays that seller the amount indicated in the bid. The winning seller is supposed to deliver the product to the buyer. However, it may decide not to deliver the product. The buyer will report the result of conducting business with the seller to the central server, registering a rating for the seller. It is precisely these ratings of the seller that can then be shared with other buyers.

In this work, we compare only three approaches: BRS, TRAVOS and the personalized approach. These three approaches are all based on the beta density function. They are also representative amongst other approaches. They cover all the four categories of "global", "local", "public/endogenous" and "private/exogenous". They are also useful for demonstrating the importance of the capabilities, which some of the approaches have and others do not. We implement the TRAVOS model and the personalized approach for modeling the trustworthiness of advisors. Note that TRAVOS does not discount older ratings of sellers. We also implement the BRS approach to filter out unfair ratings for each seller. The aggregation of fair ratings is slightly different from Equation 5 by assuming $Tr(A_i)$ is always 1 because trustworthiness of advisors is not modeled by BRS.

4.1 Simulation Setting

We simulate a marketplace operating for a period of 60 days. The marketplace involves 90 buyers. These buyers are grouped into three groups. They have different numbers of requests. Each group of buyers has a different number (20, 40 and 60) of requests. In our experiments, we assume that there is only one product in each request and each buyer has a maximum of one request each day. For the purpose of simplicity, we also assume that the products requested by buyers have the same valuation for buyers. After they finish business with sellers, buyers rate sellers. Some dishonest buyers from each group will provide unfair ratings. We allow 2 buyers from each group to leave the marketplace at the end of each day. Accordingly, we also allow 6 buyers to join the marketplace at the end of each day. Some of them may also provide unfair ratings, to keep the percentage of dishonest buyers in each group the same in each day.

There are also 6 sellers in total in the marketplace. Each 2 sellers acts dishonestly in different percentages (0%, 25% and 50%) of their business with buyers. We assume that all sellers have the same cost for producing the products because all products have the same valuation for buyers.

We also set different parameters in the experiments. We set the lower and upper boundaries for BRS to be 0.1 and 0.99 respectively, as recommended in [14]. The number of bins N_{bin} used by the TRAVOS model is chosen to produce the best results in our experiments. The weight of private reputation used by the personalized approach is also selected to produce the best performance. λ for BRS and the personalized approach is set to be 1. We set the threshold γ to be 0.7 and δ to be 0.3. Therefore, a seller is considered as trustworthy if its trust value is greater than 0.7 and untrustworthy if it is below 0.3. In our experiments, a buyer is considered to be honest if its trust value is greater than 0.5, otherwise, it is untrustworthy.

4.2 Performance Measurement

We measure the performance of an approach for coping with unfair ratings in two ways. One is its ability of detecting dishonest advisors. An effective approach should be able to correctly detect dishonest advisors. This performance can be measured by the false positive rate (FPR) and false negative rate (FNR). A false positive represents that a honest advisor is incorrectly detected as a dishonest advisor. A false negative represents that an advisor is misclassified as honest but actually is dishonest. The lower values of FPR and FNR imply better performance. We also use Matthew's correlation coefficient (MCC) [7] to measure the approaches' performance on detecting dishonest advisors. MCC is a convenient measure because it gives a single metric for the quality of binary classifications, and is computed as follows:

$$MCC = \frac{(t_p.t_n - f_p.f_n)}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{8}$$

where f_p = false positives, t_p = true positives, f_n = false negatives, t_n = true negatives. An MCC value is between -1 and +1. A coefficient of +1 represents a perfect detection, 0 an average random detection and -1 the worst possible detection.

We also measure the performance of an approach based on how much buyers can benefit if the approach is employed. We use two metrics to represent this benefit, the profit of buyers and the ratio of buyers' successful business with sellers. Eventually, the higher the ratio of successful business the buyers can have with sellers, the larger the profit they will be able to gain.

5 Experimental Results and Analysis

In this section, we present experimental results comparing the three approaches, BRS, TRAVOS and the personalized approach. We first provide the comparison of their overall performance. We then analyze how these approaches perform in different scenarios.

