On Handling Inaccurate Witness Reports

T. Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt

Intelligence, Agents, Multimedia Group, Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, UK
{tdh02r,nrj,nrs}@ecs.soton.ac.uk

Abstract. Witness reports are a key building block for reputation systems in open multi-agent systems in which agents, that are owned by a variety of stakeholders, continuously enter and leave the system. However, in such open and dynamic environments, these reports can be inaccurate because of the differing views of the reporters. Moreover, due to the conflicting interests that stem from the multiple stakeholders, some witnesses may deliberately provide false information to serve their own interests. Now, in either case, if such inaccuracy is not recognised and dealt with, it will adversely affect the function of the reputation model. To this end, this paper presents a generic method that detects inaccuracy in witness reports and updates the witness's credibility accordingly so that less credence is placed on its future reports. Our method is empirically evaluated and is shown to help agents effectively detect inaccurate witness reports in a variety of scenarios where various degrees of inaccuracy in witness reports are introduced.

1 Introduction

A wide variety of networked computer systems (such as the Grid, the Semantic Web, and peer-to-peer systems) can be viewed as multi-agent systems (MAS) in which the individual components act in an autonomous and flexible manner in order to achieve their objectives [3]. An important class of these systems are those that are *open*; here defined as systems in which agents can freely join and leave at any time and where the agents are owned by various stakeholders with different aims and objectives. From these two features, it can be assumed that in open MAS: (1) the agents are likely to be self-interested and may be unreliable; (2) no agent can know everything about its environment; and (3) no central authority can control all the agents. Given such uncertainties, trust is central to effective interactions between the agents [5]. Indeed this recognition accounts for the large number of recently developed models of trust and reputation (see [5] for a review). Although there are many differences in the way these models are implemented, the majority of them are built upon some form of *witness reports* (information about an agent's behaviour from a third-party). However, a key problem in this area, and one that is exacerbated in open MAS, is that these reports can be inaccurate. This can happen because of the differing views of the reporters. However, it can also happen due to the conflicting interests that stem from the multiple stakeholders (e.g. some witnesses may deliberately provide false information to serve their own interests). In both cases, witness reports that differ from the actual performance an agent receives are viewed as

inaccurate and their inaccuracy[1] is reflected by the magnitude of the differences. Now, since these reports are central building blocks for reputation systems, if their inaccuracy is not recognised and dealt with, it will adversely affect the function of these systems. Worse still, they may become a means for malicious agents to gain unwarranted trust which may then allow them to benefit to the detriment of others.

Given its importance, there have been several attempts to tackle the inaccurate witness reports problem (see Sect. 4), but none of them are well suited to open MAS. In particular, in order to operate as intended, they typically make assumptions about the target environment that are incompatible with the characteristics of open MAS or they require additional domain knowledge that clearly limits their applicability (see Sect. 4 for examples). To this end, we devise a witness credibility model that can be used by trust and reputation models in open MAS. In so doing, we advance the state of the art in the following ways. First, our model is able to recognise inaccurate reports based on an agent's own experience, and, therefore, no addition domain knowledge is required. As a result, an agent is able to keep track of the quality of a witness's reports and assess that witness's credibility accordingly. Second, based on witness credibility, our model provides a witness rating weight function so that the likely accuracy of witness reports can be taken into account when an agent's reputation is produced from them. Hence, lying witnesses can quickly be detected and their reports disregarded. Third, of particular relevance is the generic nature of our approach. Our model can be used for handling inaccurate reports in any trust model. Specifically, in this paper, its integration with the FIRE model [2] is given as an example of its usage, but the model is not restricted to this case. Since this paper focuses only on modelling and dealing with witness inaccuracy, we are not going to consider other design issues of a trust model. We simply assume that our model is to be applied in a working trust model that is able to evaluate trust based on experiences from direct interactions and to collect and evaluate witness reputation.

The remainder of the paper is organised as follows. In the next section, we will present our witness credibility model. The model will then be empirically evaluated in Section 3. Section 4 presents related work in the area. Finally, Section 5 concludes and outlines the future work.

2 Witness credibility model

As our witness credibility model will be integrated into FIRE in order to provide a concrete grounding for its operation, we will briefly recap the FIRE model in Sect. 2.1, before going onto present the witness credibility model in Sect. 2.2.

[1] It should be noted that in this context inaccuracy is according to the view of the agent receiving witness reports. It does not reflect the true honesty/accuracy of a witness agent. Rather it should be viewed as the subjective measure of the usefulness of witness information provided by that witness which is assessed by a particular agent.

2.1 The FIRE model

This section describes the Interaction Trust (IT) and the Witness Reputation (WR) components of FIRE[2]. The detailed description and justification for the various design choices that have been made are given in [2]. The IT and WR component although given here in the context of FIRE are broadly similar to a range of other trust components (see [5] for examples). Therefore, the use of our witness credibility model is not restricted to FIRE.

Interaction Trust. This is the trust that ensues from direct interactions between two agents. Specifically, it is derived from the experience of the agents about the performance or behaviour of their partners in those interactions. Such experience is recorded in the form of ratings which are tuples of the following form: $r = (a, b, i, c, v)$, where a and b are the agents that participated in interaction i, and v is the rating a gave b for the term c (e.g. price, quality, delivery). The range of v is $[-1, +1]$, where -1, $+1$, and 0 means absolutely negative, absolutely positive, and neutral respectively.

