Certified Reputation: How an Agent Can Trust a Stranger

Trung Dong Huynh Nicholas R. Jennings Nigel R. Shadbolt

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

ABSTRACT

Current computational trust models are usually built either on an agent's direct experience of an interaction partner (interaction trust) or reports provided by third parties about their experiences with a partner (witness reputation). However, both of these approaches have their limitations. Models using direct experience often result in poor performance until an agent has had a sufficient number of interactions to build up a reliable picture of a particular partner and witness reports rely on self-interested agents being willing to freely share their experience. To this end, this paper presents *Certified Reputation* (CR), a novel model of trust that can overcome these limitations. Specifically, CR works by allowing agents to actively provide third-party references about their previous performance as a means of building up the trust in them of their potential interaction partners. By so doing, trust relationships can quickly be established with very little cost to the involved parties. Here we empirically evaluate CR and show that it helps agents pick better interaction partners more quickly than models that do not incorporate this form of trust.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Design, Reliability, Experimentation

Keywords

Trust, Reputation, Multi-Agent Systems

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. 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 an open MAS: (1) the agents are likely to be self-interested and may not always complete tasks that are requested of them; (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 [7]. Indeed, this recognition accounts for the large number of computational models of trust (here defined as the subjective probability with which an agent a assesses that another agent b will perform a particular action, both before a can monitor such action and in a context in which it affects its own action [2]) that help agents to determine the most reliable interaction partner (see Section 4 for more details).

Although there are many differences in the way these models are implemented, the majority of them are built on an agent's direct experience of an interaction partner (interaction trust) or reports provided by third parties about their experiences with a partner (witness reputation). However, both these approaches have their shortcomings. First, when an agent first enters an environment, it has no history of interactions (with the other agents in that environment). In such circumstances, if its trust model is based solely on direct experience, it would need to explore the environment by interacting with other agents to learn about their performance. However, in so doing, the agent also inevitably risks making losses if it encounters unreliable partners. Moreover, because it can learn about only one agent per interaction, trust models using direct experience typically require a long time to be able to achieve stable performance. Second, models based on witness reports usually implicitly assume the altruism of agents in sharing their experiences. Now, this cannot be guaranteed in all cases because self-interested agents are unlikely to be willing to sacrifice their resources in order to provide witness reports. Moreover, in the distributed and open environments we consider, the required witnesses for any given agent can be difficult to locate. This is usually addressed by using some form of centralised mechanism to collect all the witness reports [13] or implementing a distributed search process to look for witnesses in an agent's social network [12]. However, the former is not compatible with an open MAS since agents representing different owners may well question the trustworthiness of a central authority and the latter may well involve high costs of time

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.

and resources to locate witnesses.

Against this background, in this paper we present a novel type of trust called *certified reputation* (CR) whose mechanism overcomes the limitations of current trust models described above. Hence, CR can be used complementarily to other sources of trust information (e.g. direct experience, witness reports) to build a versatile composite trust model (which benefits from all the advantages of its components). In more detail, the CR of an agent is the reputation that is derived from third-party references about its previous performance. Agents that adopt the CR mechanism will actively collect and present such references in order to seek the trust of their potential partners. This, in turn, moves the burden of obtaining and maintaining trust information from the trust evaluator to the agent being evaluated (who is incentivised to do so)[1]. Hence, trust information (references) about a particular agent becomes available to those who want to interact with it. Moreover, agents using CR have neither to interact first with the agent being evaluated (as they do with direct experience), nor to locate its witnesses (as they do with witness reports). The potential downside, however, is that third-party references might not be reliable (since referees can collude with particular agents by providing falsely inflated references for them and negative ones for the others). Now, this is an inherent problem of using any type of third-party information and so it is important to be able to evaluate the credibility of a referee and to use this to weigh its information. To this end, we also develop such a model that records the history of a referee's performance in terms of providing accurate references (by comparing an agent's actual observations with the references it receives). Such histories of referees are then used to derive their credibility.

In so doing, we advance the state of the art in the following ways. First, we develop a new process of obtaining information for trust evaluation. This process addresses the inherent shortcomings of interaction trust (the lack of direct experience) and witness reputation (the difficulty in finding witness reports). Second, as CR allows agents to be able to evaluate trust themselves, without relying on a central mechanism, it is compatible with a wide range of open and distributed environments. Third, from empirical evaluation, it is shown that using CR improves an agent's utility gain by helping it to quickly pick good interaction partners. Finally, it is shown that our model is robust against various types of collusion.

The rest of the paper is organised as follows. Section 2 presents our model of certified reputation. This model is then empirically evaluated in Section 3. Section 4 presents related work and Section 5 concludes.

