A Hierarchical Bayesian Trust Model based on Reputation and Group Behaviour

W. T. Luke Teacy[1], Nicholas R. Jennings[1], Alex Rogers[1] and Michael Luck[2]

[1] Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK
{wtlt,nrj,acr}@ecs.soton.ac.uk
[2] King's College London, London, WC2R 2LS, UK
michael.luck@kcl.ac.uk

Abstract. In many systems, agents must rely on their peers to achieve their goals. However, when trusted to perform an action, an agent may betray that trust by not behaving as required. Agents must therefore estimate the behaviour of their peers, so that they may identify reliable interaction partners. To this end, we present a Bayesian trust model (HABIT) for assessing trust based on direct experience and (potentially unreliable) reputation. Although existing approaches claim to achieve this, most rely on heuristics with little theoretical foundation. In contrast, HABIT is based on principled statistical techniques; can be used with any representation of behaviour; and can assess trust based on observed similarities between groups of agents. In this paper, we describe the theoretical aspects of the model and present experimental results in which HABIT was shown to be up to twice as accurate at predicting trustee performance as an existing state-of-the-art trust model.

1 Introduction

Trust constitutes an important facet of multi-agent systems research since it provides a form of distributed social control within highly dynamic and open systems whereby agents form opinions about others based on their own past interactions, as well as from the reports of other agents (1). Now, in many dynamic open systems, such as e-marketplaces, the Grid, and peer-to-peer networks, agents have to interact with one another to achieve their goals—for example by purchasing services or information from each other. Here, agents may be self-interested, and when trusted to perform an action for (or provide information to) another, may betray that trust by not performing the action as required. In addition, due to the size of such systems, agents will often have to interact with agents with which they have little or no past experience. There is thus a need for models of trust that will ensure good interactions among software agents in large scale open systems.

To this end, a number of trust models and strategies have been proposed (see (1; 2) for a full review) to deal with distinct aspects of the interactions between agents [e.g. to model the context of interaction (3), to explore the trustworthiness of unknown agents (4), and to deal with innaccurate information (5; 6)]. However, existing approaches suffer from at least one of the following three limitations. First, they may rely on heuristics that lack a strong theoretical foundation, which makes it difficult to characterise how they should perform under different conditions or to define what their optimal performance should be. Second, they may only be able to make predictions about an agent if its behaviour is represented in a specific way. For example, trust models [including (5) and (6)] typically require an agent's behaviour to be represented by a small discrete set of labels, such as {cooperate,defect} or {good,medium,bad}.

Third, they may only make predictions about an agent based on previous observations of its own behaviour. In general, these observations may be made directly by the truster, or they may be reports of third party experience, commonly known as the trustee's *reputation*. The problem with this is that a truster can only assess a trustee if it has access to a sufficient number of observations of the trustee's past behaviour. However, when a trustee enters a system for the first time, such information may not be available because it has not yet interacted with any other agent. This is a problem, particularly in systems susceptible to *whitewashing* (7), in which agents can adopt a new identity to avoid a previously obtained bad reputation.

In this paper, we address these limitations by introducing the Hierarchical And Bayesian Inferred Trust Model (HABIT), which applies Bayesian analysis to assess trustee behaviour without the need for heuristics. Unlike previous (statistically principled) trust models, HABIT does not restrict how agent behaviour is represented, and so can be easily adapted to make predictions about any number of discrete and continuous aspects of behaviour. For example, a car insurance policy might be judged on price, represented by a continuous real number, and whether or not it offers roadside repair, represented by a discrete (binary) variable. Moreover, HABIT provides a pragmatic solution to the whitewashing problem by allowing a truster to assess agents for which there is little or no previous experience. To do so, it searches for correlations in the behaviour of groups of known agents, and uses this to predict the behaviour of other agents with similar attributes. Earlier solutions to this problem, such as proposed by (7), typically suggest treating unknown agents as completely unreliable. However, this unfairly penalises potentially trustworthy agents that are yet to gain a good reputation. In contrast, HABIT can learn the reliability of newcomers in general, and so can adapt its decisions to account for the reliability of newcomers found in practice.

In the following sections, we elaborate on these claims and detail the theoretical basis for HABIT. Specifically, the rest of this paper is structured as follows: Section 2 introduces the basic notation used throughout the paper; Section 3 presents the HABIT model; Section 4 discusses how this general model can be applied to different application domains; Section 5 details a Monte Carlo sampling algorithm, which can be used to perform practical inference in a large number of possible instances of the HABIT model; Section 6 presents experimental results in which HABIT was shown to be up to twice as accurate at predicting trustee performance as an existing state-of-the-art trust model; and finally, Section 7 summarises the main properties of the model and discusses future work.

