Prob-Cog: an Adaptive Filtering Model for Trust Evaluation

Zeinab Noorian[1], Stephen Marsh[2], and Michael Fleming[1]

[1]University of New Brunswick, Canada
[2]Communications Research Centre, Canada
{z.noorian@unb.ca,stephen.marsh@crc.gc.ca,mwf@unb.ca}

Abstract. Trust and reputation systems are central for resisting against threats from malicious agents in decentralized systems. In previous work we have introduced the Prob-Cog model of multi-layer filtering for consumer agents in e-marketplaces which provide mechanisms for identifying participants who disseminate unfair ratings by cognitively eliciting the behavioural characteristics of e-marketplace agents. We have argued that the notion of *unfairness* does not exclusively refer to deception but can also imply differences in dispositions. The proposed filtering approach goes beyond the inflexible judgements on the quality of participants and instead allows environmental circumstances and the human dispositions that we call optimism, pessimism and realism to be incorporated into our trustworthiness evaluation procedures. In this paper we briefly outline the two layers before providing a detailed exposition of our experimental results, comparing Prob-Cog to FIRE and the personalized approach under various attacks and normal situations.

1 Introduction

Open e-marketplaces are uncertain places. These uncertainties contribute to misunderstandings amongst the agents that inhabit them [4]. While malicious agents exist, the recommendations of even honest agents who are unknown must be considered to be unreliable. Strategies for managing the uncertainties exist. In particular, in order to diminish the risk of being misled by unfair advisers, a consumer agent seeks advice from participants with the most similar ratings [3],[13].

In previous work [7] we amended this common view of trustworthiness [10],[15] by introducing a new definition for *unfairness*. Unfairness can be examined across two categories: 1) *intentional*, a) participants consistently act malevolently and b) participants occasionally engage in deceitful activities. And 2) *unintentional*, as a result of a) lack of personal experiences and b) various behavioural characteristics resulting in different rating attitudes.

Our algorithm uses a two-layered filtering approach combining cognitive and probabilistic views of trust [4] to mainly target the intentional group of unfair advisers. We showed that modeling the trustworthiness of advisers based on a strict judgement of the quality of their recommendations is not complete unless it is accompanied by the analysis of their dispositions. Thus, through the comprehension of their rating attitudes, a consumer agent could take appropriate steps to evaluate them.

In this paper, we provide an overview of the algorithm before presenting new experimental results to show its efficacy in filtering unfair advisers whilst still managing the problem of unfairness to new or unknown others. We then describe the adaptive approach of the Prob-Cog model in determination of the employed threshold parameters in different environmental circumstances. Our experimental results show the utility of our approach in classification of various participants and, specifically, how consumers could detect more honest advisers in a community where the majority of participants are unfair. Our Prob-Cog filtering model can therefore be seen as an effective approach in modeling the reputation of advisers in a dynamic agent-oriented e-commerce application.

2 Related Works

Several reputation systems and mechanisms have been proposed for modeling the trustworthiness of advisers and coping with the problem of unfair ratings in multi-agent online environments. Below we provide a description on two representative approaches: FIRE and the personalized model. More detailed overviews of other existing trust and reputation systems can be found in [7],[8].

The FIRE Model [3] is a decentralized trust and reputation system designed for open multi-agent systems such as e-commerce applications. In FIRE, trust is evaluated within the context of a number of different information components:1)*Interaction Trust* (IT) that is built from the direct experience of buying agents ; 2) *Witness Reputation* (WR) that is based on the direct observation of selling agents' performance by third-party agents; 3)*Certified Reputation* (CR) which consists of certified references disclosed by selling agents; and 4) *Role-based Trust* (RT), which models the trust across predefined role-based relationships between two participants. In this trust model, each component has a deterministic trust formula with a relevant rating weight function [8]. These weight functions are designed such that they reflect intrinsic characteristics of their components. For example, in the IT component, the weight function is merely based on the recency of the reputation information, whereas in WR the weight is calculated based on the credibility of its reputation providers. To evaluate the credibility of reputation providers, FIRE has developed an adaptive mechanism to detect and filter out inaccurate reports. It defines an adaptive inaccuracy tolerance threshold based on the selling agents' performance variation to specify the maximal permitted differences between the actual performance and the provided ratings. Credibility ratings of the reputation providers are tuned to be inversely proportional to the differences, i.e., the higher the differences are, the lower their credibility.

Zhang [15] proposed a personalized approach for handling unfair ratings in centralized reputation systems. It provides public and private reputation components to evaluate the trustworthiness of advisers. Depending on the availability of reputation information, consumer agents would determine the weight of private and public reputation components differently. In the personalized approach, advisers share their subjective opinions over a common set of providers. To estimate the credibility of advisers, consumers estimate the recency of their ratings

using the concept of a time window [15] and exploit a probabilistic approach to calculate the expected value of advisers' trustworthiness based on their provided ratings.

Our work differs from them in a number of ways. Unlike other models[3],[9],[10], [15] which mainly evaluate the credibility of advisers based on the percentage of unfair ratings they provided, the Prob-Cog model takes initiative steps to aggregate several parameters in deriving the trustworthiness of advisers. That is, in addition to the similarity degree of advisers' opinions, it further aggregates their behavioural characteristics and evaluates the adequacy of their reputation information in its credibility measure. In this model, every consumer with different behavioural characteristics is able to objectively evaluate the similarity degree of advisers through a multi-criterion rating approach. Consumer agents could adaptively predict the trustworthiness of advisers using different credibility measures well-suited for various kinds of advisers. Prob-Cog provides a consumer with the ability to simply adjust the influence of each view of trust based on its own preferences and determines the influence of each layer in its decision making. Besides, while in most existing models [9],[11],[13],[15] the evaluation of the adopted thresholds are not addressed explicitly, in this model we explore some determinant factors which are important in evaluating the adopted thresholds effectively.