5.1 Overall Performance Comparison

In this experiment, we vary the percentage of dishonest buyers (from 20% to 80%) in the marketplace environment. We then measure the average MCC values for TRAVOS, BRS and the personalized approach for the period of 60 days. Results are shown in Figure 2. From this figure, we can see that the personalized approach produces the highest MCC values for different percentages of dishonest buyers. TRAVOS performs better than BRS. The performance of these approaches will generally decrease when more buyers are dishonest. Note that the performance of BRS is close to random classification when 50% of buyers are dishonest and becomes much worse when the majority of buyers are dishonest. This result confirms our argument in Sections 2.1 and 3.2.

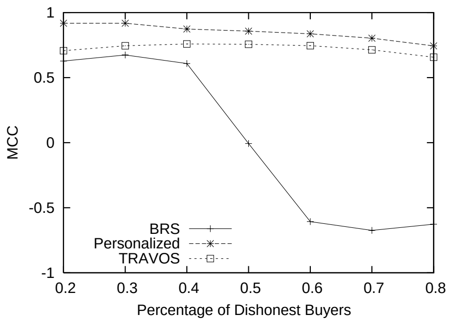

Figure 2. Detecting Dishonest Buyers

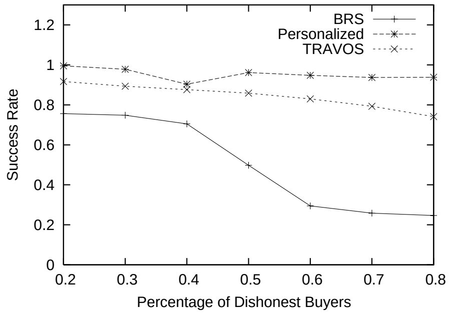

Figure 3. Ratio of Successful Business

We measure the ratio of buyers' successful business with sellers. We call a transaction between a buyer and a seller successful business if the seller delivers what it promised in its bid submitted to the buyer's auction when the seller is selected as the winner of the auction. We measure the success ratio of buyers after 60 days. We then average the success ratio over the total number of buyers in the marketplace (90 in our experiments). In this experiment, we also measure the average total profit of buyers after 60 days. The results are shown in Figures 3 and 4. These two figures are very similar and also confirm the results shown in Figure 2. Note that the performance of the personalized approach decreases when 40% of the buyers are dishonest. This is because the public reputation component of the personalized approach does not perform well when a large number of buyers are dishonest. When 40% of buyers are dishonest, the personalized approach still considers the public reputation part. Its performance is then affected by the public part. When more than 50% of buyers are dishonest, the personalized approach will rely only on the private component.

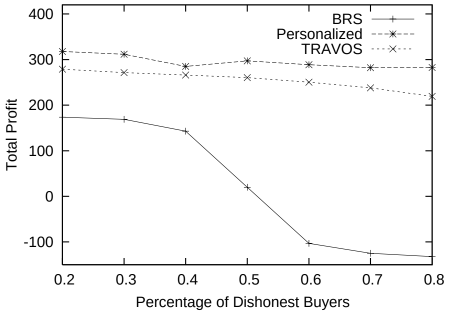

Figure 4. Total Profit of Buyer

In summary, the personalized approach performs the best. The TRAVOS model performs better than BRS, which is similar to the results in [11]. BRS performs much worse when the majority of buyers are dishonest, which will be further analyzed in depth in the next section. We will also analyze how the three approaches perform in different scenarios.

5.2 Analysis of Different Scenarios

In order to further compare the three approaches and analyze their capabilities, we simulate different scenarios where the majority of buyers are dishonest, buyers do not have much experience with sellers in the marketplace, sellers may vary their behavior widely, and buyers may provide a large number of ratings in a short period of time. Note that in this section we will only present the performance of the approaches in detecting dishonest buyers because this performance is correlated with the results of total profit and success ratio of buyers, as presented in the previous section.

5.2.1 Dishonest Majority

BRS assumes that a significant majority of the buyers are honest. This is why the performance of BRS decreases dramatically when half of the buyers are liars as shown in Figures 2, 3 and 4.