In order to calculate IT, an agent needs to record its past ratings in a (local) rating database which stores at maximum the H latest ratings the agent gave to each of its partners after interactions. Here H is called the *local rating history size*. When assessing the IT of agent b with respect to term c, agent a retrieves the set of relevant ratings (denoted by $\mathcal{R}_{\mathsf{I}}(a, b, c)$) from its rating database. The IT value of b, denoted by $\mathcal{T}_{\mathsf{I}}(a, b, c)$, is given by the following formula:

$$\mathcal{T}_{\mathsf{I}}(a, b, c) = \frac{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega_{\mathsf{I}}(r_i) \cdot v_i}{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega_{\mathsf{I}}(r_i)} \tag{1}$$

where $\omega_{\mathsf{I}}(r_i)$ is the rating weight function that calculates the relevance or the reliability of the rating r_i with respect to IT, and v_i is the value of the rating r_i.

Since older ratings may become out-of-date, recency of the ratings is used as the rating weight. Specifically, $\omega_{\mathsf{I}}(r_i)$ is a parameterised exponential decay function calculated from the time difference between the current time and the time when the rating r_i was recorded $\Delta t(r_i)$:

$$\omega_{\mathsf{I}}(r_i) = e^{-\frac{\Delta t(r_i)}{\lambda}} \tag{2}$$

where λ, the recency factor, is the parameter used to adjust the rating weight function to suit the time unit used in a particular application.

In FIRE, each trust value comes with a reliability value that reflects the confidence of FIRE in producing that trust value given the data it took into account. The reliability value is given based on the two following measures:

- $\rho_{\mathsf{RI}}(a, b, c)$ is a function that calculates the reliability of $\mathcal{T}_{\mathsf{I}}(a, b, c)$ based on the reliability (or the relevance) of all the ratings taken into account[3], which is given

[2] FIRE also has other trust components but we do not consider them in this paper because they are not affected by inaccurate witness reports.

[3] Here in ρ_{RI}, ρ stands for reliability, R for rating quality, and I for IT.

by the function $\omega_{\mathsf{I}}(r_i)$ (by its definition):

$$\rho_{\mathsf{RI}}(a,b,c) = 1 - e^{-\gamma_{\mathsf{I}} \cdot \left(\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega_{\mathsf{I}}(r_i)\right)} \tag{3}$$

where γ_{I} is a parameter used to adjust the slope of the reliability function to suit the rating weight function of each component.

– $\rho_{\mathsf{DI}}(a,b,c)$ is a reliability measure based on the variability of the rating values[4]. The greater the variability, the more volatile the other agent's behaviour. Hence, $\rho_{\mathsf{DI}}(a,b,c)$ is calculated as the deviation in the ratings' values:

$$\rho_{\mathsf{DI}}(a,b,c) = 1 - \tfrac{1}{2} \cdot \frac{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega(r_i) \cdot |v_i - \mathcal{T}_{\mathsf{I}}(a,b,c)|}{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega(r_i)}, \tag{4}$$

Finally, the reliability of $\mathcal{T}_{\mathsf{I}}(a,b,c)$, denoted as $\rho_{\mathsf{I}}(a,b,c)$, combines the two reliability measures above:

$$\rho_{\mathsf{I}}(a,b,c) = \rho_{\mathsf{RI}}(a,b,c) \cdot \rho_{\mathsf{DI}}(a,b,c) \tag{5}$$

Witness Reputation. This is built on observations about the behaviour of the target agent b by others (witnesses). Thus, in order to evaluate the WR of b, an agent a needs to find the witnesses that have interacted with b. In this component, FIRE employs a variant of the referral system in [8] to find such witnesses. In that system, each agent maintains a list of acquaintances (other agents that it knows). Then when looking for a certain piece of information, an agent can send the query to a number of its acquaintances who will try to answer the query if possible or, if they cannot, they will send back referrals pointing to other agents that they believe are likely to have the desired information (see [2] for more details).

After the set of witness ratings is collected through the referral process, they will then be aggregated using the same method as per the IT component into a WR value (denoted by $\mathcal{T}_{\mathsf{W}}(a,b,c)$). The accompanying reliability value, denoted by $\rho_{\mathsf{RW}}(a,b,c)$, and the rating weight function, $\omega_{\mathsf{W}}(r_i)$, are also defined as per the IT component. This means that accurate and inaccurate ratings are treated equally. However, as discussed in Sect. 1, this means malicious agents may take advantage of this to gain unwarranted trust. In order to prevent this type of exploitation, unreliable witnesses should be detected and disregarded, and this requires us to devise a new rating weight function for the WR component that takes into account the credibility of witnesses (see Sect. 2.2).

