TRAVOS: Trust and Reputation in the Context of Inaccurate Information Sources

W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck

Electronics & Computer Science, University of Southampton,
Southampton SO17 1BJ, UK.

{wtlt03r, jp03r, nrj, mml}@ecs.soton.ac.uk

Abstract. In many dynamic open systems, agents have to interact with one another to achieve their goals. Here, agents may be self-interested and when trusted to perform an action for another, may betray that trust by not performing the action as required. In addition, due to the size of such systems, agents will often interact with other agents with which they have little or no past experience. There is therefore a need to develop a model of trust and reputation that will ensure good interactions among software agents in large scale open systems. 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, and 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.

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, peer-to-peer networks, and the Semantic Web. In all of 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 fulfilled by its interaction partner, it is a good interaction.

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 difficulty 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, the solution to this problem is integral to the wide-scale acceptance of the Grid and agent-based VOs [5].

It is now well established that computational *trust* is important in such open systems [13, 9, 16]. Specifically, trust provides a form of social control in environments in which agents are likely to interact with others whose intentions are not known, and 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 paper, we adapt Gambetta's definition [6], 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 up over time by accumulating personal experience with others; we use this experience to judge how agents will perform in an as yet unobserved situation. However, when assessing trust in an individual with whom we have no direct personal experience, we often ask others about their experiences with that individual. This collective opinion of others regarding an individual is known as the individual's *reputation*, which we use to assess its trustworthiness, if we have no personal experience of it.

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 tailored to the domain to which they apply (see [13] 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 of the trust metric on the decision about whether 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

of opinion providers. However, existing models do not generally allow an agent to effectively assess the reliability of an opinion source and use the assessment to discount the opinion provided by that source.

To meet the above requirements, we have developed TRAVOS, a trust and reputation model for agent-based VOs, as described in this paper, which is organised as follows. Section 2 presents the basic TRAVOS model, and Section 3 then 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 presented in Section 4. Section 5 presents related work, and 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 its previous direct interactions with the trustee. Second, the truster may assess trustworthiness 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. Here, and in the rest of this discussion, we assume that all interactions take place under similar obligations. This is because an agent may behave differently when asked to provide one type of service over another, and so the best indicator of how an agent will perform under certain obligations in the future is how it performed under similar obligations in the past. Therefore, the assessment of a trustee under different obligations is best treated separately. In any case, 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} fulfils 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} (see Equation 1). We denote an outcome observed at time t as $O^t_{a_{tr},a_{te}}$, and the set of all outcomes observed from time 1 to time t as $O^{1:t}_{a_{tr},a_{te}}$. Here, each point in time is a natural number, $\{t : t \in \mathbb{Z}, t > 0\}$, in which at most one interaction

[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. This is part of our future work.

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

between any given pair of agents may take place. Therefore, $O^{1:t}_{a_{tr},a_{te}}$ is a set of at most t binary variables representing all the interactions that have taken place between a_{tr} and a_{te} up to and including time t.

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

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} (see 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}.

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

In TRAVOS, the *trust* of an agent a_{tr} in an agent a_{te}, denoted $\tau_{a_{tr},a_{te}}$, is a_{tr}'s estimate of the probability that a_{te} will fulfil its obligations to a_{tr} during an interaction. 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 below.

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 adopt a probabilistic approach to modelling trust, based on the experiences of an agent in the role of a truster. If a truster, a_{tr}, has complete information about a trustee, 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, and according to the Bayesian view [4], the best we can do is to use the *expected value* of $B_{a_{tr},a_{te}}$ given the knowledge of a_{tr}. In particular, we consider the knowledge of a_{tr} to be the set of all interaction outcomes it has observed. However, in adopting a Bayesian rather than frequentist stance, we allow for the possibility that a truster may use other prior information

in its assessment, particularly during bootstrapping, when few observations of a trustee are available (see Section 6). 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.

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

In order to determine this expected value, we need a probability distribution, defined by a *probability density function* (pdf), which is used to model the relative probability that $B_{a_{tr},a_{te}}$ 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], where $B_{a_{tr},a_{te}}$ is an example of such a variable. For this reason, beta pdfs which have also been applied in previous work in the domain of trust (see Section 5), are also used in our model.