2. CERTIFIED REPUTATION

A trust model is typically used by an agent, say agent a, to evaluate the trustworthiness of another agent, say agent b, when it considers establishing an interaction with b. In this case, we call a the *evaluator* and b the *target.* Now, most trust evaluation is based on past experience of the target agent's performance. Such experience is here 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. For instance, the quality of a news provider can be rated in terms of topicality, quality, and honesty. Here, the range of v is $[-1, +1]$, where -1, $+1$, and 0 means absolutely negative, absolutely positive, and neutral respectively. Each agent has a local *rating store* to collect the ratings it makes and the ones that it receives from others. Since ratings are context-dependent (described in the rating terms), each trust evaluation is also about a specific term. Here, we use $\mathcal{T}(a, b, c)$ to denote the trustworthiness of b in terms of c that is evaluated by a. In order to make such a trust evaluation, a needs to obtain a set of relevant ratings. We call this set $\mathcal{R}(a, b, c)$.

Now, having defined the basic notions, we turn to the CR model itself. The mechanism of CR is given in Section 2.1, which also describes how CR is calculated. Then Section 2.2 shows how the credibility of referees is evaluated.

2.1 The Mechanism of Certified Reputation

Certified reputation of a target agent b consists of a number of certified references[2] about its behaviour on particular tasks that are provided by third-party agents. Such information is obtained and stored by the target agent itself and made available to any other agent that wishes to evaluate its trustworthiness for further interactions (somewhat like a reference when a person is applying for a job). The agents giving references are called the *referees.* Here, references are in the form of ratings given by b's partners about its performance in (past) interactions between them. These ratings allow b to prove its (achievable) performance as viewed by its previous interaction partners and then to gain the trust of its potential partners. However, since b can choose which ratings to put forward, a rational agent will only present its best ones. Therefore, it should be assumed that CR information probably overestimates an agent's expected behaviour. Thus, although it cannot guarantee b's minimal performance in future interactions, the CR information does reveal a partial perspective on agent b's capabilities (which is certainly useful for trust evaluation in the absence of other sources of information).

Though CR may have lower predictive power than the other types of trust/reputation (where all bad and good ratings can be collected), it is useful because of its wide applicability. With the cooperation of its partners, agent b can obtain their references from just a small number of interactions[3]. From our evaluation, for instance, in a society where 100 agents provide a service to 500 others, agents using direct experience to evaluate trust require more than 100 interactions to achieve a reasonable level of utility gain,

[1] It can reasonably be assumed that agents want to have more interactions because of their potential utility gain. Hence, agents have incentives to provide trust information about themselves so that they can gain the trust of other agents which, in turn, will facilitate more interactions.

[2] It is assumed that some form of security mechanism (such as a public-key infrastructure) is employed to ensure that the provided references cannot be tampered with. For instance, all references could be accompanied by digital signatures from the issuers using their private keys [14]. By so doing, any change to a reference will be easily detected. Digital signatures are also a means to verify the references' origins.

[3] In many scenarios, such as those in the Internet, established service providers (e.g. news service, online merchants) usually have high volumes of interactions (at any time). Therefore, if they adopt the CR process outlined here, we can reasonably expect that such providers will have an abundance of performance ratings readily available.

which is still less than that achieved by agents using CR after 5 interactions (see Section 3 for more detail). In addition to its high availability, since references are stored by the target agent and provided directly to the evaluator, CR has a very low running cost (i.e. time, communication, processing cost) compared to sources like witness reputation.

In more detail, the process of CR is as follows:

- After every transaction, agent b asks its partners to provide their certified ratings about its performance from which it can choose the best ratings to store in its (local) rating store.
- When agent a contacts b to express its interest in using b's service, it asks b to provide references about its past performance with respect to an interested term c.
- Agent a receives the set of certified ratings of b from b, which we call $\mathcal{R}_{\mathsf{C}}(a, b, c)$ (C to denote this set is obtained via the CR mechanism), and calculates the CR of b based on this set.

In this process, since agent b relies on its interaction partner's cooperation to get references, agents may refuse to give out their ratings (as in the case of witness reputation). However, this is a much smaller problem than that in witness reputation because this information is requested far less frequently (each referee is requested to give its rating only once). Moreover, giving such information could be made a standard part of any agreement for task allocation and so agents could be forced to give it. The most notable point in this process is that when agent a makes the trust evaluation, it only involves agents a and b. Since the certified ratings about b are stored by b itself, they are immediately available to a as in the case when a uses its own experience.