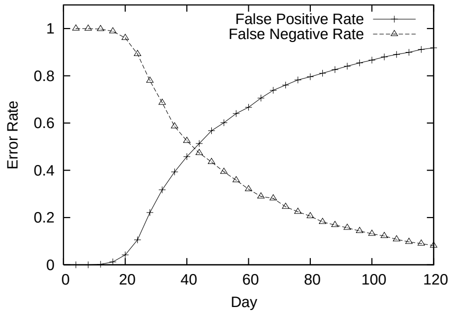

Figure 5. Error Rate of BRS

In order to better see the reasons behind this performance decrease, we show the error of BRS in detecting dishonest buyers when 50% of buyers are dishonest in a period of 120 days, in Figure 5. From this figure, we can see that the ratio of false negatives approaches 0. However, the ratio of false positives continuously increases and approaches 1. This means that BRS tends to label every buyer as dishonest.

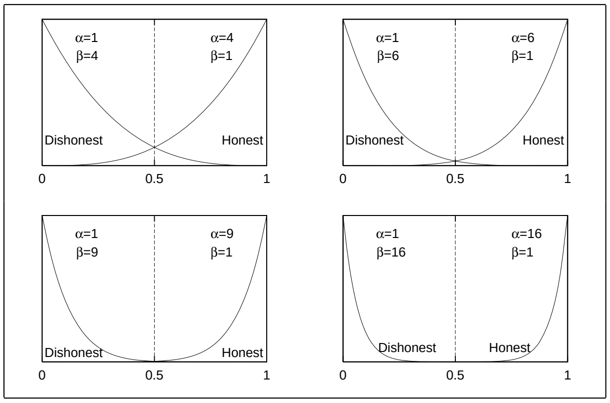

Figure 6. BRS for 50% of Dishonest Buyers

Figure 6 explains the statistical foundation of BRS's behavior when 50% of buyers are dishonest. For a honest seller, dishonest buyers provide unfairly low ratings and their Beta distributions reside near 0, according to Equation 1 when β increases. However, for the same seller, honest buyers provide high ratings that make their Beta distributions reside near 1. Overall, the expected value of the aggregated Beta distribution becomes 0.5 and it does not stay within the margins defined by the lower and upper boundaries of the buyers' Beta distributions. Hence, both the dishonest and honest buyers are regarded as dishonest.

5.2.2 Lack of Personal Experience

The TRAVOS model relies only on buyers' personal knowledge with advisors' advice, whereas BRS and the personalized approach also considers public knowledge of advisors' advice. The public knowledge is useful especially when buyers do not have much experience with sellers, and in consequence do not have much personal knowledge with advisors' advice. In this experiment, we demonstrate the performance of these three approaches in detecting dishonest buyers when 30% of buyers are dishonest. We plot the MCC values of their performance over 60 days, as shown in Figure 7. We can see that both BRS and the personalized approach perform much better than the TRAVOS model in the beginning 10 days. This confirms our argument that buyers should rely on public knowledge about advisors when they do not have much experience with sellers. We also can see from Figure 7 that the performance of BRS will decrease after 30 days and become worse than that of TRAVOS. The reason for this will be further analyzed and explained in Section 5.2.4.

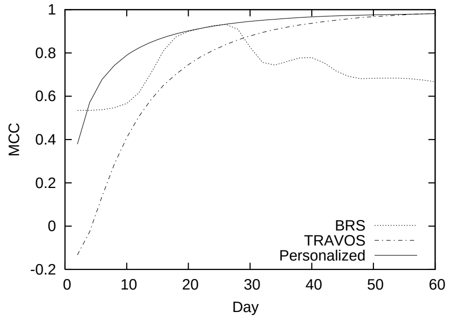

Figure 7. Detecting Dishonest Buyers

In the second experiment, we directly compare the performance of the personalized approach with that of TRAVOS in the scenario where buyers do not have much experience with sellers. In the experimental setting, 30% of buyers are dishonest. Half of all buyers have more requests for products and another half have fewer requests. Buyers having more requests will have more experience with sellers. We measure how much the personalized approach outperforms TRAVOS in detecting dishonest buyers.