Overall trust. This is produced by combining the trust values from all the various components of FIRE to give an overall picture of an agent's likely performance. The composite trust value (denoted by $\mathcal{T}(a,b,c)$) and its reliability (denoted by $\rho_{\mathcal{T}}(a,b,c)$) are calculated as follows:

$$\mathcal{T}(a,b,c) = \frac{w_{\mathsf{I}} \cdot \mathcal{T}_{\mathsf{I}}(a,b,c) + w_{\mathsf{W}} \cdot \mathcal{T}_{\mathsf{W}}(a,b,c)}{w_{\mathsf{I}} + w_{\mathsf{W}}} \tag{6}$$

[4] Here, D in ρ_{DI} stands for the deviation of the rating values.

$$\rho_{\mathcal{T}}(a,b,c) = \frac{w_{\mathsf{I}} + w_{\mathsf{W}}}{W_{\mathsf{I}} + W_{\mathsf{W}}} \tag{7}$$

where $w_{\mathsf{K}} = W_{\mathsf{K}} \cdot \rho_{\mathsf{K}}(a,b,c)$, and W_{I} and W_{W} are the coefficients corresponding to the IT and WR components. These coefficients are set by end users to reflect the importance of each component in a particular application.

2.2 Witness credibility

The credibility of a witness in reporting its ratings about another agent can be derived from a number of sources. These include knowledge about: the relationships between the witness and the rated agent (e.g. cooperating partners may exaggerate each other's performance, competing agents may underrate their opponents, no relationship may imply impartial ratings); the reputation of the witness for being honest and expert in the field in which it is doing the rating (e.g. a reputable and independent financial consultant should provide fair ratings about services of various banks); the relationships between the witness and the querying agent (e.g. agents with the same owner should provide honest reports to one another); norms in the witness's society (e.g. doctors usually recommend a drug to the benefits of patients, rather than to the benefit of its pharmaceutical companies) and so on. Unfortunately, however, these types of information are very application specific and may not be readily available in many cases. Therefore, although they could certainly be used to enhance the precision of a witness credibility measure, they are not suitable as a generic basis (although they could complement a generic measure in particular contexts).

In contrast, in our witness credibility model we view providing witness reports as a service an agent provides. Thus its performance (i.e. trustworthiness and reliability) can be evaluated and predicted by a trust model. By so doing, the credibility model can benefit from a trust model's (usually sophisticated) ability of learning and predicting an agent's behaviour (in this case, the behaviour of providing accurate reports) without having to implement its own method. Here, we use FIRE's IT component for this purpose.

In more detail, after having an interaction with agent b, agent a records its rating about b's performance, denoted by r_a ($r_a = (a, b, i_a, c, v_a)$). Now, if agent a received witness reports from agent w, it then rates the credibility of w by comparing the actual performance of b (i.e. v_a) with w's rating about b. The smaller the difference between the two rating values, the higher agent b is rated in terms of providing witness reports (mutatis mutandis for bigger differences). For each witness rating that a received from w in evaluating the WR of b (denoted by $r_k = (w, b, i_k, c, v_k)$), the credibility rating value v_w for agent w is given in the following formula:

$$v_w = \begin{cases} 1 - |v_k - v_a| & \text{if } |v_k - v_a| < \iota \\ -1 & \text{if } |v_k - v_a| \geq \iota \end{cases} \tag{8}$$

where ι is the inaccuracy tolerance threshold ($0 \leq \iota \leq 2$, 2 is the maximal difference since $v_k, v_a \in [-1, 1]$). Thus if the difference between a witness rating value and the actual performance is higher than ι, the witness is considered to be inaccurate or lying, and, therefore, receives a negative rating of -1 for its credibility. On the other hand,

if the difference is within the tolerance threshold, it can be viewed as resulting from a subjective viewpoint and is deemed acceptable. In this case, the credibility rating value v_w is set to be inversely proportional to the difference (e.g. higher difference, lower credibility). Since $0 \leq |v_n - v_a| \leq 2$, v_w is also in the range $[-1, 1]$ regardless of ι. The rating about w's credibility — $r_w = (a, w, i_w, c_{WC}, v_w)$ — is then recorded in a's rating database, where c_{WC} is the rating term of providing witness reports and i_w is the interaction of agent w providing agent a the rating r_k about agent b.

Here, as agents whose inaccuracy exceeds the tolerance threshold are considered lying and are heavily fined (by giving a -1 credibility rating), honest witnesses may be falsely penalised if their (honest) ratings are too different from that of a (e.g some agents may have substantially varying levels of performance which result in varying, though honest, ratings). However, since in the case of acceptable inaccuracy ($|v_k - v_a| < \iota$) an agent's credibility is also penalised according to the degree of its inaccuracy (i.e. $|v_k - v_a|$), it is always safe to set ι to higher values to reduce the probability of falsely classifying honest witnesses. Nevertheless, doing so inevitably allows ratings from marginally lying witnesses be accepted. In such cases, although the credibility of such witnesses may be low, it may never be low enough, or it may take a longer time, for them to be considered lying and be disregarded (see below). This means, in general, that there will be a lower performance of the witness credibility model. Therefore, it is important to choose a threshold value that closely reflects the variability of the agents' performance in the target environment and the relative costs of considering lying witnesses versus falsely classifying them.