Having obtained the references of b, a can estimate b's future behaviour, or more specifically, the expected rating value b is likely to receive in a future interaction. A common way to estimate that value is to calculate it as the arithmetic mean of all the rating values in the set. However, these ratings are usually not equally relevant when estimating the expected rating value. For example, some ratings may be older than others and, thus, may be out-of-date; and some may come from a more reliable source that suggests a higher level of credibility compared to others. Therefore, like many contemporary models [8, 10, 12], we use a *rating weight function* $\omega_{\mathsf{C}}(r_i)$ ($\omega_{\mathsf{C}}(r_i) \geq 0$) which calculates the relevance of each given certified rating r_i. Then, instead of considering all ratings equally, the trust value is calculated as the weighted mean of all the ratings available[4]:

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

where $\mathcal{T}_{\mathsf{C}}(a, b, c)$ is the CR value of b that agent a calculates with respect to term c, which is calculated from the rating set $\mathcal{R}_{\mathsf{C}}(a, b, c)$, and v_i is the value of the rating r_i. In short, the CR value is calculated as the sum of all the available ratings weighted by the rating relevance and normalised to the range of $[-1, 1]$ (by dividing the sum by the sum of all the weights).

Now, we need to determine the relevancy of a given rating ($\omega_{\mathsf{C}}(r_i)$). Since an agent's environment may change rapidly, resulting in out-of-date ratings, we use the recency of ratings as one measure of their relevance. Specifically, the recency relevance of a rating r_i is calculated by an exponential decay function based on its recency [5]:

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

where $\omega_{\mathsf{Re}}(r_i)$ is the relevance value for the rating r_i in terms of recency, $\Delta t(r_i)$ is the time difference between the current time and the time when the rating r_i is recorded, and λ is the recency factor which is used to scale $\Delta t(r_i)$ according to a particular application's time unit.

As certified ratings are digitally signed, their authenticity can be verified and their content cannot be tampered with. However, since there is no guarantee about the honesty of agents in an open MAS, a referee can still collude with the target agent and provide falsely inflated references about its performance. Moreover, even if the referee is honest in providing references, its references can be inaccurate (either because of its incapability of making accurate ratings or because it has a different view to that of the evaluator). In either case, these inaccurate references will result in inaccurate CR values. Therefore, we need measures to prevent or to minimise the adverse effects of such references. To this end, we use the credibility of referees as another measure of rating relevance. In more detail, for a reference r_i from referee w, w's credibility value is calculated by a (called $\mathcal{T}_{\mathsf{RCr}}(a, w) \in [-1, 1]$, RCr denotes referee credibility). Then, the relevance of r_i in terms of referee credibility (denoted by $\omega_{\mathsf{RCr}}(r_i)$) is defined as follows:

$$\omega_{\mathsf{RCr}}(r_i) = \begin{cases} 0 & \text{if } \mathcal{T}_{\mathsf{RCr}}(a, w) \leq 0 \\ \mathcal{T}_{\mathsf{RCr}}(a, w) & \text{otherwise} \end{cases} \tag{3}$$

Both the above measures of rating relevancy are then taken into account in weighing a rating:

$$\omega_{\mathsf{C}}(r_i) = \omega_{\mathsf{RCr}}(r_i) \cdot \omega_{\mathsf{Re}}(r_i) \tag{4}$$

It should be noted that, from Equations 3 and 4, references whose referees have negative credibility are discarded (by setting $\omega_{\mathsf{C}}(r_i) = 0$). Only references whose referees have positive credibility are taken into account when calculating CR. In such cases, they are weighted both by their referees' credibility and by their recency. In so doing, ratings from the more credible referees make a bigger impact on the CR value than those from the less credible ones. In cases where all the certified ratings collected are disregarded, due to negative credibility of their providers, no trust value will be produced. In such circumstances, agent a can give b a default low trust value (e.g. -0.5) because b fails to provide reliable references. Otherwise, agent a can re-request b for more reliable references.

2.2 Referee Credibility

The credibility of a referee in reporting its ratings about another agent can be derived from a number of sources. These include knowledge about: the relationships between the referee 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 referee 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 the services of various banks); the relationships between the referee and the querying agent (e.g.

[4] We choose the weighted mean method here because it allows us to take the relevance of each rating into account. Other aggregation methods could equally well be used if desired.

agents with the same owner should provide honest reports to one another); and so on. Unfortunately, however, such types of knowledge 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 referee credibility measure, they are not suitable as a generic basis (although they could complement a generic measure in particular contexts). Therefore, here, we base our solution on a modified version of the witness credibility model we devised for evaluating witness reputation in [4] to assess a referee's credibility. We choose this model because it does not require such domain-specific information about relationships and because providing witness reports and references are of the same broad nature. Specifically, in this model, we view providing references 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 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.