3 The Prob-Cog Filtering Algorithm

In the Prob-Cog model, consumers analyze the neighbours' trustworthiness based on two types of information. The first, which helps build the first layer of our filtering algorithm, is used to identify malicious participants with a complementary model of deception[14] who lie significantly in their ratings. It also detects newly-joined agents with insufficient personal experiences. The second helps consumers to recognize the behavioural characteristics of their neighbours. As such, it will be able to subjectively evaluate their degree of trustworthiness. Note that, in the second layer of the model, consumers take an analytical approach in order to detect deceitful participants with volatile dispositions who cheat opportunistically. By hiding their true intentions, this group of deceitful participants impose greater risk and insecurity to the system compared with those with a frequently deceptive attitude[1],[6],[8].

Below we provide an overview of the formal model. A more detailed exposition can be found in [7].

3.1 First Layer: Evaluating the Competency of Neighbours

A consumer agent C sends a query to its neighbours $N = \{N_1, N_2, ..., N_i\}$ requesting information about providers $P = \{P_1, P_2, ..., P_q\} \subseteq \{P_1, P_2, ..., P_m\}, q \leq m$ on interactions occurring before a time threshold T (which diminishes the risk of changeability in a provider's behaviour), and with a Quality of Service (QoS) threshold Ω to imply C's belief about an acceptable minimum level of trust.

Neighbour N_k responds by providing a rating vector $R_{(N_k,P_j)}$ for each provider. It contains a tuple $\langle r, s \rangle$ which indicates the number of successful (r) and unsuccessful (s) interaction results with provider P_j respectively. In the first layer

of Prob-Cog, neighbours are asked to provide merely a *binary rating*: "1" means that P_j is reputable and "0" means not reputable. Thus, considering the consumer's QoS threshold, they will send reputation reports as a collection of positive and negative interaction outcomes. Once the evidence is received, for each $R_{(N_k,P_j)}$, C calculates the expected value of the probability of a positive outcome for a provider P_j [9] as:

$$E(pr_r, P_j) = \frac{r+1}{r+s+2} \tag{1}$$

For an e-marketplace we use $E(pr_r, P_j)_{Par}$ where $Par \in \{C\} \cup N$ implies participants of the community.

Clearly, $0 < E(pr_r, P_j)_{Par} \leq 1$ and as it approaches 0 or 1, it indicates *unanimity* in the body of evidence[5]. That is, particularly large values of s or r provide better intuition about an overall tendency and service quality of providers. In contrast, $E(pr_r, P_j)_{Par} = 0.5$, (i.e, $r = s$) signifies the maximal conflict in gathered evidence resulting in increasing the uncertainty in determining the service quality of providers. Based on these intuitions, we are able to calculate the degree of reliability and certainty of ratings provided by neighbours.

Let x represent the probability of a successful outcome for a certain provider. Based on the Definitions(2) and (3) in [12], the *Reliability degree* of each $R_{(N_k,P_j)}$ is defined as:

$$c(r,s) = \frac{1}{2}\int_0^1 \mid \frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s \; dx} - 1 \mid \; dx \tag{2}$$

Similar to $E(pr_r, P_j)_{Par}$, we can use $c(r,s)_{Par}$.

Following [12], reliability is a minimum when $E(pr_r, P_j)_{Par} = 0.5$. As such, the less conflict in their ratings, the more reliable the neighbours would be. However, in Prob-Cog, C would not strictly judge the neighbours with rather low reliability in their $R_{(N_k,P_j)}$ as deceptive participants since this factor could signify both dishonesty of neighbours and the dynamicity and fraudulent behaviour of providers. That is, some malicious providers may provide satisfactory quality of service in some situations when there is not much at stake and act conversely in occasions associated with a large gain. As such, even though they retain a certain level of trustworthiness, their associated reliability degree is low.

To address this ambiguity, C computes the $E(pr_r, P_j)_C$ and $c(r,s)_C$ of its personal experiences, $R_{(C,P_j)}$, for a common set of providers. Through the comparison of neighbours' metrics with its own, it would select those with a similar rating pattern and satisfactory level of honesty as its *advisers*. More formally, it measures an average level of dishonesty of N_k by:

$$\overline{d}_{(N_k)} = \frac{\sum_{j=1}^{|P|} \mid E(pr_r, P_j)_C - E(pr_r, P_j)_{N_k} \mid}{|P|} \tag{3}$$

It may also happen that a truthful neighbour lacks in experience. Thus, despite its inherent honesty, its reliability degree is low and it is not qualified to play the role of adviser. To address this, we introduce an uncertainty function $\overline{U}_{(N_k)}$

to capture the intuition of information imbalance between C and N_k as follows:

$$\overline{U}_{(N_k)} = \frac{\sum_{j=1}^{|P|} \mid (c(r,s)_C - c(r,s)_{N_k})_{P_j} \mid}{|P|} \tag{4}$$

In light of the uncertainty function, the opinions of deceptive neighbours who attempt to mislead consumer agents by supplying a large number of ratings are discounted. Similarly, this model hinders short-term observations of newly-joined agents from having influence on a consumer agent's decision making process. Given the formulae 3 and 4, the *competency degree* of N_k is calculated by reducing its honesty based on its certainty degree:

$$Comp_{(N_k)} = (1 - \overline{d}_{(N_k)}) * (1 - \overline{U}_{(N_k)}) \tag{5}$$

Consumer C determines an incompetency tolerance threshold μ which indicates an acceptable level of a neighbour's incompetency. It further chooses the neighbours with $(1 - Comp_{(N_k)}) \leq \mu$ as its potential advisers and filters out the rest.