Given the availability of credibility ratings from the above procedure, the process of the WR component is redesigned to make use of these refined ratings. Specifically, after the referral process, for each witness rating collected from a witness w, the WR component uses the IT component to evaluate the interaction trust of w in terms of providing witness reports. The IT component, in turn, calculates the credibility trust value of w from ratings about w's credibility (retrieved from the rating database as per Sect. 2.1). If no such ratings has been recorded, and, thus, the IT value is not available, the WR component will assign the default credibility trust value, denoted by $\mathcal{T}_{DWC}$, to witness b.

$$\mathcal{T}_{WC}(a, w) = \begin{cases} \mathcal{T}_{\mathsf{I}}(a, w, c_{WC}) & \text{if } \mathcal{R}_{\mathsf{I}}(a, w, c_{WC}) \neq \emptyset \\ \mathcal{T}_{DWC} & \text{otherwise} \end{cases} \tag{9}$$

where $\mathcal{T}_{WC}(a, w)$ is the credibility of w evaluated by a. The credibility of witness w is then used to calculate the weights of ratings provided by w in calculating the WR value of the target agent b.

Hence, the rating weight function for the WR component, instead of being defined as in Equation 2, is redefined here as follows. Suppose that agent a is evaluating the WR of agent b and that the rating r_i is collected from witness w, then:

$$\omega_{\mathsf{W}}(r_i) = \begin{cases} 0 & \text{if } \mathcal{T}_{WC}(a, w) \leq 0 \\ \mathcal{T}_{WC}(a, w) \cdot \omega_{\mathsf{I}}(r_i) & \text{otherwise} \end{cases} \tag{10}$$

This means the new rating weight function disregards any rating provided by witnesses that have negative credibility (by giving 0 as the weight for their ratings). The rest are taken into account in producing the WR of b, but are weighted by the credibility of their

providers and by their recency (provided by the function $\omega_{\mathsf{I}}(r_i)$ of the IT component). In so doing, ratings from the more accurate witnesses (as judged by the accuracy of their past ratings) make a bigger impact on the WR value than those from the less accurate ones. In cases where all the witness ratings collected are disregarded, due to negative credibility of their providers, the WR component will produce no trust value (as in the case where it fails to collect witness ratings). In addition, at first, every witness receives the default credibility value since it has not provided witness ratings to agent a before. Hence, end users can set the value of $\mathcal{T}_{DWC}$ to reflect their policy towards newly encountered witnesses. For example, $\mathcal{T}_{DWC}$ can be set to 0 so that newly encountered witnesses are disregarded until they prove to be credible (by providing ratings in the acceptable accuracy threshold) or it can be set to 1 to reflect the policy that all witnesses are considered to be accurate and honest until proven otherwise.

3 Empirical evaluation

In order to empirical evaluate our new witness credibility model, we use the test bed designed in [2] with minor changes to simulate inaccuracy of witness reports. In particular, the witness credibility model will be tested under various levels of witness accuracy to determine its efficiency in filtering out inaccurate reports. The testbed is described in Sect. 3.1. The methodology and experimental settings are then presented in Sect. 3.2. Finally, Sect. 3.3 describes the experiments and discusses their results.

3.1 The testbed

The testbed is a multi-agent system consisting of agents providing services (called *providers*) and agents using those services (called *consumers*). Without loss of generality, it is assumed that there is only one type of service in the testbed. Hence, all the provider agents offer the same service. However, their performance (i.e. the quality of the service) differs. The agents are situated randomly on a spherical world whose radius is 1.0. Each agent has a *radius of operation* r_o that models its capability in interacting with others (e.g. the available bandwidth or the agent's infrastructure) and any agents situated in that range are its neighbours.

Simulations are run in the testbed in rounds (of agent interactions), and the round number is used as the time unit. In each round, if a consumer agent needs to use the service it can contact the environment to locate nearby provider agents (in terms of the distance between the agents on the spherical world). The consumer agent will then select one provider from the list to use its service. The selection process relies on the agent's trust model to decide which provider is likely to be the most reliable. Consumer agents without a trust model randomly select a provider from the list. The consumer agent then uses the service of the selected provider and gains some utility from the interaction (called UG). The value of UG is in $[-10, 10]$ and depends on the level of performance of the provider in that interaction. A provider agent can serve many users at a time.

After an interaction, the consumer agent will rate the service of the provider based on the level of performance it received. It records the rating for subsequent trust evaluations. It is also willing to share those recorded ratings (witness reports) when asked.

However, since the witness may alter its actual ratings before giving them to the querying agent, we model this phenomenon by introducing five types of witnesses. Agents in the Hon group always reveal their actual ratings truthfully. Those in Neg1 and Neg2, however, provide to the querying agent ratings that are lower than those they actually recorded. Conversely, those in Pos1 and Pos2 give falsely higher ratings. The difference between an actual rating value and its inaccurate one in Neg1 and Pos1 is randomly set in the range $[0.3, 1.0]$ (i.e. representing marginally inaccurate witnesses) and the respective range of Neg2 and Pos2 is $[1.0, 2.0]$ (i.e. representing extremely inaccurate witnesses). In the testbed, we also define levels of witness inaccuracy at the level of the overall system (-100 to 100) so that various configurations of *witness population* can be conveniently referred in our experiments. The proportions of witness types in each level of witness inaccuracy are given in Fig. 1 (where vertical dotted lines represent specific example configurations of witnesses). For example, level 0 means that all the witnesses are of the Hon group (and provide their ratings honestly). Level $+80$ means that in the witness population, the proportions of Hon, Pos1, and Pos2 are 20%, 40%, and 40% respectively. It also means that 80% of the witness population is providing positively exaggerated reports (because 80% of the witnesses are either Pos1 or Pos2). Similarly, level -60 means that 60% of the witness population (30% Neg1, 30% Neg2) is providing falsely negative reports.

