Coping with Inaccurate Reputation Sources: Experimental Analysis of a Probabilistic Trust Model

W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, Michael Luck
Electronics & Computer Science, University of Southampton
Southampton SO17 1BJ, UK.
{wtlt03r, jp03r, nrj, mml}@ecs.soton.ac.uk

ABSTRACT

This research aims to develop a model of trust and reputation that will ensure good interactions amongst software agents in large scale open systems. The following are key drivers for our model: (1) agents may be self-interested and may provide false accounts of experiences with other agents if it is beneficial for them to do so; (2) agents will need to interact with other agents with which they have little or no past experience. Against this background, we have developed *TRAVOS* (Trust and Reputation model for Agent-based Virtual OrganisationS) which models an agent's trust in an interaction partner. Specifically, trust is calculated using probability theory taking account of past interactions between agents. When there is a lack of personal experience between agents, the model draws upon reputation information gathered from third parties. In this latter case, we pay particular attention to handling the possibility that reputation information may be inaccurate.

Categories and Subject Descriptors

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

General Terms

Design, Measurement, Experimentation

Keywords

Trust, Reputation, Probabilistic Trust

1. INTRODUCTION

Computational systems of all kinds are moving toward large-scale, open, dynamic and distributed architectures, which harbour numerous *self-interested* agents. The Grid is perhaps the most prominent example of such an environment, but others include pervasive computing and the Semantic Web. In of all these environments, the concept of self-interest, is endemic and introduces the possibility of agents interacting in a way to maximise their own gain (perhaps at the cost of another). It is therefore essential to ensure good interactions between agents so that no single agent can take advantage of others. In this sense, good interactions are those in which the expectations of the interacting agents are fulfilled; for example, if the expectation of one agent is recorded as a contract that is then satisfactorily completed by its interaction partner, it is a good interaction.

In particular, we view the Grid as a multi-agent system (MAS) in which autonomous software agents, owned by various organisations, interact with each other. In particular, many of the interactions between agents are conducted in terms of Virtual Organisations (VOs), which are collections of agents (representing individuals or organisations), each of which has a range of problem-solving capabilities and resources at its disposal. A VO is formed when there is a need to solve a problem or provide a resource that a single agent cannot address. Here, the problem of assuring good interactions between individual agents is further complicated by the size of the Grid, and the large number of agents and interactions between them. Nevertheless, solutions to these problems are integral to the wide-scale acceptance of the Grid and agent-based VOs [4].

It is now well established that computational *trust* is important in such open systems [10]. Specifically, trust provides a form of social control in environments in which agents are likely to interact with others whose intentions are not known. It allows agents within such systems to reason about the reliability of others. More specifically, trust can be utilised to account for uncertainty about the willingness and capability of other agents to perform actions as agreed, rather than defecting when it proves to be more profitable. For the purpose of this work, we adapt Gambetta's definition [5], and define trust to be *a particular level of subjective probability with which an agent assesses that another agent will perform a particular action, both before the assessing agent can monitor such an action and in a context in which it affects the assessing agent's own action.*

Trust is often built over time by accumulating personal experience with others; we use this experience to judge how they will perform in an as yet unobserved situation. However, when assessing our trust in someone with whom we have no direct personal experience, we often ask others about their experiences with this individual. This collective opinion of others regarding an individual is known as the individual's *reputation*, which we use to assess its trustworthiness,

AAMAS'05, July 25-29, 2005, Utrecht, Netherlands.

if we have no personal experience.

Given the importance of trust and reputation in open systems and their use as a form of social control, several computational models of trust and reputation have been developed, each with requirements for the domain to which they apply (see [10] for a review of such models). In our case, the requirements can be summarised as follows. First, the model must provide a trust metric that represents a level of trust in an agent. Such a metric allows comparisons between agents so that one agent can be inferred as more trustworthy than another. The model must be able to provide a trust metric given the presence or absence of personal experience. Second, the model must reflect an individual's *confidence* in its level of trust for another agent. This is necessary so that an agent can determine the degree of influence the trust metric has on its decision about whether or not to interact with another individual. Generally speaking, higher confidence means a greater impact on the decision-making process, and lower confidence means less impact. Third, an agent must not assume that the opinions of others are accurate or based on actual experience. Thus, the model must be able to discount the opinions of others in the calculation of reputation, based on past reliability and consistency of the opinion providers. However, generally speaking, existing models do not allow an agent to effectively assess the reliability of an opinion source and use this assessment to discount the opinion provided by that source (see Section 5 for details). To meet the above requirements, we have developed TRAVOS, a trust and reputation model for agent-based VOs.

The remainder of this paper is organised as follows. Section 2 presents the basic TRAVOS model. Following from this, Section 3 provides a description of how the basic model has been expanded to include the functionality of handling inaccurate opinions from opinion sources. Empirical evaluation of these mechanisms is then presented in Section 4. Section 5 presents related work. Finally, Section 6 concludes.

2. THE TRAVOS MODEL

TRAVOS equips an agent (the truster) with two methods for assessing the trustworthiness of another agent (the trustee) in a given context. First, the truster can make the assessment based on the direct interactions it has had with the trustee. Second, the truster may assess the trustworthiness of another based on the reputation of the trustee.