The standard formula for beta distributions is given in Equation 4, in which two parameters, α and β define the shape of the density function when plotted.[3] 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 value of $B_{a_{tr},a_{te}}$ is the curve maximum, while the shape of the curve represents the degree 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, indicating 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.

$$f(B_{a_{tr},a_{te}}|\alpha,\beta) = \frac{(B_{a_{tr},a_{te}})^{\alpha-1}(1-B_{a_{tr},a_{te}})^{\beta-1}}{\int_0^1 U^{\alpha-1}(1-U)^{\beta-1}dU}, \quad \text{where } \alpha,\beta > 0 \tag{4}$$

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 (see Equation 6) to these parameter settings.

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

[3] The denominator in Equation 4 is a normalising constant, which is used to fulfil the constraint that the definite integral of a probability distribution must be equal to 1.

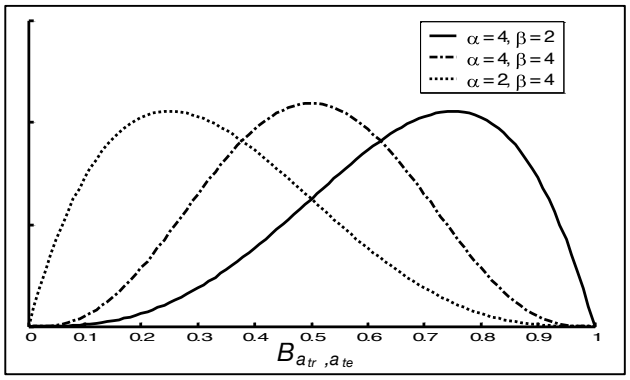

Fig. 1. Example beta plots, showing how the beta curve shape changes with the parameters α and β

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

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 of a given type of event is likely to lead to a more accurate estimate of such an event's outcome. This creates the need for an agent to be able to measure its *confidence* in its value of trust, for which we define a confidence metric, $\gamma_{a_{tr},a_{te}}$, as the posterior probability that the actual value of $B_{a_{tr},a_{te}}$ lies within an acceptable margin of error ϵ about $\tau_{a_{tr},a_{te}}$. This is calculated using Equation 7, which can intuitively be interpreted as the proportion of the probability distribution that lies between the bounds $(\tau_{a_{tr},a_{te}} - \epsilon)$ and $(\tau_{a_{tr},a_{te}} + \epsilon)$. The error ϵ influences the confidence value an agent calculates for a given set of observations. That is, for a given set of observations, a larger value of ϵ causes a larger proportion of the beta distribution to fall in the range $[\tau_{a_{tr},a_{te}} - \epsilon, \tau_{a_{tr},a_{te}} + \epsilon]$, so resulting in a large value for $\gamma_{a_{tr},a_{te}}$.

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

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, in certain circumstances, it may also be appropriate for a truster to seek third party opinions, in order to boost the information it has available on which to assess a trustee. In particular, if the truster has a low confidence level in its assessment, based only on its own experience, then seeking third party opinions may significantly boost the accuracy of its assessment. However, if the truster has significant first-hand experience with the

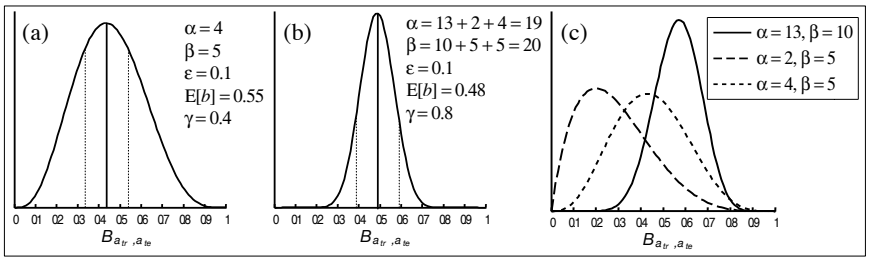

Fig. 2. Example beta distributions for aggregating opinions of 3 agents.

trustee, then the risk of obtaining misleading opinions, and any communication cost involved, may out weigh any small increase in accuracy that may be gained.