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 a reference (i.e. a certified rating) 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 w is rated in terms of providing accurate references (mutatis mutandis for bigger differences). For each certified rating that a received from w in evaluating the CR 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:

$$v_w = \begin{cases} 1 - |v_k - v_a| & \text{if } |v_k - v_a| < \iota \\ -1 & \text{if } |v_k - v_a| \geq \iota \end{cases} \qquad (5)$$

where ι is called 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 certified rating value and the actual performance is higher than ι, the referee 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 varied 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 — $r_w = (a, w, i_w, \text{term}_{\mathsf{RCr}}, v_w)$ — is then recorded by a, where $\text{term}_{\mathsf{RCr}}$ is the rating term for performance in providing references and i_w is the interaction of agent w providing agent a the certified rating r_k about agent b.

Having recorded ratings about w's performance on providing references, a can evaluate w's credibility based on those ratings. As mentioned above, a can use its own trust model for so doing. Specifically, here we calculate a's trust on w's capability of providing accurate references similarly as per CR (Equation 1), except that we use only ratings retrieved from a's ratings store (instead of obtaining them via the CR mechanism) and weigh the ratings only by their recency[5] (Equation 2). This type of trust is here called interaction trust (IT) since it is based solely on a's direct experience from its interactions. The IT value of a on b in terms of c is denoted by $\mathcal{T}_{\mathsf{I}}(a, b, c)$ and is calculated as follows:

$$\mathcal{T}_{\mathsf{I}}(a, b, c) = \frac{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega_{\mathsf{I}}(r_i) \cdot v_i}{\sum_{r_i \in \mathcal{R}_{\mathsf{I}}(a,b,c)} \omega_{\mathsf{I}}(r_i)} \qquad (6)$$

where $\omega_{\mathsf{I}}(r_i)$ is the rating weight of r_i and $\omega_{\mathsf{I}}(r_i) = \omega_{\mathsf{Re}}(r_i)$. Then, in order to determine the referee credibility of w ($\mathcal{T}_{\mathsf{RCr}}(a, w)$), we need to calculate $\mathcal{T}_{\mathsf{I}}(a, b, \text{term}_{\mathsf{RCr}})$. This is calculated from the set of $\text{term}_{\mathsf{RCr}}$ ratings — $\mathcal{R}_{\mathsf{I}}(a, b, \text{term}_{\mathsf{RCr}})$. However, if no such rating has been recorded, and, thus, the IT value is not available, we assign w the default referee credibility trust value, denoted by $\mathcal{T}_{\mathsf{DRCr}}$:

$$\mathcal{T}_{\mathsf{RCr}}(a, w) = \begin{cases} \mathcal{T}_{\mathsf{I}}(a, w, \text{term}_{\mathsf{RCr}}) & \text{if } \mathcal{R}_{\mathsf{I}}(a, w, \text{term}_{\mathsf{RCr}}) \neq \emptyset \\ \mathcal{T}_{\mathsf{DRCr}} & \text{otherwise} \end{cases} \qquad (7)$$

It should be noted here that, at first, every referee receives the default credibility value since it has not provided references to agent a before. Hence, end users can set the value of $\mathcal{T}_{\mathsf{DRCr}}$ to reflect their policy towards newly encountered referees. For example, $\mathcal{T}_{\mathsf{DRCr}}$ can be set to 0 so that newly encountered referees are disregarded until they prove to be credible (by providing ratings in the acceptable accuracy threshold) or it can be set to 1 so that all referees are considered to be accurate and honest until proven otherwise.

3. EMPIRICAL EVALUATION

In order to empirical evaluate our CR model, we use the testbed designed in [5] with a few changes. In more detail, in order to verify our intuitions about CR, it will be evaluated in ideal (honest) environments where agents do not collude with each other (Section 3.3). It will then be evaluated in (biased) environments where colluding agents give high ratings to each other (Section 3.4) to determine its robustness for open multi-agent contexts. Before doing this, however, a brief description of the testbed is given in Section 3.1 (see [5] for a complete description and for justifications of the choices we made in it) and Section 3.2 presents the methodology and experimental settings for our experiments.

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). In addition, each consumer agent has a maximum number of friend providers (N_{FP}) that it may collude with when providing references about their performance. Such providers are selected randomly from a consumer's nearby providers when it enters the testbed.

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

[5] Since a makes the ratings itself, weighing them in terms of referee credibility is irrelevant.

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 consumers at a time.

After an interaction, the consumer agent will rate the service of the provider based on the level of performance it received and provide the rating to the provider to be potentially used as a reference. A provider will select $\mathsf{N_R}$ best references from those that it receives to store and to present when requested. Since a provider knows very old references are unlikely to be valued highly by its consumers, when selecting references to store, it takes into account both a reference's rating value (i.e. saying how well the provider performed) and the time when that reference is made. In particular, when comparing two references r_1 and r_2, the value of r_1 is biased by an amount of $\frac{t_1-t_2}{\mathsf{TIMESCALE}}$, where t_i is the time that reference r_i is given, and TIMESCALE is selected to be 20.0 given the time unit used in the testbed[6].