To attain partial perception on overall quality of the environment, a consumer C evaluates an approximate dishonesty of participants based on its observation of the quality of its neighbours in this layer. It calculates an Approximate Dishonesty Coefficient $ADC_{(C)}$ as the ratio of detected incompetent neighbours to all of its neighbours as follows:

$$ADC_{(C)} = \frac{|\{N_k|(1 - Comp_{(N_k)}) > \mu \ , k \leq i\}|}{|N|} \tag{6}$$

It is worthwhile to note that, since in this layer we target the participants with a significant lying pattern, detecting fraudulent agents with oscillating rating attitudes is left for the next layer.

3.2 Second Layer: Calculating a Credibility Degree of Advisers

In the first phase of the Prob-Cog model a consumer agent has obtained a rough estimate of the honesty level of neighbours and selects a subset of them as its advisers. However, the open e-marketplace allows various kinds of participants with distinctive behavioural characteristics [2] to engage in the system. Besides, the multi-dimensional rating system provides tools for a consumer agent to objectively evaluate the performance of service providers across several criteria with different degrees of preference. Evidently, the measured QoS is mainly dependent on how much the criteria with a high preference degree are fulfilled[8]. Owing to the different purchasing behaviour of the agents, it is expected that preference degrees vary from one participant to another, resulting in dissimilar assessment of the quality of the *same* service. As such, computing the credibility of advisers regardless of their behavioural characteristics and rating attitudes, and merely based on their subjective opinions would not sufficiently ensure high quality judgements of their trustworthiness.

In the second layer of Prob-Cog, C gives credits to advisers to the extent that their evaluation of each criterion of a negotiated context is similar to its own experiences. For this purpose, it asks advisers about mutually agreed criteria over which they have bargained with *highly-reliable* providers whose reputation value has been recently released in the form of binary ratings. They also are requested to include the time of the latest interaction with such information so as to give a higher weight to more recent feedback. For this we adopt the concept of forgetting factor presented in [9],[15] and define the recency factor as:

$$T_{(C,A_k)_{P_j}} = \frac{1}{\lambda^{T_{A_k} - T_C}} \tag{7}$$

Here, T_{A_k} and T_C indicate the adviser's and consumer's time windows when they had an experience with the same provider. Also, the λ represents the forgetting parameter and $0 < \lambda \leq 1$. When $\lambda = 1$, there is no forgetting and all the ratings are treated as though they happened in the same time period. In contrast, $\lambda \approx 0$ specifies that ratings from different time windows will not be significantly taken into account. Similarly to [15], in this filtering algorithm, the recency factor is characterized with a discrete integer value where 1 is the most recent time period and 2 is the time period just prior. Also, it is presumed that the adviser's ratings are prior to those a consumer agent supplies so that $T_{A_k} \geq T_C$.

Adviser A_k responds with an interaction context $IC_{(A_k,P_j,T_A)}$ that contains a tuple of *weight* and *value*: $\{W_i.V_i | i = 1..n\}$ and the latest interaction time T_A for each provider. Given A_k's interaction context, a consumer agent would estimate the possible interaction outcomes of an adviser based on its own perspective. That is, C will examine its $IC_{(C,P_j,T_C)}$ -which contains pairs of weight and value: $\{Y_i.R_i | i = 1..n\}$- and replace A_k's preferences W_i with its personal preference degrees Y_i. Based on this, the interaction context of A_k is updated to: $IC'_{(A_k,P_j,T_A)} = \{Y_i.V_i | i = 1..n\}$. We formalize the difference of C and A_k in assessing a provider P_j as follows:

$$\mathit{Diff}_{(C,A_k)_{P_j}} = 1 - \frac{\sum_{i=1}^{n} Y_i \times R_i}{\sum_{i=1}^{n} Y_i \times V_i} \tag{8}$$

Based on Equations 7 and 8, C would calculate the *average* differences between the transaction result of A_k and its own experiences with a same set of providers as:

$$\overline{\mathit{Diff}}_{(C,A_k)} = \frac{\sum_{j=1}^{|P|} \mid \mathit{Diff}_{(C,A_k)_{P_j}} \mid * T_{(C,A_k)_{P_j}}}{|P|} \tag{9}$$

In the Prob-Cog model of filtering, we take a further step and embrace the diversity in participants as an influential factor in our credibility measures. For this purpose, C captures the overall tendency of A_k in evaluating the providers' QoS as:

$$Tendency_{(C,A_k)} = \frac{\sum_{j=1}^{|P|} Diff_{(C,A_k)_{P_j}}}{|P|} \tag{10}$$

A positive value of $Tendency_{(C,A_k)}$ indicates that an adviser has the attitude of overrating providers while a negative value declares that an adviser has a tendency to underrate providers.