Results are shown in Figure 8. In both cases when buyers have more or less experience with sellers, the personalized approach outperforms TRAVOS. From the figure, we can see that the difference is larger when buyers do not have much experience with sellers. The performance difference will decrease day after day because buyers will have more

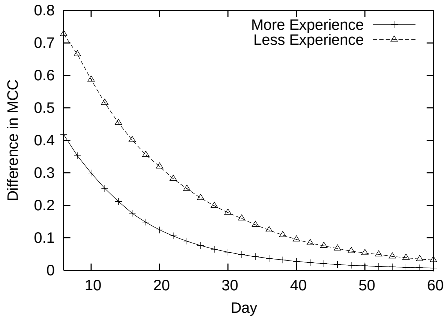

Figure 8. Personalized vs. TRAVOS

and more experience with sellers. This suggests that an approach of modeling the trustworthiness of advisors for coping with unfair ratings should rely on public knowledge of advisors' advice as well when buyers do not have much experience with sellers.

5.2.3 Seller Varying Behavior

The personalized approach introduces the concept of a time window when evaluating the trustworthiness of advisors. For example, it only compares a buyer's and an advisor's ratings if these two ratings are within the same time window when computing the private reputation of the advisor. This is to deal with the problem when sellers vary their behavior widely. However, as we point out in Section 2.2, the TRAVOS model is not able to deal with this problem. In this section, we present experimental results to confirm this argument.

We first carry out an experiment to compare the personalized approach with the TRAVOS model in the situation where sellers may change their behavior. In this experiment, the sellers that vary their behavior will be dishonest in 25% or 50% of the period of 60 days. We also have three types of sellers. The first type of sellers act dishonestly in a uniform manner. The second type of sellers is honest first and then becomes dishonest. The third type of sellers acts dishonestly first and then honestly later on. We run simulations separately 500 times for each type of seller and average the results. We then calculate the mean and standard deviation of the two approaches' performance in detecting dishonest buyers. We also set $\lambda = 0$, because the seller behavior is varying so much.

From the results shown in Figure 9, we can see that the mean performance of the personalized approach consistently increases after each day. The standard deviation of its performance stays nearly at 0, which implies that the performance of the personalized approach is not affected by sellers' varying behavior. However, the mean performance of the TRAVOS model decreases heavily after 45 days and the standard deviation of its performance is considerably large for the beginning 15 days and the ending 15 days. Therefore, TRAVOS does not perform well when sellers change their behavior widely.

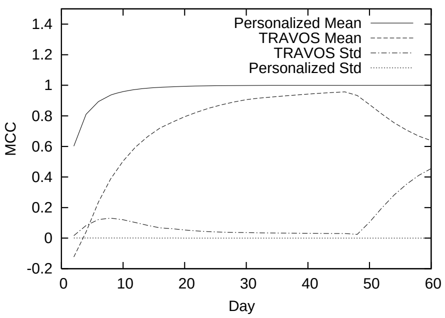

Figure 9. Personalized vs. TRAVOS

We also carry out another experiment to analyze in depth how the TRAVOS model will be affected by different types of seller varying behavior. In this experiment, we have sellers vary their behavior in different frequencies. All sellers in this experiment will act honestly first and then dishonestly later on. These different types of sellers vary their behavior for 1, 3 and 5 times respectively within the period of 60 days, as shown in Figure 10. This figure shows an example how a seller that is dishonest in 50% of the period of 60 days will vary their behavior. A seller's honesty of "1" on the vertical axis means that the seller acts honestly in the corresponding day and "0" represents dishonest behavior.

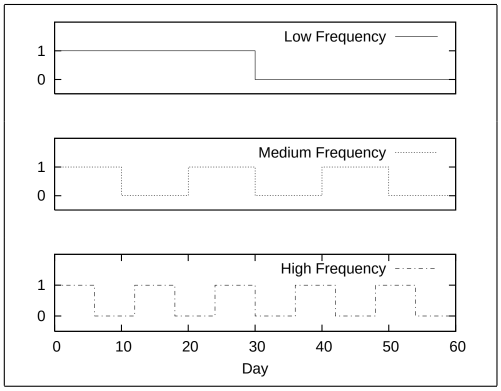

Figure 10. Seller Varying Behavior

The performance of TRAVOS for different frequencies of seller changing behavior is presented in Figure 11. When sellers change their behavior very frequently, the performance of TRAVOS will also change more often. The change of its performance is less than that when sellers vary behavior less frequently. When the sellers change their behavior only once from being honest to be dishonest, the performance decreases to a great extent to nearly a random classification.