In light of this, we use confidence values to 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, $\mathcal{R}^t_{a_{op},a_{te}} = (m^t_{a_{op},a_{te}}, n^t_{a_{op},a_{te}})$, as defined in Section 2.1. 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, for reasons of self-interest. The truster, a_{tr}, must form a single trust value from all such opinions 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 considers 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 effect of combining opinions in this way is illustrated in Figure 2. In this figure, part (a) shows a beta distribution representing one agent's opinion, along with the attributes of the distribution that have been discussed so far. In contrast to this, part (c) illustrates the differences between the distribution in part (a) and distributions representing the opinions of two other agents with different experiences. The result of combining all three opinions is illustrated in part (b), of which there are two important characteristics. First, the distribution with parameters $\alpha = 13$ and $\beta = 10$ is based on more observations than the remaining two distributions put together, and so has the greatest impact on

the shape and expected value of the combined distribution. This demonstrates how conflicts between different opinions are resolved: the combined trust value is essentially a weighted average of the individual opinions, where opinions with higher confidence values are given greater weight. Second, the variance of the combined distribution is strictly greater than any one of the component distributions. This reflects that fact that it is based on more observations overall, and so has a greater confidence value.

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

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

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. However, this also assumes that the way in which different agents assess a trustee's behaviour is consistent. That is, a truster's opinion providers categorise an interaction as successful, or unsuccessful, in the same way as the truster itself.

Condition 1 (Common Behaviour) *The behaviour of the trustee must be independent of the identity of the truster with which it is interacting. Thus:*

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

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 more than with 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 cooperative 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 arise when either Condition 1 or Condition 2 is broken, due to an opinion provider being malevolent or a trustee behaving inconsistently towards different agents. In both cases, an agent must be able to

assess the reliability of the reports passed to it, and the general solution 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.* [9] 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. [18, 3]), while the latter rely on other information to make such judgements, such as the reputation of the source or its relationship with the trustee (e.g. [1, 19, 10])[4].

Many proposals for endogenous techniques assume that inaccurate or unfair raters are generally in a minority among reputation sources, and thus 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 as follows. First, we calculate the probability that an agent will provide an accurate opinion given its past opinions and later observed[5] 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 above. In so doing, we reduce the influence that an opinion provider has on a truster's assessment of a trustee, if the provider's opinion is consistently biased in one way or another. This can be true either if the provider is malevolent, or if a significant number of trustees behave differently towards the truster than toward the opinion provider in question.

We describe this technique in more detail in the remainder of this section: first we detail how the probability of accuracy is calculated, and then we show how opinions are adjusted and the combined reputation obtained. An example of how these techniques can be used is also given with the aid of a walkthrough scenario in [12] and [16].

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, which depends on the value of the current opinion under consideration, denoted $\hat{\mathcal{R}}_{a_{op},a_{te}} = (\hat{m}_{a_{op},a_{te}}, \hat{n}_{a_{op},a_{te}})$. Specifically, if E^r is the expected value of a beta distribution D^r, such that $\alpha^r = \hat{m}_{a_{op},a_{te}} + 1$ and $\beta^r = \hat{n}_{a_{op},a_{te}} + 1$, we can estimate the probability that E^r lies within some margin of error around $B_{a_{tr},a_{te}}$, which we call the accuracy of a_{op} according to a_{tr}, denoted as $\rho_{a_{tr},a_{op}}$. To perform this estimation, we consider the outcomes of all previous interactions for which a_{op} provided an opinion similar to $\hat{\mathcal{R}}_{a_{op},a_{te}}$ about a_{te}, to a_{tr}, for each a_{te}. Using these outcomes, we construct a beta distribution,

[4] More information on such alternative techniques can be found in [16] and Section 5.
[5] These are observations made by the truster after it has obtained an opinion.