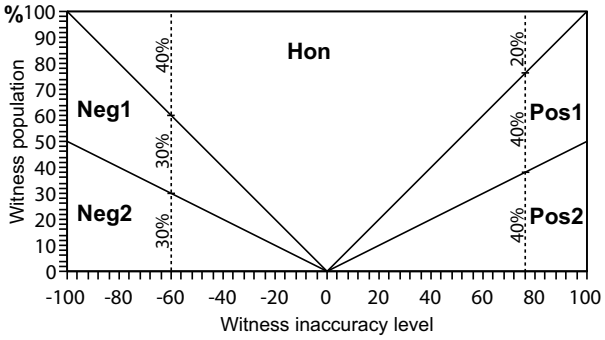

Fig. 1. The proportions of witness types at various levels of witness inaccuracy.

In our testbed, besides the accuracy level of witness reports, the only difference in each situation is the performance of the provider agents. Here, we consider three types of provider agents: good, ordinary, and bad. Each of them has a mean level of performance (μ_{P}). Its actual performance follows a normal distribution around this mean. The values of μ_{P} and the associated standard deviation (σ_{P}) of these types of providers are given in Table 1. In addition, the service quality of a provider is also degraded linearly in proportion to the distance between it and the consumer to reflect the greater uncertainties associated with service delivery (e.g. lower service quality resulting from increased delays or losses in information exchanges between two agents when they are far away from each other). Hence, from the same provider, each consumer may receive a different level of service quality depending on its location. This means honest ratings about that provider's performance by its consumers can be different; reflecting the phenomenon that every agent has its own context making its own view subjective.

Profile	Range of μ_{P}	σ_{P}
Good	[PL_GOOD, PL_PERFECT]	1.0
Ordinary	[PL_OK, PL_GOOD]	2.0
Bad	[PL_WORST, PL_OK]	2.0

Performance level	Utility gained
PL_PERFECT	10
PL_GOOD	5
PL_OK	0
PL_BAD	−5
PL_WORST	−10

Table 1. Profiles of provider agents.

3.2 Experimental methodology

In each experiment, the testbed is populated with provider and consumer agents. Each consumer is equipped with a particular trust model, which helps it select a provider when it needs to use a service. Since the only difference among consumer agents is the trust models that they use, the utility gained (UG) by each agent will reflect the performance of its trust model in selecting reliable providers for interactions. Hence, the testbed records the UG of each interaction along with the trust model used. In order to obtain an accurate result for performance comparisons between trust models, each one will be employed by a large number of consumer agents ($\mathsf{N_C}$). In addition, the average UG of agents employing the same trust models (called consumer groups) are compared with each other's using the two-sample t-test [1] (for means comparison) with the confidence level of 95%. The result of an experiment is then presented in a graph with two y-axes (see Fig. 2 for an example); the first plots the UG means of consumer groups in each interaction and the second plots the corresponding performance rankings obtained from the t-test (prefixed by "R.", where the group of rank 2 outperforms that of rank 1). The experimental variables are presented in Table 2 and these will be used in all experiments unless otherwise specified.

Simulation variable	Symbol	Value
Number of simulation rounds	N	500
Total number of provider agents:	$\mathsf{N_P}$	95
+ Good providers	$\mathsf{N_{PG}}$	10
+ Ordinary providers	$\mathsf{N_{PO}}$	40
+ Bad providers	$\mathsf{N_{PB}}$	45
Number of consumers in each group	$\mathsf{N_C}$	500

Table 2. Experimental variables.

Now, there are three groups of consumers in each experiment: one employing FIRE (IT and WR components only), one employing the SPORAS model[5] (see Sect. 4 for details), and one consisting of agents with no trust model. We name the three groups WR, SPORAS, and NoTrust. A summary of the parameters of FIRE is provided in Table 3. The component coefficients W_{I} and W_{W} are selected to reflect the belief that trust values produced by the IT component are more reliable than those produced from witness ratings (which are prone to inaccuracy). The recency factor λ is selected such

[5] SPORAS is chosen as the control benchmark for two reasons. First, it is a successful independently developed trust model which several other researchers have used for benchmarking. Second, other than SPORAS, most notable trust models make assumptions that are incompatible with open MAS, and, thus, they will not operate as intended in our testbed (see Sect. 4 for more details).

that a 5-time-unit-old rating will have a recency weight of 0.5 (to suit the time unit used in the testbed). The default witness credibility $\mathcal{T}_{DWC}$ is set to 0.5 so that all ratings from newly encountered witnesses will be taken into account in calculating WR but their weights are smaller than that of any proven accurate witness (which is typically greater than $\iota = 0.5$, see Equation 9) and larger than that of a proven inaccurate one (which is typically negative). The value of ι is handpicked based on the actual variability of honest rating values in the testbed (which never exceeds 0.5).