In addition, since a referee may make biased references, we model this phenomenon by introducing five types of referees. Agents in the Hon group always give out their actual ratings as references. *Exaggerating referees* in groups Exag1 and Exag2, however, give falsely higher ratings than those they actually recorded for their friend providers[7] (and their actual ratings for the others). In addition to giving falsely inflated references for their friends, *extreme referees* in Extr1 and Extr2 also deliberately underrate the other providers. The difference between an actual rating value and its inaccurate one in Exag1 and Extr1 is randomly set in the range $[0.3, 1.0]$ (i.e. representing marginally inaccurate referees) and the respective range of Exag2 and Extr2 is $[1.0, 2.0]$ (i.e. representing extremely inaccurate referees).

In our testbed, the only difference in each interaction situation is the performance of the provider agents. Here, we consider four types of providers: good, ordinary, bad, and intermittent. Each of them, except the last, has a mean level of performance, denoted by 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. Intermittent providers, on the other hand, yield unpredictable (random) performance levels in the range [PL_BAD, PL_GOOD]. 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 different.

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 reflects 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 a confidence level of 95%. The result of an experiment is then presented in a graph with two y-axes (see Fig. 1 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:	$\mathsf{N_P}$	100
+ Good providers	$\mathsf{N_{PG}}$	10
+ Ordinary providers	$\mathsf{N_{PO}}$	40
+ Bad providers	$\mathsf{N_{PB}}$	45
+ Intermittent providers	$\mathsf{N_{PI}}$	5
Number of consumers in each group	$\mathsf{N_C}$	500
Max. number of friend providers	$\mathsf{N_{FP}}$	4

Table 2: Experimental variables.

Now, in each experiment, we include several groups of consumers in order to compare their performance. These include one group employing CR, one employing the SPORAS model[8] (see Section 4 for details), one employing IT[9],

[6] This is obviously just one way of comparing two references but it is a way that is effective in our system. If we wanted to make rating recency less relevant then we could make TIMESCALE smaller, or if we wanted to make rating recency more relevant then we could make it bigger.

[7] This is motivated by examples where referees provided exaggerated reports about their friends.

[8] SPORAS is chosen as the control benchmark because it is a successful, independently developed trust model which several other researchers have used for benchmarking.

[9] IT is defined in Equation 6. It is calculated from a con-

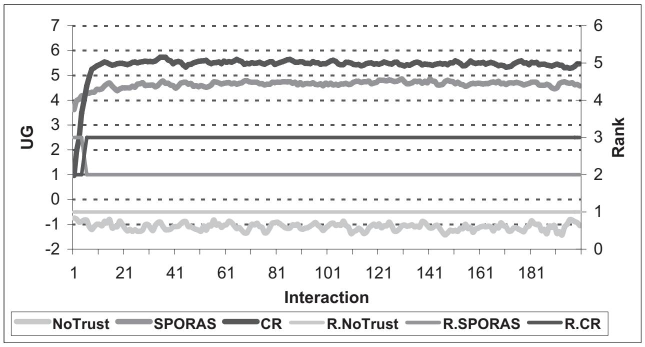

Figure 1: Performance of CR vs SPORAS and NoTrust.

and one consisting of agents with no trust model. We name these groups CR, SPORAS, IT and NoTrust. A summary of the parameters of CR is provided in Table 3. The recency factor λ is selected such 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 referee credibility $\mathcal{T}_{\mathsf{DRCr}}$ is set to 0.5 so that all ratings from newly encountered referees will be taken into account in calculating CR, but their weights are smaller than those of any proven accurate referee (which are typically greater than $\iota = 0.5$, see Equation 7) 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
Recency factor	λ	$-\frac{5}{\ln(0.5)}$
Number of stored references	$\mathsf{N_R}$	10
Referee credibility parameters:		
+ Default referee credibility	$\mathcal{T}_{\mathsf{DRCr}}$	0.5
+ Inaccuracy tolerance threshold	ι	0.5

Table 3: CR's parameters.