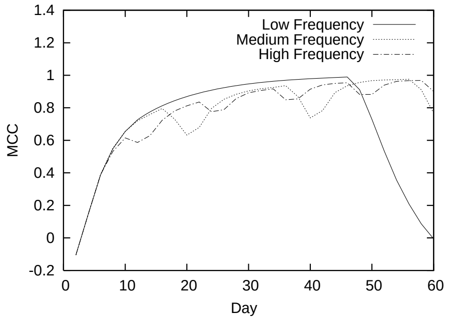

Figure 11. Performance of TRAVOS

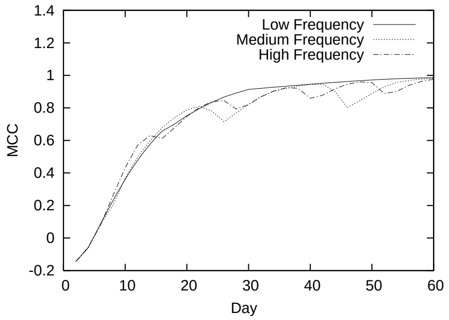

Figure 12. Performance of TRAVOS

We also show the results of the performance of TRAVOS when all sellers act dishonestly first and then honestly later on. Similarly, sellers vary their behavior in different frequencies. The results are shown in Figure 12. Comparing this figure with Figure 11, we can see that the performance of TRAVOS is affected less than that in the situation where sellers act honestly first and then dishonestly. Especially when sellers vary their behavior at a low frequency, the performance of TRAVOS does not have much change compared to that in Figure 11. In the simulation framework, sellers acting dishonestly at the beginning will have very low trust values and be prevented from joining buyers' auctions. The changes of their behavior will no longer affect the performance of detecting dishonest buyers. This also implies that a more effective varying behavior for a seller is to be honest first to build up its trustworthiness, and then acts dishonestly to exploit the marketplace (a behavior explored by such trust researchers as Tran and Cohen [12], and Sen and Banerjee [9]).

5.2.4 Buyers' Flooding

Buyers' flooding is the situation where buyers (advisors) may provide a large number of ratings for a seller in a short period of time. To deal with situation, for example, the personalized approach uses the concept of a time window and considers only a limited number of ratings from one buyer for the seller within the same time window. As discussed in Section 3.2, the BRS approach will be heavily affected by buyers' flooding. In the case where buyers provide a large number of unfair ratings, BRS will suffer from the dishonest majority problem as demonstrated in previous sections. In this section, we carry out experiments to show that BRS is affected even when buyers provide a large number of fair ratings within a short period of time.

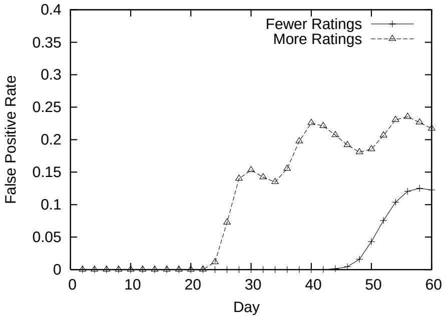

Figure 13. False Positive Rate of BRS

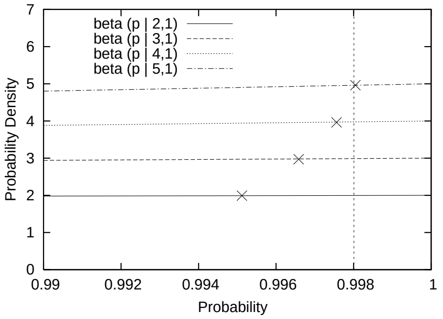

Figure 14. BRS Unable to Cope with Flooding

In this experiment, we involve two types of buyers. The first type of buyers has many more requests and therefore will provide a lot of ratings to sellers. The second type of buyers provide fewer ratings. In both cases, 20% of buyers are dishonest. We run simulations for the two cases separately and measure the false positive rate of BRS in detecting dishonest buyers. Results are shown in Figure 13. We can see that after 20 or 40 days, BRS will start incorrectly classifying honest buyers as dishonest. The false positive rate is higher when buyers provide more ratings. Therefore, BRS is even affected by the situation where buyers may provide a large number of fair ratings.