Parameters	Symbol	Value
Local rating history size	H	10
Recency factor	λ	$-\frac{5}{\ln(0.5)}$
Component coefficients:		
+ Interaction trust	W_{I}	2.0
+ Witness reputation	W_{W}	1.0
Reliability function parameters:		
+ Interaction trust	γ_{I}	$-\ln(0.5)$
+ Witness reputation	γ_{W}	$-\ln(0.5)$
Witness credibility parameters:		
+ Default witness credibility	$\mathcal{T}_{DWC}$	0.5
+ Inaccuracy tolerance threshold	ι	0.5

Table 3. FIRE's parameters.

3.3 The effect of inaccurate reports

In this section, we will look at how various levels of witness inaccuracy affect the performance of each consumer group. In this experiment, inaccurate witnesses always provide inaccurate reports. The experiment is run with the following witness inaccuracy levels: -100, -80, ..., 0, ..., $+80$, $+100$. After plotting the performance of the three consumer groups in each experiment (Fig. 2), it can be seen that the performance of NoTrust is consistently low (around -1.0). On the other hand, thanks to the trust models used, the performance of SPORAS and WR are always higher than that of NoTrust. However, the performance of WR is always superior to that of SPORAS. In this experiment, since the performance of all providers are equally exaggerated, it is still the case that good providers generally have better ratings than others. Hence, in the first few interactions, they are selected by WR, and this accounts for an increase of WR's UG in this period. Now, after several interactions with these providers, WR is able to record their actual performance, which is generally lower than the reported performance of the remaining providers (calculated only from falsely positive reports). Thus, remaining providers (i.e. ordinary and bad providers) are then selected in later interactions. As a result, WR's UG is decreased. Nevertheless, because of the witness credibility model, during these interactions, WR is able to realise that all reports are inaccurate, and, thus, future (false) reports are disregarded. Effectively, WR resorts to the IT component for evaluating providers' trustworthiness. As for SPORAS, since it cannot filter out inaccurate reports, it cannot improve its performance over time. Now, due to space constraints we can only present a general analysis of the large number of experiments and associated settings that were conducted during the evaluation of this work.

In more detail, the chart in Fig. 3 shows the average performance of the three groups in each experiment with various levels of witness inaccuracy. Here, the average performance of each group is calculated as the average utility gained in each interaction of an agent in that group. This average performance is calculated from data of the first 200 interactions in each experiment (by which time the average UG of all groups is stable in all experiments).

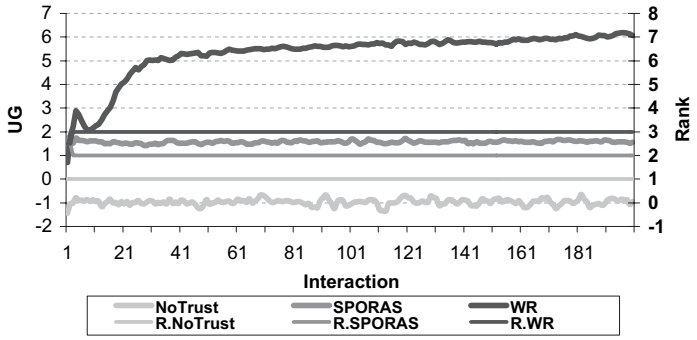

Fig. 2. Performance of NoTrust, SPORAS, and WR at witness inaccuracy level +100.

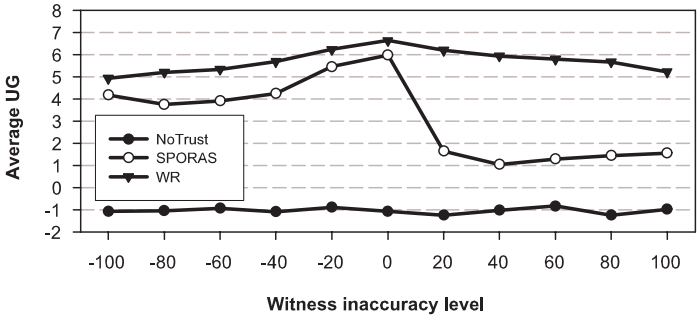

Fig. 3. Performance of NoTrust, SPORAS, and WR at various levels of inaccuracy.

From Fig. 3, it can be seen that the performance of both WR and SPORAS suffer as the witness inaccuracy increases (as we would expect). However, the performance of WR is more robust. In particular, SPORAS suffers greatly from exaggerated positive ratings (because of the reason mentioned above). On the other hand, although WR also suffers from false positive ratings at the beginning (see Fig. 2), it can gradually learn and disregard inaccurate witnesses and, generally speaking, it maintains a high performance. In the cases of falsely negative ratings (see Fig. 3, witness inaccuracy level -100 to -20), SPORAS does not suffer as much as in the cases of exaggerated positive ratings. The reason is that falsely negative ratings not only lower the rating values of good providers, but also lower those of bad and ordinary ones by similar amounts. Hence, it is still the case that good providers have better reputations in SPORAS, and, thus, they are more likely to be selected for interaction. This means that SPORAS can perform normally in scenarios where all lying witnesses provide negative ratings. However, this is because of the nature of that specific lying population rather than SPORAS's ability

of dealing with inaccurate reports. As for WR, as mentioned above, its ability to detect and penalise inaccurate witnesses also works in such scenarios and allows WR to maintain a generally high performance.