Further, an adaptive threshold β is used to determine behavioural patterns of advisers such that if A_k's $IC'_{(A_k,P_j,T_A)}$ is compatible with those experienced by C ($\overline{Diff}_{(C,A_k)} \leqslant \beta$), they will be counted as *credible* advisers. In Prob-Cog, C determines the outlook of the advisers by analyzing $\overline{Diff}_{(C,A_k)}$. If it is marginally greater than β with a negative $Tendency_{(C,A_k)}$, the corresponding adviser's attitude is identified as *pessimistic*. Similarly, in case their differences marginally exceed β with a positive $Tendency_{(C,A_k)}$, the respective adviser's attitude is recognized as *optimistic*. We define a marginal error ϵ as a ratio of β and it is subjectively determined by a consumer agent. If A_k's $IC'_{(A_k,P_j,T_A)}$ significantly deviates from the consumer agent's direct experiences, they will be detected as *malicious* advisers with *deceitful* behavioural models.

The classification mechanism of the behavioural pattern of A_k based on C's interaction context is formally presented as follows:

$$BP_{(C,A_k)} = \begin{cases} Credible: & \overline{Diff}_{(C,A_k)} \leqslant \beta \\ Optimistic: & \beta < \overline{Diff}_{(C,A_k)} \leqslant \beta + \epsilon \ \& \ Tendency_{(C,A_k)} > 0 \\ Pessimistic: & \beta < \overline{Diff}_{(C,A_k)} \leqslant \beta + \epsilon \ \& \ Tendency_{(C,A_k)} < 0 \\ Deceitful: & \overline{Diff}_{(C,A_k)} > \beta + \epsilon \end{cases} \tag{11}$$

Given the $BP_{(C,A_k)}$, the credibility measure $CR_{(C,A_k)}$ is formulated as:

$$CR_{(C,A_k)} = \begin{cases} 1 - \overline{Diff}_{(C,A_k)}: & BP_{(A_k)} = Credible \\ (1 - \overline{Diff}_{(C,A_k)}) \times e^{-\theta * \overline{Diff}_{(C,A_k)}}: & BP_{(A_k)} = Optimistic \\ (1 - \overline{Diff}_{(C,A_k)}) \times e^{-\sigma * \overline{Diff}_{(C,A_k)}}: & BP_{(A_k)} = Pessimistic \\ 0: & BP_{(A_k)} = Deceitful \end{cases} \tag{12}$$

Here, θ and σ represent the optimistic and pessimistic coefficients respectively. Coefficients θ and σ are formalized with reference to the consumer's disposition as:

$$\theta = \begin{cases} max\{|\ Diff_{(C,A_k)_{P_i}}\ |\ |i = 1...m\} & Risk\text{-}Averse\ consumer \\ min\{|\ Diff_{(C,A_k)_{P_i}}\ |\ |i = 1...m\} & Risk\text{-}Taking\ consumer \end{cases} \tag{13}$$

$$\sigma = \begin{cases} min\{|\ Diff_{(C,A_k)_{P_i}}\ |\ |i = 1...m\} & Risk\text{-}Averse\ consumer \\ max\{|\ Diff_{(C,A_k)_{P_i}}\ |\ |i = 1...m\} & Risk\text{-}Taking\ consumer \end{cases} \tag{14}$$

The coefficient parameters ensure that the recommendation of advisers with volatile behaviour who have a high variability in their opinions is heavily discounted.

4 Evaluating Threshold Parameters

In the Prob-Cog, we combine cognitive and probabilistic views of trust such that each might have different weights depending on the consumer's endogenous factors, such as willingness and preferences. That is, some consumers might assign a great deal of influence on the probabilistic evaluation results while having less interest in the inclusion of the factor of behaviour in their evaluation. The priority of either view is projected into different thresholds dedicated to each layer. In this model, the values of the adopted thresholds are attributed to many factors such as: 1) the variation of the providers' performance, 2) percentage of neighbours with dishonest attitude and 3) the influence of the cognitive approach on the consumer's perspective ($Inf_{view:cog}$).

To optimally estimate β in the second layer, C needs to acquire enough information about the potential reasons for a reporter's inaccuracy. For example, variation in providers' performance can be served as a measure of reporters' inaccuracy[3]. Thus, the inaccuracy tolerance threshold β is evaluated by capturing the mean variation in providers' quality of products. However, for the precise calculation of the provider's performance variation, in this model C only selects highly-reliable providers whose $c(r, s) > 0.50$. Based on this principle, the variation of a provider P_j can be calculated as follows:

$$dev(C, P_j) = \sqrt{\frac{\sum_{r_i \in \Re(C,P_j)} (v_i - \bar{v})^2}{|\Re(C, P_j)|}} \tag{15}$$

And β is estimated as:

$$\beta = \frac{\sum_{j=1}^{|P|} dev(C, P_j)}{|P|}$$

where $dev(C, P_j)$ is the standard deviation of provider P_j's performance in the last interactions experienced by C. v_i is the value of the rating r_i which is the rating of C provided for P_j and $0 \leq v_i \leq 1$. And $\bar{v}$ is a mean value of all the rating values in the set of ratings $\Re(C, P_j)$ which is a collection of r_i.