2.1 Basic Notation

In a MAS consisting of n agents, we denote the set of all agents as $\mathcal{A} = \{a_1, a_2, ..., a_n\}$. Over time, distinct pairs of agents $\{a_x, a_y\} \subseteq \mathcal{A}$ may interact with one another, governed by contracts that specify the obligations of each agent towards its interaction partner. An interaction between a truster, $a_{tr} \in \mathcal{A}$, and a trustee, $a_{te} \in \mathcal{A}$, is considered successful by a_{tr} if a_{te} fulfills its obligations. From the perspective of a_{tr}, the outcome of an interaction between a_{tr} and a_{te} is summarised by a binary variable[1], $O_{a_{tr},a_{te}}$, where $O_{a_{tr},a_{te}} = 1$ indicates a successful (and $O_{a_{tr},a_{te}} = 0$ indicates an unsuccessful) interaction[2] for a_{tr} (Equation 1). Furthermore, we denote an outcome observed at time t as $O^t_{a_{tr},a_{te}}$, and the set of all outcomes observed from time t_0 to time t as $O^{t_0:t}_{a_{tr},a_{te}}$.

$$O_{a_{tr},a_{te}} = \begin{cases} 1 & \text{if contract is fulfilled by } a_{te} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

At any point of time t, the history of interactions between agents a_{tr} and a_{te} is recorded as a tuple, $\mathcal{R}^t_{a_{tr},a_{te}} = (m^t_{a_{tr},a_{te}}, n^t_{a_{tr},a_{te}})$ where the value of $m^t_{a_{tr},a_{te}}$ is the number of successful interactions for a_{tr} with a_{te}, while $n^t_{a_{tr},a_{te}}$ is the number of unsuccessful interactions. The tendency of an agent a_{te} to fulfil or default on its obligations is governed by its behaviour, which we represent as a variable $B_{a_{tr},a_{te}} \in [0,1]$. Here, $B_{a_{tr},a_{te}}$ specifies the intrinsic probability that a_{te} will fulfil its obligations during an interaction with a_{tr} (Equation 2). For example, if $B_{a_{tr},a_{te}} = 0.5$ then a_{te} is expected to break half of its contracts with a_{tr}, resulting in half the interactions between a_{te} and a_{tr} being unsuccessful from the perspective of a_{tr}.

$$B_{a_{tr},a_{te}} = p(O_{a_{tr},a_{te}} = 1), \quad \text{where } B_{a_{tr},a_{te}} \in [0,1] \quad (2)$$

In TRAVOS, each agent maintains a *level of trust* in each of the other agents in the system. Specifically, the level of trust of an agent a_{tr} in an agent a_{te}, denoted as $\tau_{a_{tr},a_{te}}$, represents a_{tr}'s assessment of the likelihood of a_{te} fulfilling its obligations. The *confidence* of a_{tr} in its assessment of a_{te} is denoted as $\gamma_{a_{tr},a_{te}}$. In this context, confidence is a metric that represents the accuracy of the trust value calculated by an agent given the number of observations (the evidence) it uses in the trust value calculation. Intuitively, more evidence results in higher confidence. The precise definitions and reasons behind these values are discussed in the proceeding Section.

2.2 Modelling Trust and Confidence

The first basic requirement of a computational trust model is that it should provide a metric for comparing the relative trustworthiness of different agents. From our definition of trust, we consider an agent to be trustworthy if it has a high probability of performing a particular action which, in our context, is to fulfil its obligations during an interaction. This probability is unavoidably subjective, because it can only be assessed from the individual viewpoint of the truster, based on the truster's personal experiences.

In light of this, we have adopted a probabilistic approach to modelling trust, based on the individual experiences of any agent in the role of a truster. If a truster, agent a_{tr}, has complete information about a trustee, agent a_{te}, then, according to a_{tr}, the probability that a_{te} fulfils its obligations is expressed by $B_{a_{tr},a_{te}}$. In general, however, complete information cannot be assumed; the best we can do is to use the *expected value* of $B_{a_{tr},a_{te}}$ given the experience of a_{tr}, which we consider to be the set of all interaction outcomes it has observed. Thus, we define the level of trust $\tau_{a_{tr},a_{te}}$ at time t as the expected value of $B_{a_{tr},a_{te}}$ given the set of outcomes $O^{1:t}_{a_{tr},a_{te}}$. This is expressed using standard statistical notation in Equation 3.

$$\tau_{a_{tr},a_{te}} = E[B_{a_{tr},a_{te}} | O^{1:t}_{a_{tr},a_{te}}] \quad (3)$$

[1]Representing a contract outcome with a binary variable is a simplification made for the purpose of our model. We concede that in certain circumstances, a more expressive representation may be appropriate.

[2]The outcome of an interaction from the perspective of one agent is not necessarily the same as from the perspective of its interaction partner. Thus, it is possible that $O_{a_{tr},a_{te}} \neq O_{a_{te},a_{tr}}$.

The expected value of a continuous random variable is dependent on the *probability density function* (pdf) used to model the probability that the variable will have a certain value. In Bayesian analysis, the beta family of pdfs is commonly used as a prior distribution for random variables that take on continuous values in the interval $[0, 1]$. For example, beta pdfs can be used to model the distribution of a random variable representing the unknown probability of a binary event [2]; $B_{a_{tr},a_{te}}$ is an example of such a variable. For this reason, we use beta pdfs in our model. (Beta pdfs have also previously been applied to the domain of trust for similar reasons; see Section 5).

The general formula for beta distributions is given in Equation 4. It has two parameters, α and β, which define the shape of the density function when plotted. Example plots can be seen in Figure 1, in which the horizontal axis represents the possible values of $B_{a_{tr},a_{te}}$, and the vertical axis gives the *relative* probability that each of these values is the true value for $B_{a_{tr},a_{te}}$. The most likely of $B_{a_{tr},a_{te}}$ is the curve maximum, while the width of the curve represents the amount of uncertainty over the true value of $B_{a_{tr},a_{te}}$. If α and β both have values close to 1, a wide density plot results, representing a high level of uncertainty about $B_{a_{tr},a_{te}}$. In the extreme case of $\alpha = \beta = 1$, the distribution is uniform, with all values of $B_{a_{tr},a_{te}}$ considered equally likely.