2 Basic Notation

Before introducing our trust model, it is necessary to define some basic notation. Specifically, in a MAS consisting of n agents, we denote the set of all agents as $\{1, 2, ..., n\} = \mathcal{A}$. Over time, interactions take place between distinct pairs of agents from $\mathcal{A}$, during which one of these agents is obliged to provide a service to the other. In each case, the agent receiving the service is the truster, denoted tr, and the agent providing the service is the trustee, denoted te.

With an aim to assess trustee performance, a truster records the outcome of each interaction as it *perceives* it, which is denoted as $O_{tr \to te}$. This is the outcome of interacting with te from the perspective of tr. From this interpretation, bilateral interactions in which both parties have obligations to each other can be seen as two separate interactions in which each agent plays the role of truster and trustee in turn. If such an event occurs between agents 1 and 2, then this will result in two recorded outcomes, denoted $O_{1 \to 2}$ and $O_{2 \to 1}$. However, it is important to note

that $O_{1\to 2}$ and $O_{2\to 1}$ are not necessarily equal, as each agent may represent the outcome only in terms that are relevant to it. For example, if 1 sells high quality apples to 2, for which 2 does not pay, then from 2's perspective the interaction results in the possession of some high quality apples, while from 1's perspective, goods are lost without payment.

With this in mind, it is useful to define a number of outcome instances, and sets involving them. First, we define the set of all possible outcomes in a particular context, $\mathcal{C}$, as $\mathcal{O}^{\mathcal{C}}$. Here, a context specifies both the type of interaction from which outcomes are derived and the way it is recorded. For instance, in the example given above, we could have $O_{2\to 1} \in \mathcal{O}^{apples}$ and $O_{1\to 2} \in \mathcal{O}^{money}$, where each context is defined in terms of the services received by the respective truster. Building on this, we divide time into discrete steps starting from time 0, and denote the outcome of an interaction that occurred between tr and te at time t as $O^{t}_{tr\to te}$. In general, we wish to allow any number of interactions to occur between any agents at any time. However, to simplify our discussion, we will assume that at most one interaction can occur between a given truster and trustee in a given time step, and that each interaction is complete by the end of the time step in which it is said to occur. Furthermore, we denote the current time as t', and the set of all outcomes between tr and te from time t to $t + r$ as $O^{t:t+r}_{tr\to te}$. Thus, the history of all interactions between tr and te is given by $O^{0:t'}_{tr\to te}$.

3 The HABIT Model

Now that we have a formal language for discussing interactions between agents, we can investigate how, in general terms, a truster can assess the value of interacting with a trustee, so that it may choose between a number of competing trustees, or perhaps choose a different course of action altogether. Intuitively, a truster's aim is to choose actions that are likely to result in outcomes that it prefers, such as receiving a high quality of service from a reliable service provider. To achieve this in a principled way, we turn to decision theory (8), which states that a rational agent should always act to maximise its *expected utility* (EU). Assuming that $O_{tr\to te}$ has a continuous domain, this is calculated as follows:

$$EU = \int_{\mathcal{O}^{\mathcal{C}}} U(O_{tr\to te}) p(O_{tr\to te})\, dO_{tr\to te} \tag{1}$$

Here, $p(O_{tr\to te})$ is the probability distribution of $O_{tr\to te}$, and $U : \mathcal{O}^{\mathcal{C}} \to \mathbb{R}$ is a utility function — for which higher values indicate more preferred outcomes. Although the utility function depends entirely on the truster's domain specific goals and preferences, the calculation of $p(O_{tr\to te})$ can be addressed in more general terms. To achieve this, HABIT comprises two types of component: a *reputation* model, which accounts for group behaviour and reputation by representing the relationships that exist between the behaviour and observations of different agents; and multiple *confidence* models, one for each truster-trustee pair, which account for direct experience by representing how a trustee's behaviour is perceived by each truster. Together, these two component types form a two-layer hierarchy, in which the confidence models form the lower layer, which deals with individual agent behaviour, and the reputation model forms the higher layer, which models the connections between the behaviour of different agents (trustees and observers). These components form a Bayesian network, as illustrated in Figure 1, comprising the random variables described below.