Depending on the value of $Inf_{view:cog}$, consumers show different levels of interest in modeling the behavioural patterns of advisers. To satisfy their interests, ϵ is designed to give consumers an opportunity to detect different dispositions of advisers. However, initializing ϵ not only depends on $Inf_{view:cog}$ but also relies on β and the approximate dishonesty coefficient of participants ($ADC_{(C)}$) that is estimated in the first layer. More explicitly, in a dynamic environment where providers indicate highly-variant behaviour, a high value of β increases the risk that deceptive reporters remain undetected. In such conditions, ϵ should be automatically adjusted to a low value in order to protect consumers against spurious participants. As such, consumer C would compute ϵ as:

$$\epsilon = (1 - ADC_{(C)}) e^{-\beta} Inf_{view:cog} \tag{16}$$

As aforementioned, the incompetency tolerance threshold μ should be evaluated in such a way to be able to expel neighbours with significant dishonesty or

unreliability. Hence, we get the intuition that μ should be assigned a higher value than the second layer thresholds $(\beta + \epsilon)$ so as to be able to target only major dishonest participants. It also should be aligned with the cognitive preferences of consumers. That is, since a high value of $Inf_{view:cog}$ signifies the importance of the second layer's evaluation mechanism and behavioural modeling, a higher value of μ is desirable so as to reduce the number of filtered participants in the first layer and gives opportunity to consumers to cognitively evaluate the trustworthiness of advisers based on their behavioural characteristics in the second layer. Based on these principles, μ is calculated as follows:

$$\mu = n * (\beta + e^{-\beta} Inf_{view:cog}) \tag{17}$$

where $n > 1$.

5 Experimental and Comparison Results

In this section we explore the performance of our Prob-Cog model confronting different scenarios and attacks in comparison with two representative approaches: the FIRE model[3] and the personalized approach[15]. We picked out the important features of each model and conducted experiments to analyze how our model compares in similar conditions. For example, some experiments are dedicated to studying the effectiveness of different approaches dealing with dynamicity in an environment like the situation when providers change their behaviours. We also evaluate the accuracy of different models coping with a majority of unfair participants and indicate how exploitation of the cognitive view of trust could improve the performance in such situations. We further compare various approaches in addressing the bootstrapping problem of newcomers having limited experiences.

5.1 Performance Measurement

To measure the performance of different approaches, we have used the Matthews Correlation Coefficient (MCC)[16] to evaluate the quality of various approaches in differentiation between honest and dishonest participants. MCC is a precise metric and it gives a single value for the quality of binary classification and is calculated as follows:

$$MCC = \frac{(t_p.t_n) - (f_p.f_n)}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{18}$$

where in this paper t_p represents the number of advisers correctly detected as dishonest, t_n signifies advisers correctly detected as honest. Also, f_p represents honest advisers misclassified as dishonest and f_n signifies dishonest advisers that are incorrectly classified as honest advisers. The MCC value is between $[-1, +1]$. A coefficient of $+1$ indicates accurate detection, a coefficient of 0 indicates an average detection quality and -1 indicates the worst possible detection.

5.2 Cold Start

Consider the scenario when consumer agent C has recently joined the system and intends to bootstrap its relationship with its neighbours. In the Prob-Cog model,

C relies only on its few personal experiences. On the other side, in the personalized approach[15], C relies heavily on the public knowledge component when its personal knowledge is scarce. Exploiting public knowledge sounds promising when the majority of participants are honest. However, when an environment is controlled by a majority of deceptive participants, using public knowledge would be misleading.

The first experiment demonstrates C's performance in classification of participants when it has a limited number of personal experiences using Prob-Cog model and the personalized approach. It involves 100 advisers and 3 providers. We vary the percentage of dishonest advisers from 0% to 95% in an e-marketplace. We then measure an average MCC value for the Prob-Cog model and the personalized approach with 10 and 40 experiences commonly rated by both consumer C and advisers for the same set of providers. Results are presented in Figure 1. It indicates that the personalized approach produces high MCC when the majority of participants are honest while its performance degrades as the percentage of dishonest participants increases in an e-marketplace. It also implies that as the amount of personal knowledge increases, the resistance of C against misleading opinions of the majority of participants increases considerably. Figure 1 also shows that the Prob-Cog model consistently yields high performance in every condition since public knowledge does not have any influence on its evaluation mechanism.

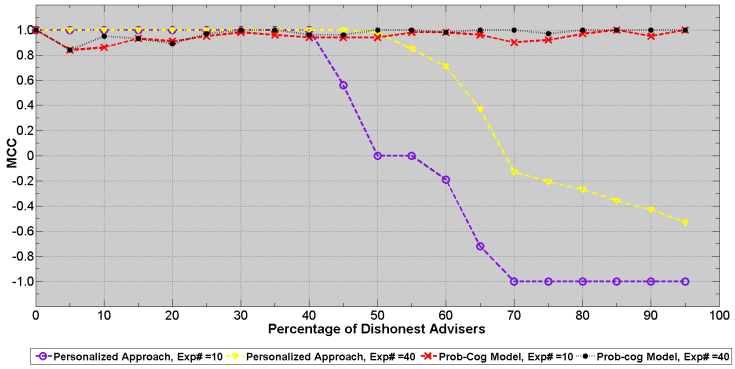

Fig. 1: Performance comparison of personalized & Prob-Cog approaches in classification of advisers

Dealing with insufficiency of personal experiences should not be restricted to consumer agents. Rather, it may happen that advisers with inadequate experiences unintentionally disseminate inaccurate information throughout an e-marketplace. Trust models must provide consumers with a mechanism to detect advisers with few experiences and reduce their influence in a consumer's decision making process.

The next experiment demonstrates how the Prob-Cog model evaluates the competency level of an intrinsically *honest* but inexperienced adviser with various amounts of experience. It involves one consumer C asking adviser N about its common experiences with 2 and 50 providers. N provides various percentages (0% to 100%) of differences in number of common experiences, presented by a

normal distribution. The results indicate that the competency of even an honest adviser degrades as its number of experiences decreases (Figure 2). We also observe that C can effectively evaluate the competency level of N even with a limited set of providers.