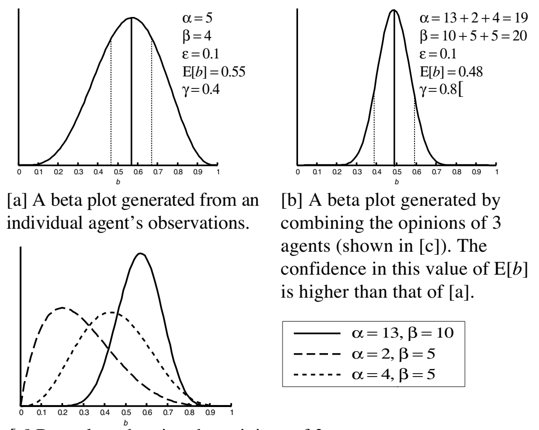
$$f(b|\alpha, \beta) = \frac{b^{\alpha-1}(1-b)^{\beta-1}}{\int U^{\alpha-1}(1-U)^{\beta-1}dU}, \quad \text{where } \alpha, \beta > 0 \tag{4}$$

Against this background, we now show how to calculate the value of $\tau_{a_{tr},a_{te}}$ based on the interaction outcomes observed by a_{tr}. First, we must find values for α and β that represent the beliefs of a_{tr} about a_{te}. Assuming that, prior to observing any interaction outcomes with a_{te}, a_{tr} believes that all possible values for $B_{a_{te}}$ are equally likely, then a_{tr}'s initial settings for α and β are $\alpha = \beta = 1$. Based on standard techniques, the parameter settings in light of observations are achieved by adding the number of successful outcomes to the initial setting of α, and the number of unsuccessful outcomes to β. In our notation, this is given in Equation 5. Then the final value for $\tau_{a_{tr},a_{te}}$ is calculated by applying the standard equation for the expected value of a beta distribution (Equation 6) to these parameter settings.

$$\begin{array}{l} \alpha = m^{1:t}_{a_{tr},a_{te}} + 1 \quad \text{and} \quad \beta = n^{1:t}_{a_{tr},a_{te}} + 1 \\ \text{where } t \text{ is the time of assessment} \end{array} \tag{5}$$

$$E[B_{a_{tr},a_{te}}|\alpha, \beta] = \frac{\alpha}{\alpha + \beta} \tag{6}$$

On its own, $\tau_{a_{tr},a_{te}}$ does not differentiate between cases in which a truster has adequate information about a trustee and cases in which it does not. Intuitively, observing many outcomes for an event is likely to lead to a better estimate for the future probability for that event (assuming all other things are equal). This creates the need for an agent to be able to measure its *confidence* in its value of trust. Therefore, we define a confidence metric $\gamma_{a_{tr},a_{te}}$ as the posterior probability given the evidence that the actual value of $B_{a_{tr},a_{te}}$ lies within an acceptable level of error ϵ about $\tau_{a_{tr},a_{te}}$ (Equation 7). This error ϵ influences the confidence of an agent given the same number of observations. For example, if the number of observations remains constant, a larger value of ϵ causes an agent to be more confident in its calculation of trust than a lower value.



[a] A beta plot generated from an individual agent's observations.

[b] A beta plot generated by combining the opinions of 3 agents (shown in [c]). The confidence in this value of E[b] is higher than that of [a].

[c] Beta plots showing the opinions of 3 separate agents, about a particular agent.

Figure 1: Example beta distributions for aggregating opinions of 3 agents.

$$\gamma_{a_{tr},a_{te}} = \frac{\int_{\tau_{a_{tr},a_{te}}-\epsilon}^{\tau_{a_{tr},a_{te}}+\epsilon} (B_{a_{tr},a_{te}})^{\alpha-1}(1 - B_{a_{tr},a_{te}})^{\beta-1} dB_{a_{tr},a_{te}}}{\int_0^1 U^{\alpha-1}(1-U)^{\beta-1}dU} \tag{7}$$

2.3 Modelling Reputation

Until now, we have only considered how an agent uses its own direct observations to calculate a level of trust. However, by using confidence, we can specify a decision-making process in an agent to lead it to seek more evidence when required. In TRAVOS, an agent a_{tr} calculates $\tau_{a_{tr},a_{te}}$ based on its personal experiences with a_{te}. If this value of $\tau_{a_{tr},a_{te}}$ has a corresponding confidence $\gamma_{a_{tr},a_{te}}$ which is below that of a predetermined *minimum confidence level*, denoted θ^{γ}, then a_{tr} will seek the opinions of other agents about a_{te} to boost its confidence above θ^{γ}. These collective opinions form a_{te}'s reputation and, by seeking it, a_{tr} can effectively obtain a larger set of observations.

The *true* opinion of a source $a_{op} \in \mathcal{A}$ at time t, about the trustee a_{te}, is the tuple, $\mathcal{R}^t_{a_{op},a_{te}} = (m^t_{a_{op},a_{te}}, n^t_{a_{op},a_{te}})$, defined in Section 2.1. In addition, we denote the *reported* opinion of a_{op} about a_{te} as $\hat{\mathcal{R}}^t_{a_{op},a_{te}} = (\hat{m}^t_{a_{op},a_{te}}, \hat{n}^t_{a_{op},a_{te}})$. This distinction is important because a_{op} may not reveal $\mathcal{R}^t_{a_{op},a_{te}}$ truthfully. The truster, a_{tr}, must form a single trust value from all such opinions that it receives. Assuming that opinions are independent, then an elegant and efficient solution to this problem is to enumerate the successful and unsuccessful interactions from all the reports it receives, where p is the total number of reports (see Equation 8). The resulting values, denoted $N_{a_{tr},a_{te}}$ and $M_{a_{tr},a_{te}}$ respectively, represent the reputation of a_{te} from the perspective of a_{tr}. These values can then be used to calculate shape parameters (see Equation 9) for a beta distribution, to give a trust value determined by opinions provided from others. In addition, the truster takes on board any direct experience it has with the trustee, by adding its own values for $n_{a_{tr},a_{te}}$ and $m_{a_{tr},a_{te}}$ with the same equation. The confidence value $\gamma_{a_{tr},a_{te}}$ for this combined distribution will be higher than for any of the component opinions, because more observations will have been taken into account (see Figure 1).

$$N_{a_{tr},a_{te}} = \sum_{k=0}^{p} \hat{n}_{a_k,a_{te}}, \quad M_{a_{tr},a_{te}} = \sum_{k=0}^{p} \hat{m}_{a_k,a_{te}} \qquad (8)$$

$$\alpha = M_{a_{tr},a_{te}} + 1 \quad \text{and} \quad \beta = N_{a_{tr},a_{te}} + 1 \qquad (9)$$

The desirable feature of this approach is that, provided Conditions 1 and 2 hold, the resulting trust value and confidence level is the same as it would be if all the observations had been observed directly by the truster itself.