D^o for which, if its expected value E^o is close to E^r, then a_{op}'s opinions are generally correlated to what is actually observed, and we can judge a_{op}'s accuracy to be high. Conversely, if E^r deviates significantly from E^o, then a_{op} has low accuracy.

The process of achieving this estimation is illustrated in Figure 3, in which the range of possible values of E^r and E^o is divided into five intervals (or bins), $bin_1 = [0, 0.2], ..., bin_5 = [0.8, 1]$. These bins define which opinions we consider to be similar to each other, such that all opinions that lie in the same bin are considered alike. This is necessary because we may never see enough opinions from the same provider to assess an opinion based on identical opinions in the past. Instead, the best we can do is consider the preceived accuracy of past opinions that do not deviate significantly from the opinion under consideration. In the case illustated in the figure, the opinion provider, a_{op}, has provided a_{tr} with an opinion with an expected value in bin_4. Now, if we therefore consider all previous interaction outcomes for which a_{op} provided an opinion to a_{tr} in bin_4, the portion of successful outcomes, and thus E^o, is also in bin_4, so $\rho_{a_{tr},a_{op}}$ is high. If subsequent outcome-opinion pairs were also to follow this trend, then D^o would be highly peaked inside this interval, and $\rho_{a_{tr},a_{op}}$ would converge to 1. Conversely, if subsequent outcomes disagreed with their corresponding opinions, then $\rho_{a_{tr},a_{op}}$ would approach 0.

More specifically, we divide the range of possible values of E^r into N disjoint intervals $bin_1, ..., bin_n$, then calculate E^r, and find the interval, bin^o, that contains the value of E^r. Then, if $\mathcal{H}_{a_{tr},a_{op}}$ is the set of all pairs of the form (O_{a_{tr},a_x}, $\hat{\mathcal{R}}_{a_{op},a_x}$), where $a_x \in \mathcal{A}$, and O_{a_{tr},a_x} is the outcome of an interaction for which, prior to being observed by a_{tr}, a_{op} gave the opinion $\hat{\mathcal{R}}_{a_{op},a_x}$, we can 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's expected value falls in bin^o. We then 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 those for which it was not (denoted C_{fail}). Based on these frequencies, the parameters for D^o can be defined 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 (see Equation 10).

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

Each truster performs these operations to determine the probability of accuracy of reported opinions. However, one implication of this technique is that the number (and size) of bins effectively determines an acceptable margin of error in opinion provider accuracy: the estimated accuracy of a larger set of opinion providers converges to 1 with large bin sizes, as opposed to small sizes.

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

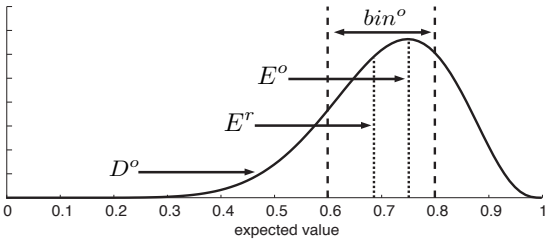

Fig. 3. Illustration of $\rho_{a_{tr},a_{op}}$ Estimation Process

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.

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, $\hat{\mathcal{R}}_{a_{op},a_{te}}$, is omitted. Third, let $\bar{D}$ be the result of adjusting the opinion distribution D^r, according to the process described below. 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. In essence, by adding $\hat{\mathcal{R}}_{a_{op},a_{te}}$ 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 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, the relatively small parameter values of d_1 compared to d_3 in the latter case, means that d_1 has virtually no impact on the combined result. Obviously, this is due to our method of reputation combination (see Equation 8), in which the parameter values are summed. This is important because it shows how, if left unchecked, an unfair rater could deliberately increase the weight an agent places on 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, respectively, the expected value and standard deviation of D^r, and the expected value and standard deviation of the uniform distribution, $\alpha = \beta = 1$, which represents a state of no information (see 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 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.