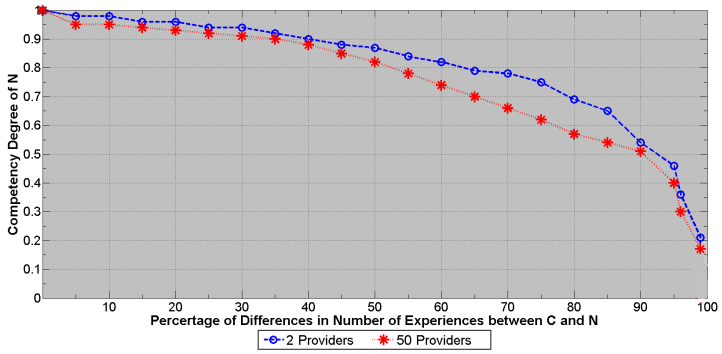

Fig. 2: Competency degradation of adviser N having different percentages of common experiences

5.3 Flooding

From [12] we get the intuition that the degree of reliability and confidence in advisers' opinions is directly related to the quantity of evidence they provide. This issue motivates advisers to provide a large number of ratings regarding certain providers upon the request of consumers. That is, deceitful advisers could manipulate a consumer by flooding it with a large number of ratings to increase their reliability substantially. Also, newcomers may exaggeratively increase their number of ratings so as to conceal their lack of experiences. This flooding problem affects the robustness and efficiency of trust models and should be dealt with effectively. To address such a problem, the Prob-Cog model discounts the num-

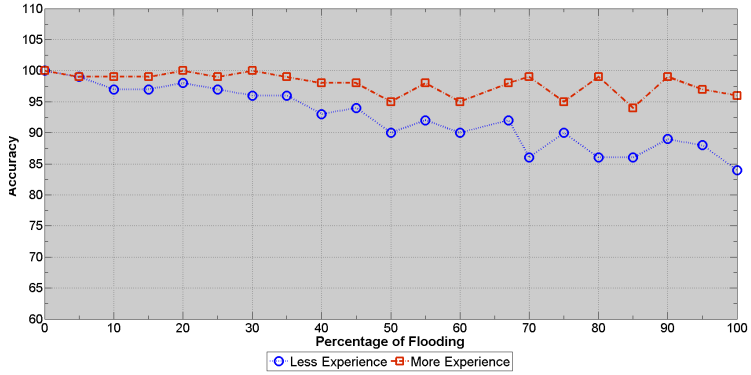

Fig. 3: The accuracy degree of Prob-Cog model dealing with different percentages of flooding

ber of ratings provided by advisers and degrades their reliability degree using Equations 2 and 4. The personalized model uses a different approach and exploits the concept of time window. It considers those ratings of advisers that are

provided in a same period of time as the ratings of the consumer more valuable and underestimates others.

The next class of experiment involves two consumers: 1) consumer C_1 with few experiences and 2) consumer C_2 with sufficient experiences. We examine the accuracy of the Prob-Cog model in classifying the advisers in a situation where different percentages of advisers from 0% to 100% flood consumers by providing 100-150 more ratings than C_1 and C_2. From the results in Figure 3 we can see that C_2 with sufficient experiences is more robust than C_1 with few experiences against the flooding attack of advisers. However, this attack does not have any significant negative impact on the overall performance of the Prob-Cog model.

5.4 Providers with Varying Behaviors

Our Prob-Cog model adopts a promising mechanism for capturing the variation in providers' performances. We have conducted specific experiments to demonstrate how our approach can accurately classify advisers in such an environment where providers continuously change their behaviour. This set of experiments involves 100 advisers which are divided in three groups: 1) 50% honest, 2) 25% deceitful who lie a small percentage of the time, and 3) 25% malicious with significant dishonesty. In this environment, providers have different levels of variation in their performance ranging from 5% to 100%.

To examine the effectiveness of our approach compared with other models, we consider four different threshold evaluation mechanisms: 1) fixed low threshold $\beta = 0.5$ as is used in the personalized approach [15], 2) fixed high threshold $\beta = 1.0$, 3) auto β which is exploited in the FIRE model [3] and 4) adaptive $\beta + \epsilon$ which is used by the Prob-Cog model in the second layer [1].

The results of this experiment are plotted in Figure 4, indicating the consumer's accuracy in classifying different groups of advisers. Specifically, the second approach with fixed $\beta = 1.0$ performs worst when the mean variation of the providers' performance is low. In contrast, the performance of the first approach with fixed $\beta = 0.5$ degrades as providers change their performance significantly. By having the ability to monitor the actual variations of the providers' performances, only the third and the last approach can maintain a high level of classification accuracy. However, the Prob-Cog model outperforms other approaches as it fairly achieves maximum classification accuracy by integrating the factor of behaviour in its evaluation.

The next experiment aims to highlight the influences of the cognitive view of trust in advisers' classifications. We learn that different $Inf_{view:cog}$ values yield different performances in various conditions. More explicitly, in case consumers deal with low-variance providers, a high value of $Inf_{view:cog}$ degrades the classification accuracy significantly. However, it shows better results as the providers' performance variation rose from 50% to 100%. As it is implied in Figure 5, assigning an average influence $Inf_{view:cog} = 0.5$ to the cognitive approach ensures consistent high performance throughout the experiment.