We further analyze the statistical foundation of this phenomenon, as shown in Figure 14. The vertical line on the figure represents the expected value (trustworthiness) of a seller when there are 500 positive ratings and 0 negative ratings provided by buyers for the seller. This figure also shows the Beta distributions for buyers that provide 1, 2, 3, and 4 positive ratings respectively, and 0 negative ratings for the seller. The "×" symbols on the distributions represent the cut-off points of upper bounds of these distributions. We can see from the figure that the seller's expected value only falls within the upper bounds of the distribution with 4 positive ratings. Therefore, the honest buyers that have only provided 1, 2 or 3 positive ratings will be incorrectly classified as dishonest buyers. This therefore increases the false positive rate of BRS.

5.3 Summary of Results

We have carried out experiments to compare the overall performance of the three representative approaches, TRAVOS, BRS and the personalized approach. We measure their accuracy in detecting dishonest buyers, the ratio of buyers' successful business with sellers when these approach are employed, and the total profit of buyers. Results show that the personalized approach performs the best, TRAVOS performs better than BRS, and BRS performs much worse when the majority of buyers are dishonest.

We also analyze how these three approaches perform in different scenarios. Results show that the personalized approach performs much better than TRAVOS especially when buyers do not have much experience with sellers. In this case, BRS also performs better than TRAVOS when the majority of buyers are honest. TRAVOS suffers from the situation where sellers may vary their behavior, and is heavily affected especially when sellers first build up their trust by being honest and then act dishonestly. BRS is shown to be ineffective when buyers provide a large number of ratings for a seller.

6 Related Work

In the work of Zhang and Cohen [16], they also provide a detailed survey of existing approaches for coping with unfair ratings. Our work extends their survey and is focused on probabilistic approaches. We also provide more extensive discussion of the categorization of these approaches and their capabilities. In their work, they focus on the development of the personalized approach and provide only some preliminary results to show the effectiveness of the personalized approach in and of itself. We conduct experiments to compare the personalized approach with other two representative approaches, BRS and TRAVOS, in a more dynamic e-marketplace environment. How the approaches perform in different scenarios is also analyzed in great depth.

The ART Testbed [2] is proposed to provide unified performance benchmarks for comparing trust and reputation modeling approaches. The current testbed specification is in an artwork appraisal domain where appraisers want to buy artwork about which they may have limited knowledge. They may then seek information about artwork from other appraisers (opinion providers). Opinion providers may choose to lie about the true value of the artwork. The appraisers will model the trustworthiness of opinion providers based on their own knowledge about the opinion providers or reputation opinion of other appraisers (reputation providers). These reputation providers may choose to lie about opinion providers' true trust values. An approach for coping with untruthful reputation opinions from opinion providers may then be integrated and evaluated by the ART Testbed. However, integrating TRAVOS, BRS and the personalized approach into the testbed is challenging. These approaches are developed for a rather simpler e-marketplace environment. They allow only binary ratings to represent simple and objective results of transactions between sellers and buyers (advisors). Advisors modeled by these approaches do not make profit from providing advice or pay cost to generate advice. Overly simplifying the ART Testbed may lose its advantages, and adapting these approaches to the complicated testbed may change their original design. Furthermore, the winning approach IAM [10] for the 2006 ART Testbed competition does not even consider reputation opinions from other appraisers. This decision raises the concern about the importance of an approach for coping with untruthful reputation opinions in this testbed, and whether the results of comparing the approaches based on this testbed will be significant.

7 Conclusions and Future Work

In this paper, we focus on probabilistic approaches for coping with the unfair rating problem. We survey some existing approaches, characterize their capabilities, and categorize them in terms of three main dimensions: public versus private, global versus local, and endogenous versus exogenous. These discussions provide a deep understanding of differences amongst these approaches and inspire empirical studies in our paper.