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

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

Once we have determined the values of $\bar{E}$ and $\bar{\sigma}$, we use Equations 13 and 14 to find the parameters $\bar{\alpha}$ and $\bar{\beta}$ of the adjusted distribution,[6] and from these we calculate adjusted values for $\hat{m}_{a_{op},a_{te}}$ and $\hat{n}_{a_{op},a_{te}}$, denoted as $\bar{m}_{a_{op},a_{te}}$ and $\bar{n}_{a_{op},a_{te}}$ respectively (see Equation 15). These scaled versions of $\hat{m}_{a_{op},a_{te}}$ and $\hat{n}_{a_{op},a_{te}}$ are then used in their place to calculate the combined trust value, as in Equation 8. Strictly speaking, $\bar{m}_{a_{op},a_{te}}$ and $\bar{n}_{a_{op},a_{te}}$ are not frequencies as are their unadjusted counterparts, but have the same affect on the combined trust value as an equivalent set of observations made by the truster itself. In general, as $\rho_{a_{tr},a_{op}}$ approaches 0, both $\bar{m}_{a_{op},a_{te}}$ and $\bar{n}_{a_{op},a_{te}}$ will also approach 0. Thus, if $\rho_{a_{tr},a_{op}}$ is 0 then no observation reported by a_{op} will affect a_{tr}'s decision making in any way.

$$\bar{\alpha} = \frac{\bar{E}^2 - \bar{E}^3}{\bar{\sigma}^2} - \bar{E} \tag{13}$$

$$\bar{\beta} = \frac{(1-\bar{E})^2 - (1-\bar{E})^3}{\bar{\sigma}^2} - (1-\bar{E}) \tag{14}$$

$$\bar{m}_{a_{op},a_{te}} = \bar{\alpha} - 1 \quad , \quad \bar{n}_{a_{op},a_{te}} = \bar{\beta} - 1 \tag{15}$$

4 Empirical Evaluation

In this section we present the results of the 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; and Section 4.3 investigates the overall performance of TRAVOS when both direct experience and reputation are taken into account.

[6] A derivation of these equations is provided in [16].

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 $p \in \mathcal{P}$. Before each experiment the behaviour of each provider and reputation source agent is set. 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 calculated the consumer's mean estimation error for the episode (see Equation 16), giving us a measure of the consumer's performance on assessing the provider population as a whole. Note that the value of this metric varies depending on the distribution of values of $B_{c,p}$ over the provider population. So, 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, that is, the set $\{0, 0.01, \ldots, 0.99, 1\}$.

$$avg\_estimate\_err = \frac{1}{N} \sum_{i=1}^{N} abs(\tau_{c,p_i} - B_{c,p_i}),$$
$$\text{where } N \text{ is the no. providers.} \tag{16}$$

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 discussed below have been tested for statistical significance using *Analysis of Variance* techniques and *Scheffé* tests. It should also be noted that although the results presented are obtained from computer simulations relating to our marketplace scenario, their scope extends to real world computer systems such as large scale open systems and peer-to-peer networks.

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) (discussed in Section 5). Like TRAVOS, this uses the beta family of probability functions to calculate the posterior probability of a trustee's behaviour holding a certain value, given past