CONDITION 1 (COMMON BEHAVIOUR). *The behaviour of the trustee must be independent of the identity of the truster it is interacting with. Thus:*

$$\forall a_{te} \quad \forall a_{op}, B_{a_{te},a_{tr}} = B_{a_{op},a_{tr}}$$

CONDITION 2 (TRUTH TELLING). *The reputation provider must report its observations accurately and truthfully. Thus:*

$$\forall a_{te} \quad \forall a_{op}, \quad \mathcal{R}^t_{a_{op},a_{te}} = \hat{\mathcal{R}}^t_{a_{op},a_{te}}$$

Unfortunately, however we cannot expect these conditions to hold in a broad range of situations. For instance, a trustee may value interactions with one agent over another, so it might therefore commit more resources to the valued agent to increase its success rate, thus introducing a bias in its perceived behaviour. Similarly, in the case of a rater's opinion of a trustee, it is possible that the rater has an incentive to misrepresent its true view of the trustee. Such an incentive could have a positive or a negative effect on a trustee's reputation; if a strong co-operative relationship exists between trustee and rater, the rater may choose to overestimate its likelihood of success, whereas a competitive relationship may lead the rater to underestimate the trustee. Due to these possibilities, we consider the methods of dealing with inaccurate reputation sources an important requirement for a computational trust model. In the next section, we introduce our solution to this requirement, building upon the basic model introduced thus far.

3. FILTERING INACCURATE REPUTATION

Inaccurate reputation reports can be due to opinion providers being malevolent or having incomplete information. In both cases, an agent must be able to assess the reliability of the reports passed to it. The general solution to coping with inaccurate reputation reports is to adjust or ignore opinions judged to be unreliable (in order to reduce their effect on the trustee's reputation). There are two basic approaches to achieving this that have been proposed in the literature; Jøsang et al. [8] refer to these as *endogenous* and *exogenous* methods. The former attempt to identify unreliable reputation information by considering the statistical properties of the reported opinions alone (e.g. [12, 3]). The latter rely on other information to make such judgements, such as the reputation of the source or the relationship with the trustee (e.g. [1, 13]).

Many proposals for endogenous techniques assume that inaccurate or unfair raters will generally be in a minority among reputation sources. Based on this assumption, they consider reputation providers whose opinions deviate in some way from mainstream opinion to be those most likely to be inaccurate. Our solution is exogenous, in that we judge a reputation provider on the perceived accuracy of its past opinions, rather than its deviation from mainstream opinion. Moreover, we define a two step-method: First, we calculate the probability that an agent will provide an accurate opinion given its past opinions and later observed interactions with the trustees for which opinions were given. Second, based on this value, we reduce the distance between a rater's opinion and the prior belief that all possible values for an agent's behaviour are equally probable. Once all the opinions collected about a trustee have been adjusted in this way, the opinions are aggregated using the technique described in Section 2.3. Below we describe this technique in more detail: Section 3.1 details how the probability of accuracy is calculated and Section 3.2 shows how opinions are adjusted and the combined reputation obtained.

3.1 Estimating the Probability of Accuracy

The first stage in our solution is to estimate the probability that a rater's stated opinion of a trustee is accurate. The way in which we do this depends on the value of the current opinion under consideration, which we denote $\mathcal{R}^r_{a_{op},a_{te}} = (m^r_{a_{op},a_{te}}, n^r_{a_{op},a_{te}})$. Specifically, let E^r be the expected value of a beta distribution D^r, such that $\alpha^r = m^r_{a_{op},a_{te}} + 1$ and $\beta^r = n^r_{a_{op},a_{te}} + 1$. Our goal is then to estimate the probability that $E^r = B_{a_{tr},a_{te}}$; we denote this as $\rho_{a_{tr},a_{op}}$ — the accuracy of a_{op} according to a_{tr}.

To perform this estimation, we consider the outcomes of all previous interactions for which a_{op} provided an opinion to a_{tr}, similar to $\mathcal{R}^r_{a_{op},a_{te}}$. Using these outcomes, we construct a beta distribution, denoted D^o; if E^r is close to the expected value of D^o, denoted E^o, then this suggests that a_{op}'s opinionsf are generally correlated to what is actually observed. We can therefore judge a_{op}'s accuracy to be high. Similarly, if E^r deviates significantly from E^o, then we judge a_{op} to have low accuracy.

Let $\mathcal{H}_{a_{tr},a_{op}}$ be the set of all pairs of the form $(O_{a_{tr},a_x}, \hat{\mathcal{R}}_{a_{op},a_x})$, where a_x is any member of $\mathcal{A}$, and O_{a_{tr},a_x} is the outcome of an interaction for which, prior to observing this outcome, a_{op} gave the opinion $\hat{\mathcal{R}}_{a_{op},a_x}$. Second, divide the range of possible values of E^r into N disjoint intervals (or bins) $bin_1, ..., bin_n$. Third, calculate E^r, and find the interval bin^o that contains the value of E^r. Fourth, find the subset $\mathcal{H}^r_{a_{tr},a_{op}} \subseteq \mathcal{H}_{a_{tr},a_{op}}$, which comprises all pairs for which the opinion falls in bin^o. From this set, count the total number of pairs in $\mathcal{H}^r_{a_{tr},a_{op}}$ for which the interaction outcome was successful (denoted $C_{success}$) and, similarly, for those which were not successful (denoted C_{fail}). Based on these frequencies, the parameters for D^o are given as $\alpha^o = C_{success} + 1$ and $\beta^o = C_{fail} + 1$. Using D^o, we now calculate $\rho_{a_{tr},a_{op}}$ as the portion of the total mass of D^o that lies in the interval bin^o (Equation 10).