More specifically, for each truster, tr, and trustee, te, the role of the confidence model is to represent the probability distribution, $p(O_{tr\to te}|\theta_{tr\to te})$, of all observations $O_{tr\to te}$, where

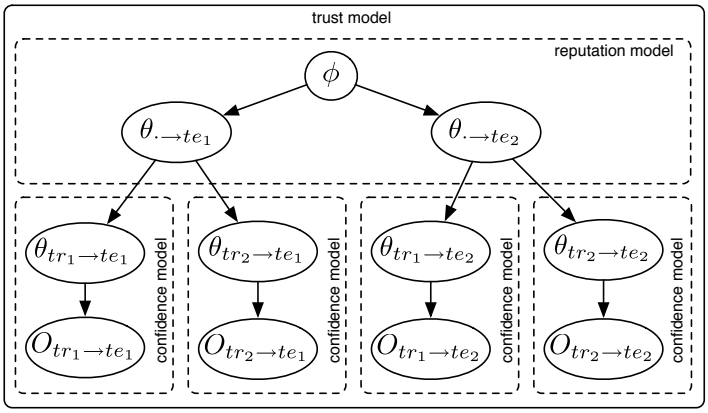

Fig. 1. Bayesian Network Inference with Correlated Behaviour Distributions

$\theta_{tr \to te}$ is a parameter vector[3] that specifies the distribution. From tr's perspective, this parameter vector is of primary interest because it characterises how te is likely to behave during an interaction and, consequently, what utility tr can expect to receive. For example, suppose that te is a search engine from which tr requests information, and $O_{tr \to te}$ is a real number specifying the time taken to respond to a request. If, over multiple requests, $O_{tr \to te}$ is assumed to follow a Gaussian distribution, then $\theta_{tr \to te}$ could comprise the mean, μ, and variance, σ^2, of the distribution. Small values of μ would imply that, on average, te is quick to respond to a request, while small values of σ^2 would imply that it does so consistently. Similarly, large values for μ and σ^2 would result in long average response times that vary greatly from order to order. The effect of these values on an agent's expected utility would depend on the precise definition of its utility function. Intuitively, however, a truster is likely to derive greater expected utility by interacting with a trustee that delivers low mean and variance than by interacting with an agent with high mean and variance.

Moreover, it is not necessary that every truster represents trustee behaviour using the same parameter model. For instance, while one truster may represent behaviour only in terms of response time (which would be reasonable if its utility function only depended on this factor), another truster may also have preferences involving the number of relevant hits. In this case, the joint distribution of these two aspects of behaviour would need to modelled, possibly using a multivariate Gaussian distribution, or some more appropriate combination of conditional distributions. Unfortunately, an agent is unlikely to know the true values of these parameters in practice, and so must perform inference given the evidence available. From a Bayesian perspective, this is achieved by treating the parameters themselves as random variables, and modelling their distributions based on an agent's beliefs and observations. In the case of a truster assessing a trustee in a single context using only its direct experience with the trustee in that context, this process is straightforward and can be achieved using standard techniques (9). The diffi-

[3] In this paper, we define HABIT in terms of parameter vectors, rather than sets, so that all equations involving parameters have their intended interpretation according to linear algebra. However, in some cases, we also use set notation to define new parameter vectors in terms of others.

cultly arises when a truster has little or no direct experience of a trustee's behaviour in a given context, and so must rely on observations of other agents, or third party observations.

In these cases, it is difficult to determine how much (if any) information such experience can give about an agent. For example, third party observations of a search engine may be unreliable if the source of those observations is lying, if it assesses trustee behaviour according to different criteria, or if the search engine delivers varying quality of service to different users. Likewise, there is no guarantee that any search engines will offer similar quality of service, and so an agent's experience of one service may not provide useful information about the likely behaviour of another. Nevertheless, some search engines may provide a similar quality of service (for example, if they employ similar technology) and most probably offer similar quality of service to different users. The key challenge — and the main contribution of this paper — is to determine precisely what these relationships are, so that an agent can make valid generalisations to assess an agent based on all available observations from different (but related) sources and contexts. This is achieved automatically, based on the data observed in any given context, by applying the reputation model illustrated in Figure 1. Here, each $\theta_{\cdot \to j}$ is a vector of all parameters used to model trustee j by all known observers. That is, $\theta_{\cdot \to j}$ is formed by concatenating all parameter vectors, $\theta_{i \to j}$, where $i \in \mathcal{A}$ (see Table 1). In Figure 1, for example, $\theta_{\cdot \to te_1}$ therefore contains $\theta_{tr_1 \to te_1}$ and $\theta_{tr_2 \to te_1}$, hence they are dependent as represented by the connecting vertices. As described above, the figure also shows that, for each i and j, an interaction outcome $O_{i \to j}$ depends on the corresponding parameter vector, $\theta_{i \to j}$. However, we now introduce an additional vector, ϕ, that specifies the *joint distribution* of all parameter vectors for each pair of agents, where each $\theta_{\cdot \to j}$ is independent and identically distributed (i.i.d.) according to ϕ. Intuitively, this means that ϕ characterises the relationship that exists between the distributions of observations made by different sources of different trustees. This allows a truster to perform inference about a *specific* trustee, given observations of *any* trustee from any source (direct or third party). However, just as an agent is unlikely to know the precise value of any of the parameter vectors, $\theta_{i \to j}$, it is also unlikely to know the value of ϕ. Nevertheless, it is possible for a truster to learn about ϕ using Bayesian techniques, just as it can learn about $\theta_{tr \to te}$ through repeated interaction with te. Moreover, it can then apply its knowledge of ϕ to make more informed inferences about te based on all available evidence.