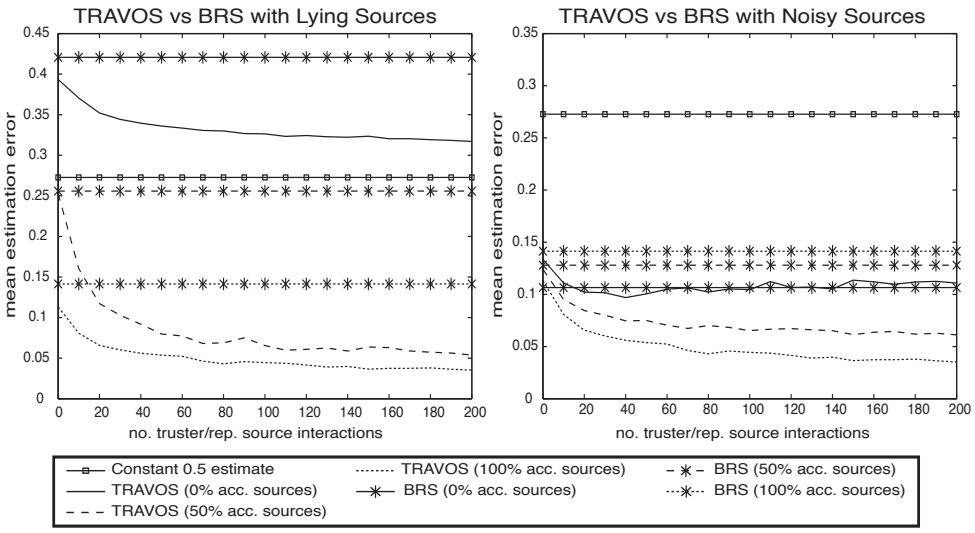

Fig. 4. TRAVOS Reputation System vs BRS

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

interactions with that trustee. 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, while BRS assumes that the majority of reputation sources provide an accurate opinion, and 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 in which consumers have no personal experience, and must therefore rely on reputation alone.

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% (Table 2). Figure 4 shows the mean estimation error of TRAVOS and 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 $p \in \mathcal{P}$. This is 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 less than or equal to that of BRS. Moreover, estimation errors decrease significantly for TRAVOS as the number of consumer to reputation source interactions increases, while 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, however performance was better because noisy source opinions are generally not as misleading as lying source opinions. TRAVOS still outperforms 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 is less than 10. We believe this occurs because the gaussion noise added to such opinions had a mean of 0, so noisy sources still provided accurate information on average. Thus, the BRS approach of removing outlying opinions may be successful at removing those noisy opinions that deviate significantly from the mean on any given cycle. However, this advantage decreases as TRAVOS learns which opinions to avoid.

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 5. 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.[7] This

[7] 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.

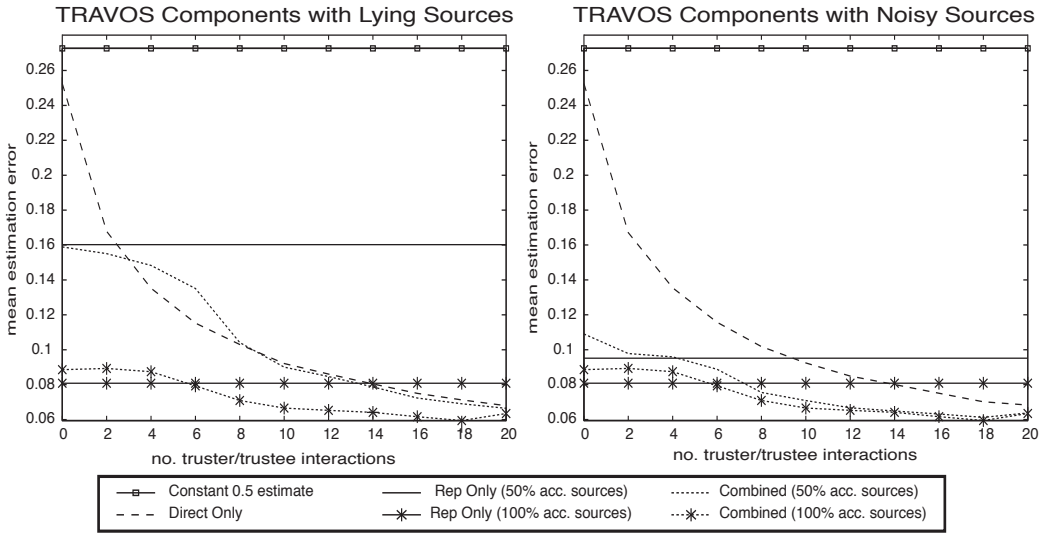

Fig. 5. TRAVOS Component Performance

can be attributed to the fact that TRAVOS will usually 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, but these results are not illustrated in this paper.

5 Related Work

There are many computational models of trust, a review of which can be found in [13]. A more detailed comparision of TRAVOS to related work can also be found in [16]. Generally, however, models not based on probability theory (e.g. [7, 14, 20]) calculate trust from hand-crafted formulae that yield the desired results, but that can be considered somewhat ad hoc (although approaches using information theory [15] and Dempster-Shafer theory [19] also exist).