3.3 Honest Environments

Having defined the testbed and the evaluation methodology, we now turn to the experiments themselves. The first thing to test is whether CR helps consumer agents select profitable providers (i.e. those yielding positive UG) from the population and, by so doing, helps them gain better utility than without using CR. Hence, in this experiment we evaluate the performance of CR against that of NoTrust. SPORAS is also included in this experiment to compare the performance of CR with that of an independent benchmark that does not use CR. Here, Fig. 1 shows that the NoTrust group, selecting providers randomly without any trust evaluation, performs consistently the lowest (as we would expect). On the other hand, both SPORAS and CR prove to be beneficial to consumer agents, helping them obtain significantly high UG. This shows that the tested trust models can learn about the provider population and allow their agents to select profitable providers for interactions. In particular, SPORAS, being a centralised service, is able to gather ratings about all interactions in the system. This allows agents using it to achieve high performance right from the first interactions. In contrast, since each provider only shows a small number of ratings to agents using CR, they spend the first few interactions learning about their environment. Hence their initially lower performance than that of SPORAS. However, agents using CR quickly catch up with those in SPORAS (after 5 interactions) and they then maintain a higher stable performance thereafter (after the first 5 interactions, the average UG per interaction of SPORAS and CR are 4.65 and 5.48, respectively). The t-test also confirms that this difference is statistically significant as the ranks of SPORAS and CR switch after the first 5 interactions. Here, it should be noted that consumers may have different views on the same provider's performance because of the different distances between them and the provider (see Section 3.1). Therefore, although all the agents are honest, their ratings about that provider can be different due to their particular experience. Now, SPORAS takes all ratings equally and, thus, such differences are merely noise to it. This is the main reason that limits its performance. In contrast, the referee credibility model of CR compares each referee's ratings with an agent's own ratings and gives higher weights to those referees that have similar views to it. By so doing, the relevancy of each rating is determined and taken into account in calculating CR, making it more accurate than SPORAS.

Having shown that CR can outperform a centralised trust model based on witness reports, we now test CR against IT, which uses an agent's direct experience. Here, there are two groups of consumer agents, IT and CR, using IT and CR, respectively, as their trust models. The t-test results in Fig. 2 shows that CR outperforms IT in all interactions. The chart also shows that while it only takes agents using CR 5 interactions to achieve a stable level of performance, agents using IT need more than 100 interactions to obtain a reasonably stable level of performance. This is clearly because agents in IT do not have adequate ratings to quickly learn about their environment (i.e. the performance of providers). Moreover, even the maximum UG of IT (4.85) is lower than the average UG of CR (around 6.0). The reason is that agents using IT can miss identifying the best providers because ratings about them are less available than in the case of CR and because agents in our testbed will stick to choosing the highest trusted provider after a period of learning about the environment[10] (resulting in locally sub-optimal provider selections). Meanwhile, agents using CR always receive the most recent and the best references that are actively presented by providers. Moreover, the high availability of these references helps CR greatly increase its learning speed and, thus, avoid the problems associated with IT.

In summary, we have shown that in honest environments, CR provides a robust trust measure that is also highly available. However, we now go onto consider the more realistic situations in which collusion can occur.

3.4 Biased Environments

In this section, we evaluate CR in a wider range of environments. In particular, the consumers in each experiment consist of honest referees and those from one of the four colluding referee groups (Exag1, Exag2, Extr1, and Extr2 — see Section 3.1). For example, Fig. 3 presents the result from the experiment where the consumers consist of 20% honest referees and 80% colluding referees from the Extr2 group.

sumer's ratings about a provider's performance which that consumer collects in its rating store after every interaction.

[10] Agents in our testbed use a Boltzmann exploration strategy to explore (i.e. try the service of) unknown providers.

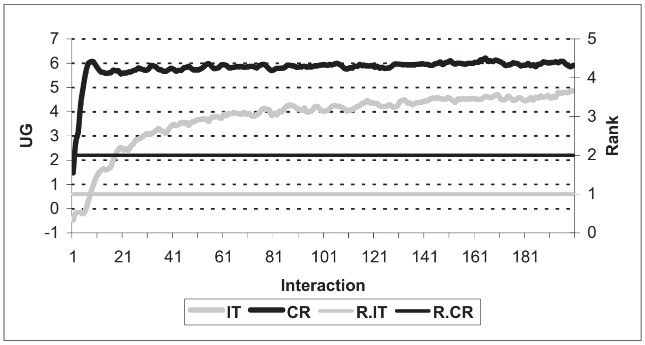

Figure 2: Performance of CR vs IT.