Vector	Set Definition
θ	$\{\theta_{i \to j} \mid i \in \mathcal{A}, j \in \mathcal{A}\}$
$\theta_{i \to \cdot}$	$\{\theta_{i \to j} \mid j \in \mathcal{A}\}$
$\theta_{\cdot \to j}$	$\{\theta_{i \to j} \mid i \in \mathcal{A}\}$
Φ	$\theta \cup \{\phi\}$
$\Phi_{\cdot \to j}$	$\theta_{\cdot \to j} \cup \{\phi\}$

Table 1. Parameter vectors defined in terms of the sets of parameters they comprise.

4 Applying the HABIT Model to Specific Domains

So far, we have discussed the theoretical aspects of HABIT in general terms, independent of any particular scenario. However, since each application places its own unique requirements on how trust should be modelled, the parameters and probability distributions that define HABIT must be instantiated to best suit the target scenario. Therefore, in this section, we outline the steps required to apply HABIT to a given domain, and discuss the issues that should be considered

when fulfilling these steps and how they can be addressed. As described in the previous section, the aim of HABIT is to enable a truster to estimate the expected utility of interacting with a trustee in a specific context (see Equation 1). To achieve this, a truster can make use of its own personal observations $O_{tr\rightarrow te}^{0:t'}$, and all observations $O_{i\rightarrow j}^{0:t'}$ reported by an arbitrary observer, i, about an arbitrary trustee, j. More precisely, if $\mathcal{R}$ is the set of all pairs (i, j), including (tr, te), such that $O_{i\rightarrow j}^{0:t'}$ is known to tr, then the goal is to estimate:

$$EU|\mathcal{E} = \int_{\mathcal{O}^C} U(O_{tr\rightarrow te})p(O_{tr\rightarrow te}|\mathcal{E})\ dO_{tr\rightarrow te} \tag{2}$$

where $\mathcal{E} = \bigcup_{(i,j)\in\mathcal{R}} O_{i\rightarrow j}^{0:t'}$. Here, the *predictive* distribution, $p(O_{tr\rightarrow te}|\mathcal{E})$, is defined in terms of the hidden parameter variables, Φ, and their relationship to the observed outcomes. However, precisely how these parameters are defined and how they affect the observed outcomes is domain dependent, and so is not stipulated by the generic HABIT model. Instead, these must be instantiated to suit the specific requirements of the target domain.

In particular, these requirements may comprise constraints on the computational resources available to perform inference with the model, the level of accuracy required in estimating expected utilities and the aspects of trustee behaviour that affect a truster's utility. In any case, to fully instantiate the model, four sets of probability distributions must be defined along with their associated domains, probability density functions (p.d.f.s) and parameters: (1) for each confidence model (i.e. each truster-trustee pair), the conditional distribution of interaction outcomes, $O_{i\rightarrow j}$, given a chosen parameter vector, $\theta_{i\rightarrow j}$; (2) the prior distribution (that is, without knowledge of any observed outcomes) of each parameter vector $\theta_{i\rightarrow j}$; (3) the conditional distribution of all joint parameter vectors, $\theta_{\cdot\rightarrow j}$, given the hyperparameter vector ϕ; and (4) the prior distribution of the hyperparameter vector, ϕ. Although having this number of unspecified components may seems like a weakness of the model, this is the minimum required to allow HABIT the flexibility to be adapted to any domain in an unconstrained way. Furthermore, choosing these distributions is a straightforward matter, which can be achieved by matching the specific requirements of an application to the well known properties of standard distributions. Moreover, from these four sets of distributions, all other necessary conditional and marginal distributions can be derived by applying the standard rules of Bayesian analysis. In particular, the following aspects should be considered when instantiating the distributions above:

Model Sophistication and Time Complexity: HABIT can be instantiated by combining a number of existing parameter models, ranging from discrete distributions (which are efficient, but can only make simple predictions), to infinite mixture models (10) (which require more computational resources, but can make more sophisticated predictions). Thus, the choice of model depends on the requirements and resources of the target domain.