Probabilistic approaches are not commonly used in the field of computational trust, but there are some models in the literature (e.g. [11, 8, 18, 10]). In particular, the Beta Reputation System (BRS) [8] is a probabilistic trust model like TRAVOS, which is based on the beta distribution. The system is centralised and specifically designed for online communities. It works by users giving ratings to the performance of other users in the community, where ratings consist of a

single value that is used to obtain positive and negative feedback values. These 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.* [18] 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 which case the only agents that will provide an opinion are those with an incentive to lie. In TRAVOS, opinion 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 [19]. 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. 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 in which 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 gaining experience, 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 will consider 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 as prior information to incorporate into trust assessment within the Bayesian framework. 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. Using this as prior information should not only improve the overall accuracy of trust assessment, but should also handle bootstrapping. That is, when neither the truster or its opinion providers have previous experience with a trustee, the truster can still assess the trustee based on other information it may have available.

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) and earlier versions of this paper appeared in EUMAS 2005 and [17].

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, 3rd edition, 2002.
3. C. Dellarocas. Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems. In *Proceedings of the 21st International Conference on Information Systems*, pp. 520–525, Brisbane, Australia, December 2000.
4. D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, 2002.
5. I. Foster, N. R. Jennings, and C. Kesselman. Brain meets brawn: Why grid and agents need each other. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems*, pp. 8–15, New York, USA, July 2004.
6. D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, chapter 13, pp. 213–237. Basil Blackwell, 1988. Reprinted in electronic edition from Department of Sociology, University of Oxford.
7. T. D. Huynh, N. R. Jennings, and N. Shadbolt. Developing an integrated trust and reputation model for open multi-agent systems. In *Proceedings of the 7th International Workshop on Trust in Agent Societies*, pp. 62–77, New York, USA, 2004.

8. A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Conference on Electronic Commerce*, Bled, Slovenia, June 2002.
9. A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 2005. (to appear).
10. T. Klos and H. L. Poutré. Using reputation-based trust for assessing agent reliability. In *Proceedings of 7th International Workshop on Trust in Agent Societies*, pp. 75–82, New York, USA, 2004.
11. L. Mui, M. Mohtashemi, and A. Halberstadt. A computational model of trust and reputation. In *Proceedings of the 35th Hawaii International Conference on System Science*, volume 7. IEEE Computer Society Press, 2002.
12. J. Patel, W. T. L. Teacy, N. R. Jennings, and M. Luck. A probabilistic trust model for handling inaccurate reputation sources. In P. Hermann, V. Issarny, and S. Shiu, editors, *Proceedings of the 3rd International Conference on Trust Management*, volume 3477 of *LNCS*, pp. 193–209, Rocquencourt, France, May 2005. Springer-Verlag.
13. S. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19(1):1–25, 2004.
14. J. Sabater and C. Sierra. Regret: A reputation model for gregarious societies. In *Proceedings of the 4th Workshop on Deception Fraud and Trust in Agent Societies*, pp. 61–70, 2001.
15. C. Sierra and J. Debenham. An information-based model for trust. In *Proceedings of 4th International Joint Conference on Autonomous Agents and MultiAgent Systems*, pp. 497–504, Utrecht, the Netherlands, 2005.
16. W. T. L. Teacy. An investigation into trust & reputation for agent-based virtual organisations. Technical report, ECS, University of Southampton, 2005.
17. W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model. In *Proceedings of 4th International Joint Conference on Autonomous Agents and MultiAgent Systems*, pp. 997–1004, Utrecht, the Netherlands, 2005.
18. A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the 7th International Workshop on Trust in Agent Societies*, New York, USA, 2004.
19. B. Yu and M. P. Singh. Detecting deception in reputation management. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems*, pp. 73–80, Melbourne, Australia, July 2003. ACM Press.
20. G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms in online marketplaces. In *Proceedings of 32nd Hawaii International Conference on System Sciences*, volume 8. IEEE Computer Society Press, 1999.