$$\rho_{a_{tr},a_{op}} = \frac{\int_{min(bin^o)}^{max(bin^o)} X^{\alpha^o-1}(1-X)^{\beta^o-1}dX}{\int_0^1 U^{\alpha^o-1}(1-U)^{\beta^o-1}dU} \qquad (10)$$

The intuition behind this process is illustrated in Figure 3.1. Here, the range of possible values of E^r has been divided into five intervals, $bin_1 = [0, 0.2], ..., bin_5 = [0.8, 1]$. The opinion provider, a_{op}, has provided a_{tr} with an opinion for which the expected value is in bin_4; thus, we consider all previous interaction outcomes for which a_{op} provided an opinion in bin_4 to a_{tr}. In this case, the portion of successful outcomes, and thus E^o, is also in bin_4, and so $\rho_{a_{tr},a_{op}}$ is high. If subsequent outcome-opinion pairs were also to fol-

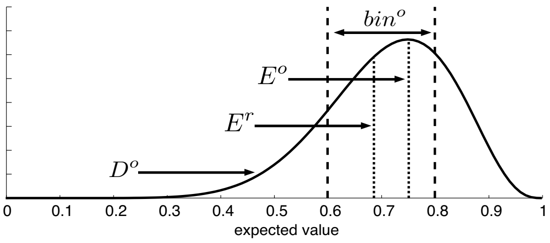


Figure 2: Illustration of $\rho_{a_{tr},a_{op}}$ Estimation Process

low this trend, then D^o would be highly peaked inside this interval; therefore $\rho_{a_{tr},a_{op}}$ would converge to one. On the other hand, if subsequent outcomes disagreed with their corresponding opinions, then $\rho_{a_{tr},a_{op}}$ would approach 0. One implication of this technique is that the number of bins effectively determines an acceptable margin of error in opinion provider accuracy: a larger set of opinion providers will have their estimated accuracy converge to 1 if bin sizes are large, compared to if bin sizes are small.

3.2 Adjusting Reputation Source Opinions

To describe how we adjust reputation opinions, we must introduce some new notation. First, let D^c be the beta distribution that results from combining all of a trustee's reputation information (using Equations 8 and 9). Second, let D^{c-r} be a distribution constructed using the same equations, except that the opinion under consideration, $\mathcal{R}^r_{a_{op},a_{te}}$, is omitted. Third, let $\bar{D}$ be the result of adjusting the opinion distribution D^r, according to the process we describe here. Finally, we refer to the standard deviation (denoted σ), expected value and parameters of each distribution by using the respective superscript; for instance, D^c has parameters α^c and β^c, with standard deviation σ^c and expected value E^c.

Now, our goal is to reduce the *effect* of unreliable opinions on D^c. Effectively, by adding $\mathcal{R}^r_{a_{op},a_{te}}$ to a trustee's reputation, we move E^c in the direction of E^r. The standard deviation of D^r contributes to the confidence value for the combined reputation value but, more importantly, its value relative to σ^{c-r} determines how far E^c will move towards E^r, . This effect has important implications: Consider as an example three distributions d_1, d_2 and d_3, with shape parameters, expected value and standard deviation as shown in Table 1; the results of combining d_1 with each of the other two distributions are shown in the last two rows. As can be

Distribution	α	β	E	σ
d_1	540	280	0.6585	0.0165
d_2	200	200	0.5000	0.0250
d_3	5000	5000	0.5000	0.0050
$d_1 + d_2$	740	480	0.6066	0.0140
$d_1 + d_3$	5540	5280	0.5120	0.0048

Table 1: Combination of beta distributions.

seen, distributions d_2 and d_3 have identical expected values with standard deviations of 0.025 and 0.005 respectively. Although the difference between these values is small (0.02), the result of combining d_1 with d_2 is quite different from combining d_1 and d_3. Whereas the expected value in the first case falls approximately between the expected values for d_1 and d_2, in the latter case, the relatively small parameter values of d_1 compared to d_3 mean that d_1 has virtually no impact on the combined result. Obviously, this is due to our method of reputation combination (Equation 8), in which the parameter values are summed. This is important because it shows how, if left unchecked, an unfair rater could purposely increase the weight an agent places in its opinion by providing very large values for m and n which, in turn, determine α and β.

In light of this, we adopt an approach that significantly reduces very high parameter values unless the probability of the rater's opinion being accurate is very close to 1. Specifically, we reduce the distance between the expected value and standard deviation of D^r, and the uniform distribution, $\alpha = \beta = 1$, which represents a state of no information (Equations 11 and 12). Here, we denote the standard deviation of the uniform distribution as $\sigma_{uniform}$ and its expected value as $E_{uniform}$. By adjusting the standard deviation in this way, rather than changing the α and β parameters directly, we ensure that large parameter values are decreased more than smaller more conservative values. We adjust the expected value to guard against cases where we do not have enough reliable opinions to mediate the effect of unreliable opinions; if we did not adjust the expected value, then in the absence of any other information, we would take an opinion source's word as true, even if we did not consider its opinion reliable.

$$\bar{E} = E_{uniform} + \rho_{a_{tr},a_{op}} \cdot (E^r - E_{uniform}) \quad (11)$$
$$\bar{\sigma} = \sigma_{uniform} + \rho_{a_{tr},a_{op}} \cdot (\sigma^r - \sigma_{uniform}) \quad (12)$$