Dynamic Behaviour: In most situations, it is reasonable to expect agent behaviour to change over time. In HABIT this can be modelled by instantiating the confidence model parameters with any off-the-shelf model of time-dependent phenomena, including hidden Markov models (11) and Gaussian processes (12). The reputation model need not be affected because it only needs to refer to current agent behaviour.

Whitewashers and Group Behaviour: HABIT's ability to assess group behaviour can be applied in various ways. For example, it can maintain a single reputation model for all agents it encounters, thereby enabling predictions about relatively unknown agents by generalising from the observed behaviour of all other trustees. However, a more significant possibility

Algorithm 1 General Monte Carlo Algorithm for Expected Utility Estimation.

Require: $n > 0$ {Larger values of n result in more accurate expected utility estimates.}
$EU \Leftarrow 0$
for $k = 1$ to n **do**
  for all $(i, j) \in (\theta/\theta_{tr\to te}) \times (\theta/\theta_{tr\to te})$ **do**
    $\theta_{i\to j} \Leftarrow$ sample from $p(\theta_{i\to j}|O^{0:t'}_{i\to j})$
  end for
  $\phi \Leftarrow$ sample from $p(\phi|\theta/\theta_{tr\to te})$
  $\theta_{tr\to te} \Leftarrow$ sample from $p(\theta_{tr\to te}|\theta_{\cdot\to te}/\theta_{tr\to te}, \phi, O^{0:t'}_{tr\to te})$
  $O_{tr\to te} \Leftarrow$ sample from $p(O_{tr\to te}|\theta_{tr\to te})$
  $EU \Leftarrow EU + U(O_{tr\to te})/n$
end for{EU is now an estimate of tr's expected utility for interacting with te.}

is to first cluster agents into non-overlapping groups and maintain a *separate* reputation model for each group. For example, to deal with whitewashing, agents can be grouped based on the length of time they have been in a system. Thus, if whitewashing is an issue, groups containing new agents will be less reliable, and so could be treated appropriately.

5 Performing Inference with the HABIT Model

As with most nontrivial Bayesian models, performing all inference analytically with HABIT is unfeasible in general.[4] Instead, tractable algorithms must be sought that can approximate the optimal Bayesian solution within a reasonable amount of time. In this section, we propose one such algorithm that, through the application of Monte Carlo Sampling, can be applied to any instance of the general model. In line with the previous section, the aim of this algorithm is to estimate the expected utility for interacting with a trustee, given a truster's own personal observations and reported reputation (Equation 2). This is usually intractable to evaluate analytically because the calculation of the predictive distribution, $p(O_{tr\to te}|\mathcal{E})$, involves integration over all the parameters in the model. Despite this, it is typically possible to draw a set of n samples, $\{O_1, \ldots, O_n\}$, from the predictive distribution, such that $EU|\mathcal{E} \approx \sum_{i=1}^{n} U(O_i)$ with the accuracy of the estimate increasing as n becomes large (13). To achieve this, we take advantage of the conditional independence relations in HABIT to decompose the task of sampling from $p(O_{tr\to te}|\mathcal{E})$ into a number of simpler sampling problems. This is achieved in three steps. First, from the standard properties of random variables, we know that sampling from $p(O_{tr\to te}|\mathcal{E})$ is equivalent to sampling from the joint distribution $p(O_{tr\to te}, \Phi|\mathcal{E})$ (see Table 1); the generated values for Φ are simply discarded because they are not required. Second, we express this joint distribution in terms of simpler conditional distributions as follows:

$$
\begin{aligned}
p(O_{tr\to te}, \Phi|\mathcal{E}) &= p(O_{tr\to te}|\Phi, \mathcal{E})p(\Phi|\mathcal{E}) \qquad (3)\\
&= p(O_{tr\to te}|\theta_{tr\to te})p(\theta_{tr\to te}|\Phi_{\cdot\to te}/\theta_{tr\to te}, O^{0:t'}_{tr\to te})p(\phi|\theta/\theta_{tr\to te})p(\theta/\theta_{tr\to te}|\mathcal{E})\\
&= p(O_{tr\to te}|\theta_{tr\to te})p(\theta_{tr\to te}|\Phi_{\cdot\to te}/\theta_{tr\to te}, O^{0:t'}_{tr\to te})p(\phi|\theta/\theta_{tr\to te}) \prod_{\substack{\theta_{i\to j}\in\\ \theta/\theta_{tr\to te}}} p(\theta_{i\to j}|O^{0:t'}_{i\to j})
\end{aligned}
$$

[4] However, under certain circumstances, analytical solutions for this model are possible; for example, see Section 6.

Finally, according to standard theory, sampling from the full joint distribution can be achieved by sampling from each of the component distributions shown in Equation 3, and using the generated samples from the rightmost p.d.f.s in the equation to satisfy the conditional variables for the p.d.f.s to the left. This process is summarised in Algorithm 1.[5] At this level of detail, the algorithm is completely general, and can be applied (without modification) to any choice of parameter models that allows sampling from the distributions referred to in Algorithm 1. Of these, $p(O_{tr\to te}|\theta_{tr\to te})$ can be chosen directly to suit the target application,[6] while the other three distributions should be derived according to Bayes rule, with suitable prior distributions chosen from ϕ and each $\theta_{i\to j}$. Ideally, each of these distributions will have forms that allow independent sampling. That is, it is desirable to draw samples from these distributions that are independent of each other and identically distributed according to the desired distribution. If this is possible, the number of samples required to accurately estimate the expected utility can be very low[7], and it is straightforward to calculate the estimation error (w.r.t. the utility) using the standard deviation of the generated samples.

In most cases, this can be achieved by choosing conjugate prior distributions, which lead to simple analytical formulas for the posterior parameter distributions given the evidence (14). However, efficient i.i.d. sampling is unlikely to be possible for $p(\theta_{tr\to te}|\Phi_{\cdot\to te}/\theta_{tr\to te}, O^{0:t'}_{tr\to te})$ because, apart from the trivial case where $O^{0:t'}_{tr\to te} = \emptyset$ (i.e. a truster has no direct experience with a trustee), it is difficult to ensure that $p(\theta_{tr\to te}|\Phi_{\cdot\to te}/\theta_{tr\to te})$ is conjugate with respect to $O^{0:t'}_{tr\to te}$. In such cases, there are two existing types of solution to choose from: (1) Markov Chain Monte Carlo MCMC methods, which are a class of algorithms for generating a sequence of samples, where each sample depends on the previous sample in the sequence (13); and variational methods, which are used to estimate complicated probability distributions using one of a number of simpler types of distribution (13). Where they exist, both types of solution can readily be integrated into our sampling algorithm without modification. In the case of variational methods, these can be used to approximate the problematic distribution(s), and subsequently, the approximate distributions can be used to generate i.i.d. samples in the normal way. For MCMC methods, the situation is similar; for example, suppose that an MCMC algorithm is used to simulate $p(\phi|\theta/\theta_{tr\to te})$ by generating a sequence of values labelled $\phi_1, \ldots, \phi_k$, such that, for each $i > 1$, $\phi_i \sim p(\phi_i|\theta/\theta_{tr\to te}, \phi_{i-1})$. It is perfectly fine to use these in Algorithm 1, in place of independent samples from $p(\phi|\theta/\theta_{tr\to te})$, with each ϕ_i being generated using different samples for $\theta/\theta_{tr\to te}$: convergence will still be guaranteed, albeit more slowly in terms of the total number of samples (13).

6 Empirical Evaluation

In principle, the innumerable ways in which HABIT can be instantiated allow for a wide range of properties to suit a variety of different applications. As such, we do not advocate any specific instantiation, but it is nevertheless useful to evaluate the general properties of HABIT by analysing its empirical performance in some specific cases. To this end, we now present experimental results that measure the performance of three instances of the generic HABIT model,

[5] In these equations, the symbol '/' is the set difference operator. Thus, x/y should be interpreted as a parameter vector consisting of all elements in x except for those in y.

[6] For example, if $O_{tr\to te}$ is the number of relevant hits returned by a search engine, then $p(O_{tr\to te}|\theta_{tr\to te})$ could be modelled as a Poisson distribution (14) with unknown mean $\theta_{tr\to te} = \lambda$.