[1] In this experiment, we assume that none of the advisers are filtered in the first layer so that $ADC_{(C)} = 0$

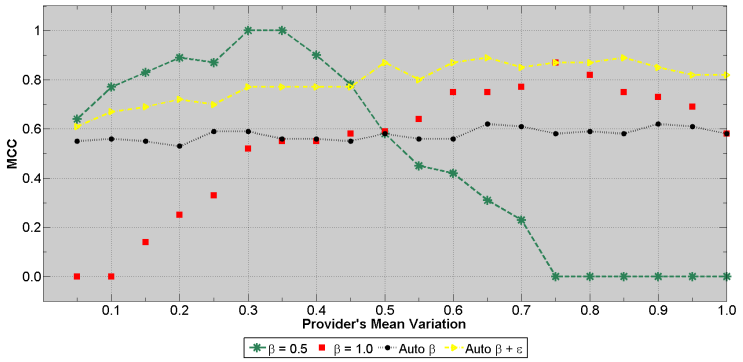

Fig. 4: Classification of advisers across different variations in providers' behaviours

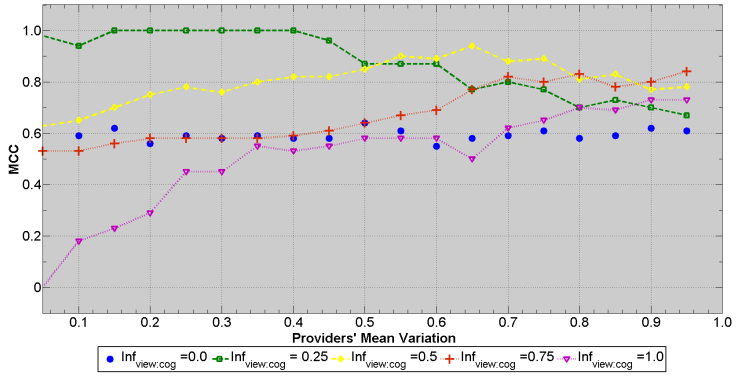

Fig. 5: Accuracy of advisers' classification having different cognitive preferences

5.5 Unbalanced Environment with Dishonest Majority

Trust models should be able to effectively cope with the problem of unfair ratings. They specifically should perform ideally in a situation where a majority of participants act dishonestly in an environment. In the following series of experiments we examine the performance of the personalized approach and the Prob-Cog model in classifying participants when 5% to 90% of them are dishonest. In this environment providers change their performance from 35% to 45%. We consider two patterns for dishonesty: 1) a deceitful pattern which includes 50% dishonest participants and 2) a complementary pattern which covers the remaining percentage of dishonest participants. We also assume that in both models consumer C has an adequate number of experiences[2] and advisers' ratings are provided in the same time window as consumer C. We adjust the $Inf_{view:cog}$ of the Prob-Cog model and the trustworthiness threshold of the personalized approaches to 0.5.

[2] Based on this assumption, the weight of the public knowledge component in the personalized approach is negligible.

Results are shown in Figure 6. We can see that the classification performance of the Prob-Cog model is higher than the personalized approach across different percentages of dishonest participants. This is mainly due to the static approach of the personalized approach in determination of the trustworthiness threshold. That is, it might happen that in the environment where providers vary their QoS, this model labels honest participants as dishonest. This issue shows its significant deficiency when the majority of dishonest participants prevail in the environment, resulting in the detection of fewer honest advisers.

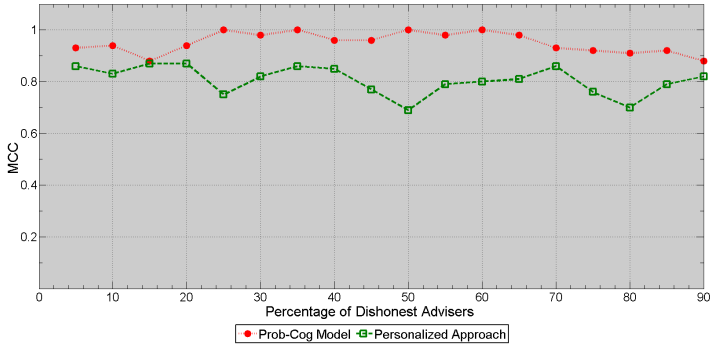

Fig. 6: The Prob-Cog model vs Personalized approach Performance

To better perceive the reasons behind the performance leaks, we compute the error rates (FPR and FNR) of these two models in detecting honest and dishonest participants dealing with different population tendency. Based on Figure 7, we observe that in the Prob-Cog model, when the majority of advisers are honest, the low value of $ADC_{(C)}$ in the first layer increases the effect of the cognitive dimension. As such, the high value of ϵ relatively amplifies the probability of misclassification of dishonest advisers as honest so that $FNR \geq 0$.

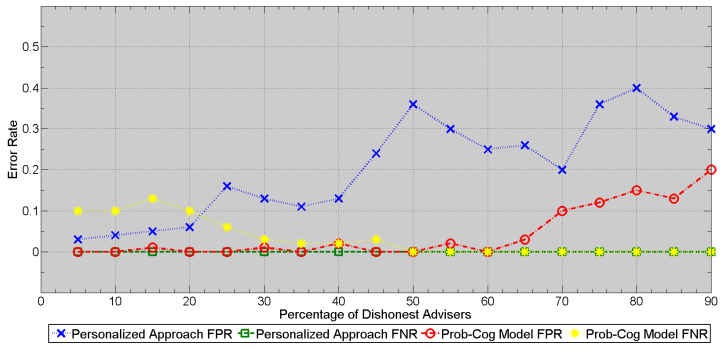

Fig. 7: The Error Rate of Prob-Cog model vs Personalized approach

On the other hand, in this model when a majority of advisers turn out to be dishonest, it adaptively reduces ϵ so as to degrade the influence of the cognitive dimension and behavioural modeling in trustworthiness evaluation. Even though this strategy helps consumer C to detect a high percentage of deceitful participants, there is a chance that honest advisers with low credibility will be

misclassified as dishonest, resulting in a reduction of the classification performance ($FPR \geq 0$).