Once all reputation opinions have been adjusted in this way, we sum the ratings as normal according to Equation 8, by calculating the adjusted values for $\hat{m}_{a_{op},a_{te}}$ and $\hat{n}_{a_{op},a_{te}}$. It can be shown that the adjusted parameter values, $\bar{\alpha}$ and $\bar{\beta}$, can be calculated according to Equation 13 and Equation 14. The new values for $\hat{m}_{a_{op},a_{te}}$ and $\hat{n}_{a_{op},a_{te}}$ are then given by subtracting the prior parameter settings from the adjusted distribution parameters (Eqn. 15).

$$\bar{\alpha} = \frac{\bar{E}^2 - \bar{E}^3}{\bar{\sigma}^2} - \bar{E} \quad (13)$$
$$\bar{\beta} = \frac{(1-\bar{E})^2 - (1-\bar{E})^3}{\bar{\sigma}^2} - (1-\bar{E}) \quad (14)$$
$$\bar{m}_{a_{op},a_{te}} = \bar{\alpha} - 1 \quad , \quad \bar{n}_{a_{op},a_{te}} = \bar{\beta} - 1 \quad (15)$$

4. EMPIRICAL EVALUATION

In this section we present the results of our empirical evaluation performed on TRAVOS. Our discussion is structured as follows: Section 4.1 describes our evaluation testbed and overall experimental methodology; Section 4.2 compares the reputation component of TRAVOS to the most similar model found in the literature; Section 4.3 investigates the overall performance of TRAVOS when both direct experience and reputation are taken into account.

4.1 Experiment Methodology

Evaluation of TRAVOS took place using a simulated marketplace environment, consisting of three distinct sets of agents: provider agents $\mathcal{P} \subset \mathcal{A}$, consumer agents $\mathcal{C} \subset \mathcal{A}$, and reputation source agents $\mathcal{S} \subset \mathcal{A}$. For our purposes, the role of any $c \in \mathcal{C}$ is to evaluate $\tau_{c,p}$ for all $p \in \mathcal{P}$. The behaviour of each provider and reputation source agent is set before each experiment. Specifically, the behaviour of a provider $p_1 \in \mathcal{P}$ is determined by the parameter B_{c,p_1} as described in Section 2.1. Here, reputation sources are divided into three types that define their behaviour: *accurate* sources report the number of successful and unsuccessful interactions they have had with a given consumer without modification; *noisy* sources add gaussian noise to the beta distribution determined from their interaction history, rounding the resulting expected value if necessary to ensure that it remains in the interval $[0, 1]$; and *lying* sources attempt to maximally mislead the consumer by setting the expected value $E[B_{c,p}]$ to $1 - E[B_{c,p}]$.

Against this background, all experiments consisted of a series of episodes in which a consumer was asked to assess its trust in all providers $\mathcal{P}$. Based on these assessments, we calculate the consumer's mean estimation error for the episode (Eqn. 16). This gives us a measure of the consumer's performance on assessing the provider population as a whole. The value of this metric will vary depending on the distribution of values of $B_{c,p}$ over the provider population. For simplicity, all the results described in the next sections have been acquired for a population of 101 providers with values of $B_{c,p}$ chosen uniformly between 0 and 1 at intervals of 0.01.

$$avg\_estimate\_err = \frac{1}{N}\sum_{i=1}^{n} abs(\tau_{c,p_i} - B_{c,p_i}) \qquad (16)$$

In each episode, the consumer may draw upon both the opinions of reputation sources in $\mathcal{S}$ and its own interaction history with both the providers and reputation sources. However, to ensure that the results of each episode are independent, the interaction history between all agents is cleared before every episode, and re-populated according to set parameters. All the results that we will discuss have been tested for statistical significance using Analysis of Variance techniques and Scheffé tests.

4.2 TRAVOS vs. the Beta Reputation System

Of the existing computational trust models in the literature, the most similar to TRAVOS is the Beta Reputation System (BRS) (see Section 5 for more detail). Like TRAVOS, this uses the beta family of probability functions to calculate the posterior probability of an agent a_{te}'s behaviour holding a certain value, given past interactions with a_{te}. However, the models differ significantly in their approach to handling inaccurate reputation. TRAVOS assesses each reputation source individually, based on the perceived accuracy of past opinions. In contrast, BRS assumes that the majority of reputation sources provide an accurate opinion, and it ignores any opinions that deviate significantly from the average. Since BRS does not differentiate between reputation and direct observations, we have focused our evaluation on scenarios were consumers have no personal experience, and must therefore rely on reputation only.

To show variation in performance depending on reputation source behaviour, we ran experiments with populations containing accurate and lying reputation sources, and populations containing accurate and noisy sources. In each case, we kept the total number of sources equal to 20, but ran separate experiments in which the percentage of accurate sources was set to 0%, 50% and 100% (see Table 2). Now figure 3 shows the mean estimation error of TRAVOS and