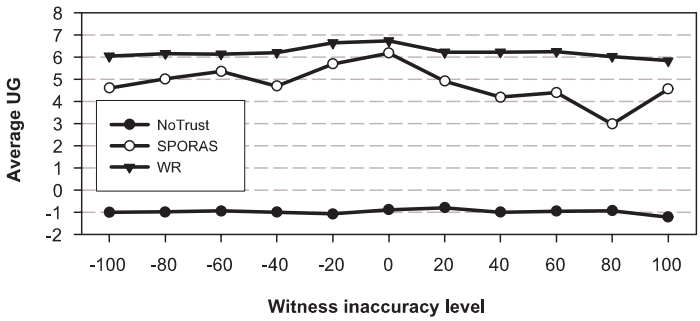

Fig. 4. Lying 25% of the time.

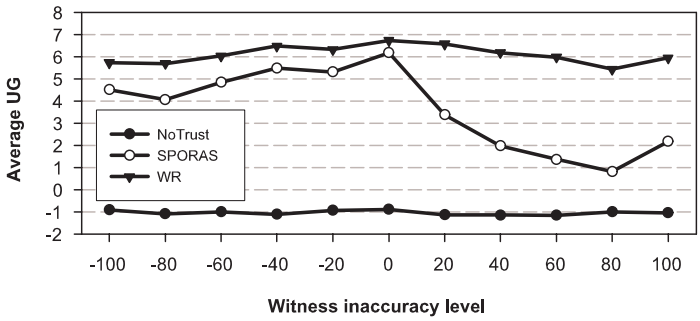

Fig. 5. Lying 75% of the time.

Next, we seek to determine whether our model can cope with situations where witnesses have more subtle lying behaviour — they lie sometimes and provide their honest ratings at other times. Here, we investigate two scenarios: the lying witnesses provide false ratings: (1) 25% of the time (i.e. being mostly honest, lying sometimes) and (2) 75% of the time (i.e. mostly lying, being honest sometimes). The two cases are interesting because some agents may try to fool a reputation system by lying only a few times to maintain their credibility (case 1) or by giving reports honestly to increase its (bad) credibility (case 2). The set of experiments are rerun for the two scenarios and their results are presented in Figures 4 and 5. From these, it can be seen that the performance of all groups have a broadly similar pattern in scenarios of negative lying. However, as in the scenarios of lying 100% of the time, SPORAS suffers adversely from exaggerated positive reports (as in the previous experiment). It can also be seen that SPORAS's performance decreases in proportion to the amount of lying (i.e 25%, 75%, 100%). In contrast, in all scenarios presented, our witness credibility model can learn the witnesses' lying behaviour (thanks to the adaptive nature of FIRE's IT component), and this accounts for the robust performance of WR throughout in these scenarios.

4 Related work

Many trust and reputation models have been devised in recent years due to the increasing recognition of their roles in controlling social order in open systems [5]. Now, SPORAS [9] is one of the notable models. In this model, each agent rates its partner after an interaction and reports its ratings to the centralised SPORAS repository. The ratings received are then used to update the global reputation values of the rated agents. The model uses a learning function for the updating process so that the reputation value can closely reflect an agent's performance. In addition, it also introduces a reliability measure based on the standard deviations of the rating values. However, it has been designed without considering the problem of inaccurate reports and so it suffers disproportionately when false information is collected (as shown in Sect. 3). Speaking more generally, as many trust models are built on witness reports, the unreliable reporting problem has come to the fore of several recent works on trust and we explore the most notably approaches in the remainder of this section.

Regret [6] models a witness' credibility based on the difference between that witness' opinion and an agent's past experience. In particular, before taking a witness opinion into account, agent a compares that witness opinion with the direct trust it has on the target agent based on its previous experiences (i.e. Interaction Trust in FIRE). If both the witness and agent a are confident about their opinions on the target agent, the difference between the two opinions (i.e the direct trust of a and w on b) is used to determine the credibility of that witness. The higher the difference, the less credible the witness is. Then the witness's credibility is used to weight that witness's opinion in aggregating it with those of other witnesses. Now, although this method appears to be similar to our approach, there are two important differences. First, suppose agent a receives an opinion from witness w about agent b, in order determine w's credibility for future interactions, our model compares w's opinion with the actual performance of b *after* a interacts with b, not with a's past experiences about b before the interaction (as in Regret). This allows agent a to determine the credibility of w without depending on the availability of its past experiences with b[6]. More importantly, in our model, in the case that b changes its behaviour and a's experience about b becomes out-of-date, w's credibility will not be judged based on out-of-date information (as it would be if Regret is used). Second, our model views providing witness information as a normal service and assess a witness' performance (i.e. its credibility) in all its (witness information providing) interactions whenever its information is used. By so doing, a witness' credibility is deduced from the history of its service, not just a single assessment as in the case of Regret. Alternatively, Regret also uses fuzzy rules to deduce the weight for each witness rating. However, agents are assumed to have a social network that models the social relationships in the agent's world. Then, in order to determine the trustworthiness of a witness, Regret applies simple rules over the relationships between the witness and