In the personalized approach when C has sufficient personal experiences and the threshold is assigned to a low value, it can perfectly detect dishonest participants so that ($FNR = 0$). However, since this model does not capture the providers' variations, C would highly misclassify honest advisers as malevolent ($FPR > 0$). In these series of experiments we observed that with the employment of the Prob-Cog approach we are able to detect *more* honest advisers compared with other approaches.

6 Summary of the Results

We have carried out a set of experiments to compare overall performance of three representative approaches: FIRE, the personalized approach and the Prob-Cog model in different scenarios. We measure their accuracy in detecting honest advisers when a majority of advisers are unfair, providers vary their behaviours in different degrees and consumers lack in personal experiences. We notice that the Prob-Cog model performs the best as it is able to better classify advisers in the aforementioned situations in comparison with other approaches. We have shown that, owing to the limited observation of the environment, it is not a sensible idea to exploit public knowledge as the environment might be controlled by a majority of dishonest participants. We noticed that the Prob-Cog model could successfully differentiate honest participants from dishonest ones in the cold start problem. We also verified how our approach effectively detects advisers with insufficient experiences and reduces their competency degree proportionately.

7 Conclusion and Future Work

In this paper, we present an adaptive multi-layered filtering algorithm that enables consumers with different behavioural attitudes to subjectively evaluate the trustworthiness of a variety of advisers in an e-marketplace. In the Prob-Cog model, the genuine beliefs and behavioural characteristics of participants are cognitively modeled and integrated in their trustworthiness evaluation metrics.

The principles of the two-layer filtering algorithm detect and disqualify various types of participants such as: malicious agents with a complementary rating pattern, newcomers with insufficient experiences, and fraudulent participants who retain a minimum level of trust to cheat opportunistically. In the Prob-Cog model, consumers can dynamically adjust the influence of the cognitive view of trust pertaining to their behavioural patterns to amplify or reduce the effect of different dimensions on credibility measures. This model also provides consumers with a mechanism to adaptively determine the value of the thresholds' parameters based on the observations of the quality of providers and environmental conditions. To demonstrate the effectiveness and capabilities of our approach, we focused on the experimental comparison with two representative approaches: the personalized approach and FIRE. We specifically examined some prominent scenarios, including ones dealing with participants' lack of experience, advisers flooding consumers with lots of ratings, providers' dynamicity and an environment with a majority of dishonest participants. Such empirical studies are useful for highlighting the importance of the capabilities of our Prob-Cog model.

Notably, results indicate that through the proper adaptation of the employed thresholds, consumers are able to identify more honest advisers compared with other approaches in different environmental circumstances, specifically when a majority of participants are unfair and when reliable advisers are scarce. One possible avenue for future work is to develop a provider classification mechanism which exploits the Prob-Cog model to evaluate the qualification of the participating providers in an e-marketplace.

References

1. K. Barber, Karen Fullam, and Joonoo Kim. Challenges for trust, fraud and deception research in multi-agent systems. In *Trust, Reputation, and Security: Theories and Practice*, pages 167–174. 2003.
2. Cristiano Castelfranchi, Rino Falcone, and Michele Piunti. Agents with anticipatory behaviors: To be cautious in a risky environment. In *ECAI*, 2006.
3. Huynh T. D, Jennings N. R., and Shadbolt N. R. An integrated trust and reputation model for open multi-agent systems. *AAMAS*, pages 119–154, 2006.
4. Rino Falcone and Cristiano Castelfranchi. Generalizing trust: Inferencing trustworthiness from categories. In *AAMAS-TRUST*, pages 65–80, 2008.
5. Chung-Wei Hang, Yonghong Wang, and Munindar P. Singh. An adaptive probabilistic trust model and its evaluation. In *AAMAS (3)*, pages 1485–1488, 2008.
6. Reid Kerr and Robin Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *AAMAS (2)*, pages 993–1000, 2009.
7. Zeinab Noorian, Stephen Marsh, and Michael Fleming. Multi-layer cognitive filtering by behavioral modeling. In *Proceedings of the 10th international conference on Autonomous agents and Multiagent systems (AAMAS'11)*. ACM, 2011.
8. Zeinab Noorian and Mihaela Ulieru. The state of the art in trust and reputation systems: A framework for comparison. *J. Theor. Appl. Electron. Commer.*, 5(2), 2010.
9. Audun Jøsang and Roslan Ismail. The Beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.
10. W. T. L Teacy, J Patel, N. R. Jennings, and M. Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Journal of Autonomous Agents and Multi-Agent Systems*, 2006.
11. Yao Wang and Julita Vassileva. Bayesian network-based trust model. In *Web Intelligence*, pages 372–378, 2003.
12. Yonghong Wang and Munindar P. Singh. Formal trust model for multiagent systems. In *IJCAI*, 2007.
13. Li Xiong and Ling Liu. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering*, 16:843–857, 2004.
14. Bin Yu and Munindar P. Singh. Detecting deception in reputation management. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 73–80, 2003.
15. Jie Zhang and Robin Cohen. Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. *Electronic Commerce Research and Applications*, 2008.
16. Moti Zviling, Hadas Leonov, and Isaiah T. Arkin. Genetic algorithm-based optimization of hydrophobicity tables. *Bioinformatics*, 21(11):2651–2656.