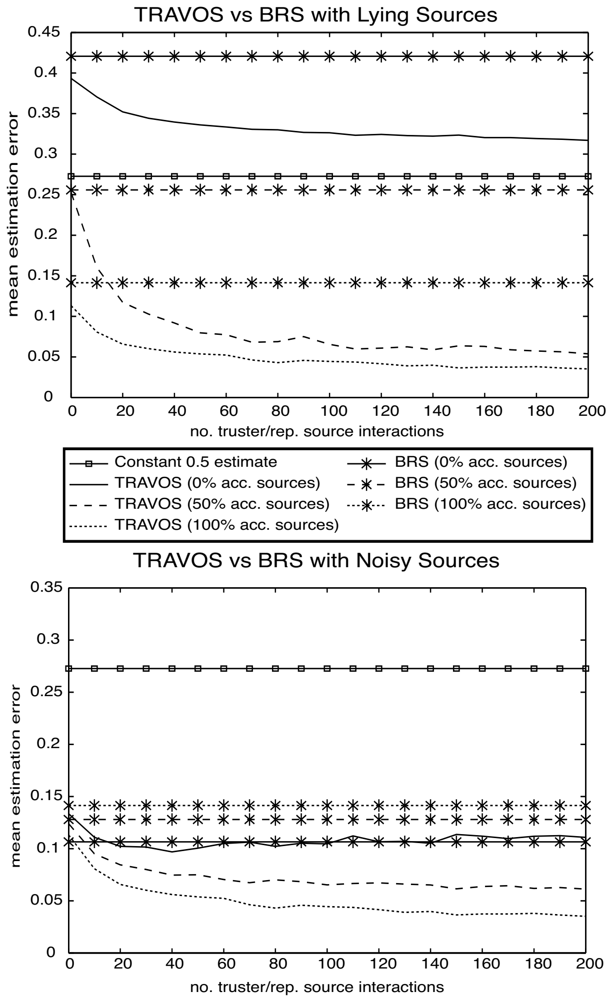


Figure 3: TRAVOS Reputation System vs BRS

experiment	no. lying	no. noisy	no. accurate
1	0	0	20
2	0	10	10
3	0	20	0
4	10	0	10
5	20	0	0

Table 2: Reputation Source Populations

BRS with these different reputation source populations averaged over 50 independent episodes in each experiment. To provide a benchmark, the figure also shows the mean estimation error of a consumer $c_{0.5}$, which keeps $\tau_{c_{0.5},p} = 0.5$ for all $p \in \mathcal{P}$. Results are plotted against the number of previous interactions that have occurred between the consumer and each reputation source.

As can be seen, in populations containing lying agents, the mean estimation error of TRAVOS is consistently equal to or less than that of BRS. Moreover, estimation errors decrease significantly for TRAVOS as the number of consumer to reputation source interactions increases. In contrast, BRS's performance remains constant, since it does not learn from past experience. Both models perform consistently better than $c_{0.5}$ in populations containing 50% or 0% liars. However, in populations containing only lying sources, both models were sufficiently misled to perform worse than $c_{0.5}$, but TRAVOS suffered less from this effect than BRS. Specifically, when the number of past consumer to reputation interactions is low, TRAVOS benefits from its initially conservative belief in reputation source opinions. The benefit is enhanced further as the consumer becomes more skeptical with experience.

Similar results can be seen in populations containing noisy sources. In general, performance is better because noisy source opinions are not as misleading as lying source opinions on average. TRAVOS still out performs BRS in most cases, except when the population contains only noisy sources. In this case, BRS has a small but statistically significant advantage when the number of consumer to reputation source interactions are less than 10.

4.3 TRAVOS Component Performance

To evaluate the overall performance of TRAVOS, we compared three versions of the system that used the following information respectively: direct interactions between the consumer and providers; direct provider experience and reputation; and reputation information only. In these experiments, we varied the number of interactions between the consumers and providers, and kept the number of consumer to reputation source interactions constant at 10. We used the same reputation source populations as described in Section 4.2. The mean estimation errors for a subset of these experiments are shown in Figure 4. Using only direct consumer to provider experience, the mean estimation error decreases as the number of consumer to provider interactions increases. As would be expected, using both information sources when the number of consumer to provider interactions is low, results in similar performance to using reputation information only. However, in some cases, the combined model may provide marginally worse performance than using reputation only.[3] This can be attributed to the fact that TRAVOS will always put more faith in direct experience than reputation.

With a population of 50% lying reputation sources, the combined model is misled enough to temporarily increase its error rate above that of the direct only model. This is a symptom of the relatively small number of consumer to reputation source interactions (10), which is insufficient for the consumer to completely discount all the reputation information as unreliable. The effect disappears when the number of such interactions is increased to 20; however, these results are not illustrated graphically in this paper.

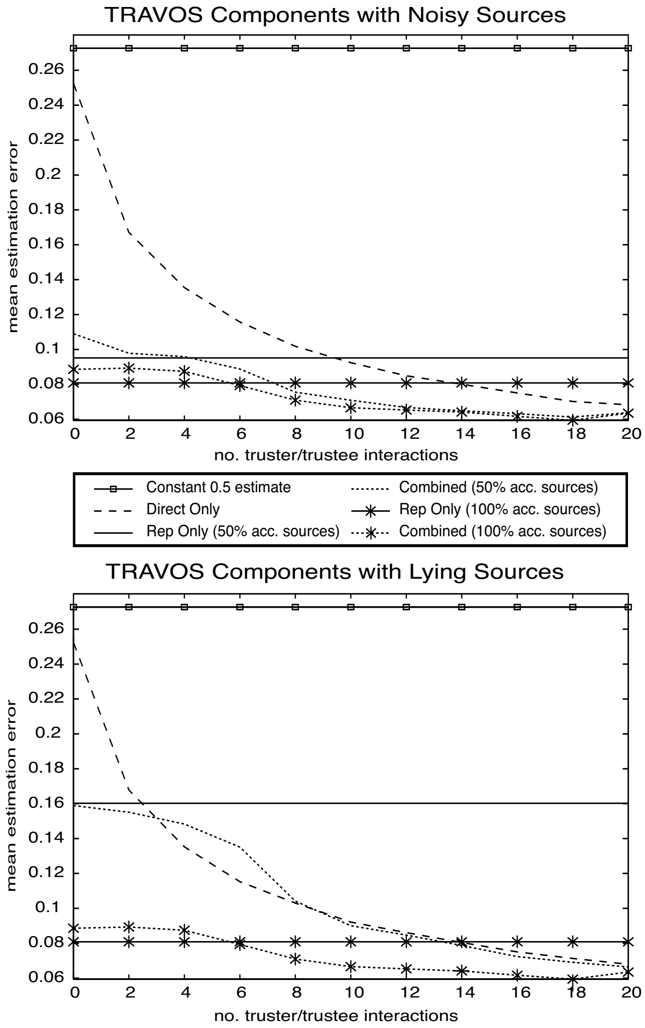


Figure 4: TRAVOS Component Performance

5. RELATED WORK

There are many computational models of trust, a review of which can be found in [10]. Generally speaking, however, models not based on probability theory (e.g. [6, 11, 14]) calculate trust from hand-crafted formulae that yield the desired results, but that can be considered somewhat ad hoc.