[6] In addition, Regret requires that both a and w are confident about their assessment of b's performance in order to determine the credibility of w. This, in turn, requires a and w to have a sufficient number of ratings about b. In our opinion, if a is already confident about its assessment of b's performance, witness reputation is less important. This view is also adopted by Regret considering its method of combining direct trust and its various types of reputation.

the target agent (e.g. cooperating or competing). Thus, this approach requires significantly more knowledge about the agent environment than our model and, moreover, no evaluation has been presented to demonstrate the effectiveness of their method.

Jurca and Faltings [4] attempt to eliminate the lying witness problem by introducing a reputation mechanism such that the agents are incentivised to share their ratings truthfully. However, in so doing, they also introduce a payment scheme for witness reports that requires an independent monetary system and this may not always be available in the context we consider. Moreover, their reputation information is managed by third-party agents which cannot be guaranteed to be entirely impartial due to the self-interested character of an open MAS.

In Whitby et al.'s system [7] the "true" rating of an agent is defined by the majority's opinions. In particular, they model the performance of an agent as a beta probability density function (PDF) which is aggregated from all witness ratings received. Then a witness is considered unreliable and filtered out when the reputation derived from its ratings is judged to be too different from the majority's (by comparing the reputation value with the PDF). Since this method bases it decisions entirely on PDFs of witness reports, if these reports are scarce and/or too diverse it will not be able to recognise lying witnesses. Moreover, it is possible that a witness can lie in a small proportion of their reports without being filtered out.

Yu and Singh [8] propose a similar approach to that of Whitby et al. Specifically, they use a weighted majority algorithm to adjust the weight for each witness over time. Although the weights of the deceitful agents are reduced, these agents are never disregarded completely. Several successful applications of this approach have been demonstrated, but only for agent populations where deceitful agents are in the minority and are balanced between negative and positive lying agent.

In sum, in contrast to the above mentioned models, our witness credibility model does not require additional domain knowledge, makes no assumptions about the environment, and is evaluated in a broad range of scenarios.

5 Conclusions and future work

This paper has presented a novel model of witness credibility which is designed to recognise inaccurate reports and ensure that they are disregarded in making assessments about reputation. Moreover, our model also ensures accurate witnesses are recognised and that their information is highly valued. Through empirical evaluation, we show that our model helps FIRE filter out inaccurate reports and perform robustly in a variety of scenarios of inaccuracy. Moreover, our witness credibility model can be adjusted to suit an agent's stance/policy to witness reports via its parameters (i.e. the default witness credibility and the inaccuracy tolerance threshold). This means it is suitable for a wide range of applications.

As our model considers witness credibility as an agent's performance in providing witness reports, it can be used in any trust model that is able to predict an agent's performance based on performance ratings. Obviously, this ability is the intended function of almost all trust models, and, as a result, our model can be used alongside most existing trust models. Specifically, in this paper, FIRE's IT component is employed only

as a means so that our model can learn about the lying behaviour of witnesses, and its integration with FIRE is to provide a concrete grounding. It can equally well be plugged into other models in a similar way.

In future work, we aim to devise a method to automatically adjust the accuracy tolerance threshold during the system's operation (instead of handpicking a value as at present). This can be achieved by analysing the recorded performance levels of service providers that an agent has interacted with to determine the likely variability of honest ratings. By so doing, the inaccuracy can be adjusted to closely reflect the actual variability of performance levels in an agent's environment, and, as a result, the precision of inaccuracy detection is improved (e.g. more marginal lying cases can be detected, and honest witnesses will not be falsely classified as lying because of an increased fluctuation in a provider's performance).

References

1. P. R. Cohen. *Empirical Methods for Artificial Intelligence*. The MIT Press, 1995.
2. T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. Developing an integrated trust and reputation model for open multi-agent systems. *Proc. of the 7th Int. Workshop on Trust in Agent Societies*, 2004.
3. N. R. Jennings. An agent-based approach for building complex software systems. *Communications of the ACM*, 44(4):35–41, April 2001.
4. R. Jurca and B. Faltings. An incentive compatible reputation mechanism. In *Proceedings of the IEEE Conference on E-Commerce CEC03*, Jun 24-27 2003.
5. S. D. Ramchurn, T. D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1), 2004.
6. J. Sabater. *Trust and Reputation for Agent Societies*. Phd thesis, Universitat Autònoma de Barcelona, 2003.
7. A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. *Proc. of the 7th Int. Workshop on Trust in Agent Societies*, 2004.
8. B. Yu and M. P. Singh. Detecting deception in reputation management. In *Proc. of the 2nd Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*. ACM Press, 2003.
9. G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9), 2000.