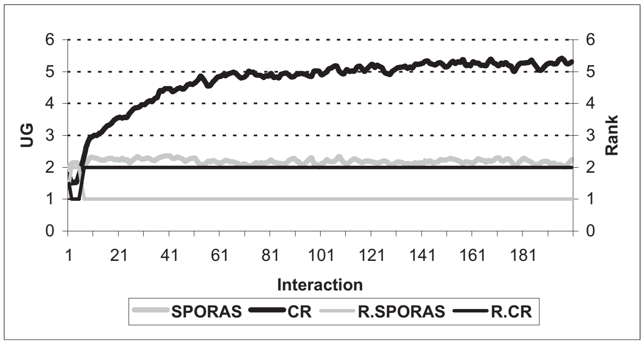

Figure 3: 80% **Extr2 referees.**

Since NoTrust performs consistently poorly in all the experiments, its results are omitted from our charts for the sake of simplicity. Moreover, because of space constraints, we cannot provide a detailed result of every experiment as in Fig. 3. Instead, we plot the average UG per interaction of SPORAS and CR in each experiment on the summary charts in Figs. 4 and 5. In these charts, the plots are named as GroupName.RefereeType. For example, SPORAS.Extr1 is the plot for the average UG of agents in the SPORAS group when the colluding referees in the testbed are of type Extr1.

The first thing these charts show is that collusion adversely affects the performance of trust models (as we would expect). For instance, Fig. 3 shows that it takes longer for CR to reach a stable level of performance (i.e. to learn about the colluding agents) and this performance is also lower than that in an honest environment. In such circumstances, SPORAS also yields a low UG and, without a credibility model, it cannot improve its performance over time. We can also see that the average performance of both SPORAS and CR generally decreases when the number of colluding agents increases (Figs. 4 and 5). However, CR always outperforms SPORAS except in the case when 100% of the consumers are Extr2 agents. In this particular experiment, since all consumers are identified by CR as lying (because all of them provide highly distorted references for all the providers), CR stops using their references, thus, depressing performance. However, this particular case (i.e. 100% extreme collusion) is highly unlikely to happen in practice (and if it did one might just retract trust from the population altogether). In the remaining experiments, CR can easily detect referees providing highly distorted references and maintain a generally high performance (see plots CR.Exag2 and CR.Extr2). CR is less effective in filtering out the colluding agents in Exag1 and Extr1 since their colluded references are less distorted than in the case of Exag2 and Extr2 (i.e. more difficult to detect lying). This suggests that the inaccuracy tolerance threshold ι should be carefully selected to reflect the nature of biased behaviours in a particular environment, or better, learning techniques could be used to enable an agent to adjust this parameter according to the prevailing context.

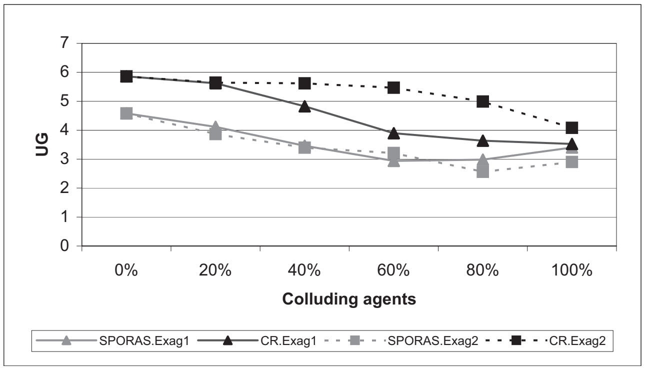

Figure 4: Performance with exaggerating referees.

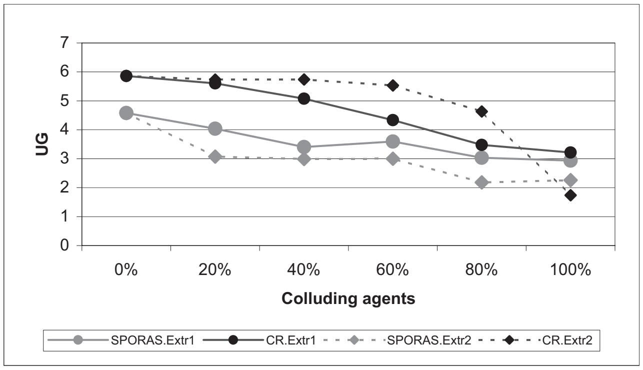

Figure 5: Performance with extreme referees.

In sum, this section shows that the credibility model of CR enables it to outperform SPORAS in dealing with biased behaviours. Specifically, it allows agents using CR to maintain a robust and high performance in a variety of cases, especially when the level of collusion is less than 50% (which is, in our opinion, the most likely case in realistic 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 [7]. SPORAS [13] is one of the most notable of these models. In this model, each agent rates its partner after an interaction and reports its ratings to the centralised SPORAS repository. The received ratings 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 Section 3).