Probabilistic approaches are not commonly used in the field of computational trust, but there are a couple of such models in the literature (e.g. [9, 7, 12]. In particular, the Beta Reputation System (BRS) [7] is a probabilistic trust model like TRAVOS, which is based on the beta distribution. The system is specifically designed for online communities and is centralised. It works by users giving ratings to the performance of other users in the community. Here, ratings consist of a single value that is used to obtain positive and negative feedback values. The feedback values are then used to calculate shape parameters that determine the reputation of the user the rating applies to. However, BRS does not show how it is able to cope with misleading information.

Whitby et al. [12] extend the BRS and show how it can be used to filter unfair ratings, either unfairly positive or negative, towards a certain agent. It is primarily this extension that we compare to TRAVOS in Section 4.2. However their approach is only effective when a significant majority of available reputation sources are fair and accurate, and there are potentially many important scenarios where this assumption does not hold. One example occurs when no opinion providers have previously interacted with a trustee. In this case, the only agents that will provide an opinion will be those with an incentive to lie. In TRAVOS, opin-

[3]This effect was not considered significant under a Scheffé test, but was considered significant by Least Significant Difference Testing. The latter technique is, in general, less conservative at concluding that a difference between groups does exist.

ion providers that continually lie will have their opinions discarded, regardless of the proportion of opinions about a trustee that are inaccurate.

Another method for filtering inaccurate reputation is described by [13]. This is similar to TRAVOS, in that it rates opinion source accuracy based on subsequent observations of trustee behaviour. However, at this point the models diverge, and adopt different methods for representing trust, grounding trust in trustee observations, and implementing reputation filtering. Further experimentation is required to compare this approach to TRAVOS.

6. CONCLUSIONS AND FUTURE WORK

This paper has presented a novel model of trust for use in open agent systems, such as the Grid. Its main benefits are that it provides a mechanism for assessing the trustworthiness of others in situations both in which the agents have interacted before and share past experiences, and in which there is little or no past experience between them. Establishing the trustworthiness of others, and then selecting the most trustworthy, gives an agent the ability to maximise the probability that there will be no harmful repercussions from the interaction.

In situations were an agent's past experience with a trustee is low, it can draw upon reputation provider opinions. However, in doing so, the agent risks lowering, rather than increasing, assessment performance due to inaccurate opinions. TRAVOS copes with this by having an initially conservative estimate in reputation accuracy. Through repeated interactions with individual reputation sources, it learns to distinguish reliable from unreliable sources. By empirical evaluation, we have demonstrated that this approach allows reputation to be used to significantly improve performance while guarding against the negative effects of inaccurate opinions. Moreover, TRAVOS can extract a positive influence on performance from reputation, even when 50% of sources are intentionally misleading. This effect is increased significantly through repeated interactions with individual reputation sources. When 100% of sources are misleading, reputation has a negative effect on performance. However, even in this case, performance is increased by learning, and it outperforms the most similar model in the literature, in the majority of scenarios tested.

As it stands, TRAVOS assumes that the behaviour of agents does not change over time, but in many cases this is an unsafe assumption. In particular we believe that agents may well change their behaviour over time, and that some will have time-based behavioural strategies. Future work will therefore include the removal of this assumption and the use of functions that allow an agent to take into account the fact that very old experiences may not be relevant in predicting the behaviour of an individual. Further extensions to TRAVOS will include using the rich social metadata that exists within a VO environment in the calculation of a trust value. Thus, as described in Section 1, VOs are social structures, and we can draw out social data such as roles and relationships that exist both between VOs and VO members. The incorporation of such data into the trust metric should allow for more accurate trust assessments to be formed.

7. ACKNOWLEDGEMENTS

This work is part of the CONOISE-G project, funded by the DTI and EPSRC through the Welsh e-Science Centre, in collaboration with the Office of the Chief Technologist of BT. The research in this paper is also funded in part by the EPSRC Mohican Project (Reference no: GR/R32697/01).

8. REFERENCES

[1] S. Buchegger and J. Y. L. Boudec. A robust reputation system for mobile ad-hoc networks ic/2003/50. Technical report, EPFL-IC-LCA, 2003.

[2] M. DeGroot and M. Schervish. *Probability & Statistics*. Addison-Wesley, 2002.

[3] C. Dellarocas. Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems. In *ICIS*, pages 520–525, 2000.

[4] I. Foster, N. R. Jennings, and C. Kesselman. Brain meets brawn: Why grid and agents need each other. In *AAMAS '04*, pages 8–15, 2004. IEEE Computer Society.

[5] D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Basil Blackwell, 1988.

[6] T. D. Huynh, N. R. Jennings, and N. Shadbolt. Developing an integrated trust and reputation model for open multi-agent systems. In *Proc. 7th Int. Workshop on Trust in Agent Societies*, pages 62–77, 2004.

[7] R. Ismail and A. Jøsang. The beta reputation system. In *Proc. 15th Bled Conf. on Electronic Commerce*, 2002.

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

[9] L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *Proc. of the 35th Hawaii Int. Conf. on System Science*, pages 280–287, 2002.

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

[11] J. Sabater and C. Sierra. Regret: A reputation model for gregarious societies. In *4th Workshop on Deception Fraud and Trust in Agent Societies*, pages 61–70, 2001.

[12] 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.

[13] B. Yu and M. P. Singh. Detecting deception in reputation management. In *AAMAS '03*, pages 73–80, 2003. ACM Press.

[14] G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms in electronic marketplaces. In *Proc. 32nd Annual Hawaii Int. Conf. on System Sciences*, page 8026. IEEE Computer Society, 1999.