[7] Typically, 30 to 100 (i.i.d.) samples will be sufficient for most applications (13).

labelled *DP, GD-Improper*, and *GD-Conjugate*. For evaluation purposes, all three adopt a discrete representation of trustee behaviour, which enables objective comparison between HABIT and existing trust models that are limited to such representations. Most notably, these include BLADE (6), which we use here as a benchmark because it is representative of the state-of-the-art among statistically principled trust models.

With regard to direct experience, all three instances model trust in the same way, by instantiating their confidence models such that each $O_{tr \to te} \in \{O_i\}_{i=1}^{k}$ is a discrete random variable, where $p(O_i) = \theta_{tr \to te}^{(i)}$, $\theta_{tr \to te} = \langle \theta_{tr \to te}^{(1)}, \ldots, \theta_{tr \to te}^{k} \rangle$, and $\theta_{tr \to te}$ is assigned a conjugate Dirichlet prior. In this respect, all three instances are not only equivalent to each other, but are also equivalent to many existing models of trust, including BLADE. Thus, in the special case where a truster has only its direct experience with which to assess a trustee, its beliefs will be identical if it uses any of these existing models, or one of the instances of HABIT described here. However, where the three instances differ (both from each other and existing models of trust) is in how they achieve the more complex task of assessing trust based on reputation and group behaviour. For this purpose, each instance uses a different reputation model:

- In DP, reputation is modelled by assuming that, for each trustee, $\theta_{\cdot \to te}$ is drawn from a *Dirichlet Process* (15) (which plays the role of ϕ). Significantly, this allows the predictive distribution (see Section 4) to be calculated efficiently and analytically without the need for Monte Carlo sampling. However, using this approach, if the number of observations of each trustee is high, relative to the number of encountered trustees, then a trustee's reputation may have little impact on inference, even if it provides useful information.
- In GD-Improper, reputation is modelled by assuming that, for each trustee, $\theta_{\cdot \to te}$ is drawn from a multivariate Gaussian distribution with unknown mean and covariance represented by ϕ, which in turn is assigned an *improper* prior distribution.
- GD-Conjugate is equivalent to GD-Improper, except that it places a conjugate prior distribution on ϕ, representing the prior belief that reputation provides no useful information about a trustee's behaviour.

In the following subsections, we outline the methodology used in our experiments (see Section 6.1) and discuss the performance of the models when used to perform inference based on group behaviour (see Section 6.2) and reputation (see Sections 6.3 and 6.4). In all of these experiments, the effect of direct experience and issues such as dynamic behaviour are not addressed because these are largely determined by the choice of parameter models. For most parameter models, these properties are already well understood, so we choose to focus on the issues directly related to the general HABIT model.

6.1 Experimental Design

To determine performance, all experiments were conducted in a simulated environment in which five trusters were asked to estimate their expected utility for interacting with a single *test* trustee based on group behaviour and reputation. Each truster represented one of five inference models: one for each instance of HABIT described above, one for BLADE, and one labelled *Prior*, which ignored all available evidence and instead relied on the prior assumption (shared by all agents) that all possible behaviours were equally likely.

To form their estimates, each of these trusters was presented with a variable number of direct observations and reputation reports about a number of *training* trustees. This provided a basis

on which trusters could (potentially) learn the average behaviour of a group of agents and the reliability of the *single* source from which reputation was obtained. Multiple reputation sources were not considered because, assuming that reports from different sources are independent, models that can extract more information form a single source will naturally do better with information from multiple sources. Moreover, the same observations and reputation information were always presented to all trustees to minimise excess variance in the results and, in particular, no direct observations were ever available for the test trustee, forcing the trusters to rely on reputation and group behaviour.

To measure performance, all experiments were run multiple times under fixed control conditions (including fixed numbers of training trustees and observations), where each run was based on a different randomly generated set of observations of a different set of randomly generated trustees. More specifically, the true parameter vectors for each trustee were randomly sampled in each run from a fixed Dirichlet distribution determined by the control conditions. Thus, any sampling bias due to a particular set of trustees or observations was avoided. At the end of each run, the absolute error in the expected utility estimate was recorded for each trustee in order to calculate confidence bounds on the mean error for each model.

In the following subsections, all results are plotted with error bars representing 95% confidence intervals on the mean absolute error. These are based on the standard assumption that the sampling distribution of the mean is a *t* distribution with degrees of freedom determined by the number of runs.[8] In addition, all claims that are made in the text are statistically significant (with p-values greater than 0.95) according to t-tests and analysis of variance (16).