Speaking more generally, as many trust models are built on witness reports, the problem of disinformation has come to the fore in several recent works on trust. Regret [8] models a witness' credibility based on the difference between that witness' opinion and an agent's past experience. This differs to our approach in that it depends on the availability of an agent's past experience. Moreover, this approach cannot deal with the situations where the target agent's behaviour changes since it does not take the new behaviour into account in its witness credibility assessment. Thus, it can falsely punish honest witnesses (who report the target agent's new behaviour). In Whitby et al.'s system [11], 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). Due to the dependency on PDFs of witness reports, if these reports are scarce and/or too diverse it is not able to recognise lying witnesses. Moreover, it is possible that a witness can lie in a small proportion of their reports without being detected. In addition, isolated honest opinions (i.e. different than that of the majority) can be falsely punished. To rectify this, TRAVOS uses an approach that is similar to ours [10]. However, as it uses beta PDFs for representing trust derived from binary outcomes (i.e. 0 for 'failed', 1 for 'success'), it is not suitable for our CR model because we require a more fine-grained and continuous range for trust values. In our earlier work [5], we presented a preliminary model of CR. However, this model could not cope with collusion and did not take the variance of referees situations (see Section 3.3) into account.

The mechanism of CR has similarities to trust policy management engines such as PolicyMaker [3] and Trust-Serv [9]. These engines grant rights (i.e. trust) to an agent based on its certificates of its identity according to predefined policies (i.e. rules, such as 'if a is a registered user and it possesses a valid credit card then it can book flights'). However, these engines are designed to determine the access rights of agents, rather than to determine the expected performance of these agents (i.e. how the agents behave after they can access a particular service). Similarly, the PGP web of trust model [14] allows a person to prove the authenticity of his public key by others' digital signatures, but not that person's behaviour. Certified ratings are also similar to the concept of endorsements in [6] — certificates endorsing that a service (provider) is trusted and preferred by their issuers. However, such endorsements are one-value ratings and therefore cannot show how good that service is (as our certified ratings can by having values in $[-1, +1]$ in a context sensitive way using rating terms). Moreover, the work in [6] does not consider the problem of collusion.

5. CONCLUSIONS AND FUTURE WORK

This paper has presented a novel mechanism for a new type of trust — certified reputation. This model provides a number of advantages over current approaches. First, its mechanism addresses the problem of the lack of direct experience (since agents can typically collect a large number of references themselves and they are incentivised to present these to establish new trust relationships). Second, agents are freed from the various costs involved in locating witness reports (e.g. resource, time, and communication costs). Third, since CR allows agents to evaluate trust for themselves it does not require a centralised service and, thus, is compatible with open multi-agent environments. In addition, in our evaluation we have shown CR performs significantly better than both a direct experience only model and a successful centralised trust mechanism. We have also shown that our CR model is robust against various types of collusion between agents.

In the future, 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. With respect to the CR mechanism itself, we plan to elevate the current presenting references process into a full dialog between the evaluator and the target agent. In so doing, the evaluator can specify what type of reference it wants to see (e.g. only the most recent ones, only references about transactions of high values, or those from a selected set of referees). This dialog can be a full negotiation process in which the evaluator can ask why a reference is bad and the target agent can explain a plausible reason for that. This will further allow both agents to learn more about each other, which clearly facilitates the trust evaluation process.

6. REFERENCES

[1] P. R. Cohen. *Empirical Methods for Artificial Intelligence.* The MIT Press, 1995.

[2] D. Gambetta. *Trust: Making and Breaking Cooperative Relations.* Dept. of Sociology, University of Oxford, 2000.

[3] T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Comm Surveys & Tutorials*, 3(4), 2000.

[4] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. On handling inaccurate witness reports. In *Proc. 8th Int. Workshop on Trust in Agent Societies*, pages 63–77, 2005.

[5] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of AAMAS*, 2006. (in press).

[6] E. M. Maximilien and M. P. Singh. Reputation and endorsement for web services. *ACM SIGEcom Exchanges*, 3(1):24–31, 2002.

[7] S. D. Ramchurn, T. D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, March 2004.

[8] J. Sabater. *Trust and Reputation for Agent Societies.* PhD thesis, Universitat Autònoma de Barcelona, 2003.

[9] H. Skogsrud, B. Benatallah, and F. Casati. Model-driven trust negotiation for web services. *IEEE Internet Computing*, 7(6):45–52, 2003.

[10] 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 *Proc. 4th Int Joint Conf on AAMAS*, pages 997–1004, 2005.

[11] A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proc. 7th Int Workshop on Trust in Agent Societies*, 2004.

[12] B. Yu and M. P. Singh. Detecting deception in reputation management. In *Proc. 2nd Int Joint Conf on Autonomous Agents and Multi-Agent Systems*, pages 73–80, 2003.

[13] G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14(9):881–908, 2000.

[14] P. R. Zimmermann. *The Official PGP Users Guide.* MIT Press, Cambridge, MA, 1995.