We then focus on experimental comparison of the representative approaches, including BRS, TRAVOS and the personalized approach. We propose a framework that simulates a dynamic electronic marketplace environment involving possibly deceptive buying and selling agents. These three approaches are compared for the first time in terms of their capabilities for detecting dishonest buyers. Total profit of buyers is also the most direct and important measure used in the comparison between these approaches. We further specifically examine different scenarios, including ones where the majority of buyers are dishonest, buyers lack personal experience with sellers, sellers may vary their behavior, and buyers may provide a lot of ratings. Such an empirical study is useful for highlighting the importance of the capabilities of the approaches.

We have compared the performance of different approaches in detecting the possible dishonest buyers that are being fairly consistent in lying. For future work, in our evaluations, it would be worthwhile to explore the case where some dishonest buyers lie only for some sellers while being honest for other sellers. It would also be worthwhile to consider other types of dishonest buyers from the literature, such as the Exaggerated Positive and Exaggerated Negative types defined in [15]. The performance of detecting these types of dishonest buyers would then be evaluated and compared for those approaches.

We may want to investigate more advanced dishonest buyers that are strategic. For example, some dishonest buyers may have mixed lying types. Inspired by the evaluation in [14], a marketplace may involve some buyers that have an adaptive lying strategy where buyers may learn from the marketplace and build some strategies to adapt their lying types or lying frequency. A similar idea can be found in the work of Sen and Banerjee [9], where strategic agents may exploit the marketplace. We are interested in demonstrating how the existing approaches perform in this kind of marketplace environment. We are also interested in seeing how well they handle marketplaces where strategic agents collude with each other.

Coping with unfair ratings from advisors in e-marketplaces by a modeling of their trustworthiness has some similarity with the challenge of addressing shilling attacks in recommender systems. The research of [6] suggests that the general algorithms used by attackers (i.e. the kind of attacks) may be useful to model and that the areas being attacked (e.g. low use items) may influence the possible damage that can be inflicted. For future work, it would be useful to simulate these attacks and to compare the robustness of the approaches against the attacks.

Introducing innovation to the design of trust modeling systems used in agent-oriented e-marketplaces is a crucial concern, as part of the ongoing effort to promote electronic commerce to businesses and organizations. This paper has demonstrated some key shortcomings of existing trust modeling systems and has discussed the advantages introduced by our particular personalized approach. As a result, specific directions are now available for users who are selecting trust modeling algorithms to run in e-marketplaces.

References

[1] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM Conference on Electronic Commerce (EC)*, Minneapolis, MN, 2000.

[2] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. A specification of the agent reputation and trust (art) testbed: Experimentation and competition for trust in agent societies. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2005)*, 2005.

[3] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[4] A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[5] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *(to appear). Decision Support Systems*, 2005.

[6] S. K. Lam and J. Riedl. Shilling recommender systems for fun and profit. In *Proceedings of the 13th International Conference on World Wide Web*, 2004.

[7] B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, 405:442–451, 1975.

[8] 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.

[9] S. Sen and D. Banerjee. Monopolizing markets by exploiting trust. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2006.

[10] W. T. L. Teacy, H. T. D., D. R. K., J. N. R., P. J., and L. M. The art of iam: The winning strategy for the 2006 competition. In *The Workshop on Trust in Agent Societies at The Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2007)*, 2007.

[11] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model. In *Proceedings of Fourth International Autonomous Agents and Multiagent Systems (AAMAS)*, 2005.

[12] T. Tran and R. Cohen. Improving user satisfaction in agent-based electronic marketplaces by reputation modeling and adjustable product quality. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-04)*, 2004.

[13] Y. Wang and J. Vassileva. Bayesian network-based trust model. In *Proceedings of the 6th International Workshop on Trust, Privacy, Deception and Fraud in Agent Systems*, 2003.

[14] A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. *The Icfain Journal of Management Research*, pages 48–64, February 2005.

[15] B. Yu and M. Singh. Detecting deception in reputation management. In *Proceedings of International Autonomous Agents and Multi Agent Systems (AAMAS)*, 2003.

[16] J. Zhang and R. Cohen. A personalized approach to address unfair ratings in multiagent reputation systems. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS) Workshop on Trust in Agent Societies*, 2006.