6.2 Learning from Group Behaviour

To demonstrate the effect of group behaviour on performance, we ran a series of experiments in which trusters had to assess the test trustee, based solely on their direct experience with a number of training trustees. That is, in the absence of information pertaining directly to the trustee, the trusters had to rely on the reasonable *a priori* assumption that the test trustee would behave similarly to the training trustees, and so use any observed correlation between the behaviour of different training trustees to predict the behaviour of the test trustee. Here, there are three control variables that can impact performance: (1) the number of observations per training trustee, which dictates how certain a truster can be about an individual's behaviour; (2) the number of training trustees from which to infer the distribution of behaviours exhibited by the trustee population as a whole; and (3) the amount of similarity that exists between trustee behaviour, which determines how informative the behaviour of others is about a specific trustee. During this set of experiments, we controlled the first two factors directly, keeping the number of observations the same for each trustee in the interest of simplicity. To control the third, we generated trustee behaviour parameters from Dirichlets, using the magnitude of the Dirichlet hyperparameters as a proxy for the similarity between agents. More specifically, we allowed the mean of the Dirichlet to vary randomly by choosing it from a uniform distribution[9] at the start of each run. This mean was then multiplied by a chosen factor to form the α vector used to specify the distribution. Generally, high factor values (and thus higher values for α) would

[8] The number of runs performed for each experiment varied according to the compute time available to run the simulation, but typically ranged between 300 and 2000.

[9] The uniform distribution used here was equivalent to a Dirichlet distribution with all hyperparameters set to 1.

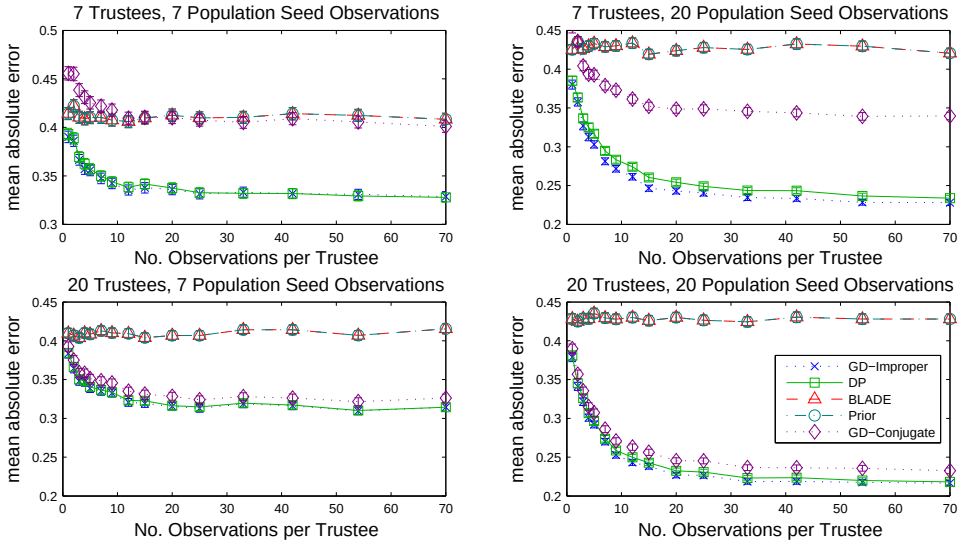

Fig. 2. Group Behaviour

result in trustee parameters that deviate less from the mean. Therefore, by increasing the magnitude, we increase the amount of information the training trustees provide about the unknown test trustee.

The ability to decipher this information in demonstrated in Figure 2, which (for example) shows the average error of each truster given varying numbers of observations per trustee, and values of 7 or 20 for both $\sum_{i=1}^{k} \alpha$ and the number of trustees. What is important about these results is that, as the evidence for behaviour correlation increases, all three instances of HABIT are able to perform significantly better than the prior, while at the same time perform no worse than the prior when no evidence for correlation exists. This follows as a direct result of the application of Bayesian inference in HABIT: the behaviour of known trustees is only allowed to influence predictions about other trustees to the extent supported by the evidence.

In addition to this, two other conclusions can be drawn from the figure. First, since BLADE does not allow for possible dependencies between trustees' behaviour, it performed no better than the prior agent. Second, although there was little difference between the predictions made by DP and GD-Improper, GD-Conjugate generally required more data to overcome its stronger prior that trustees' behaviour is generally dissimilar. However, as we shall see in the next section, strong priors do not always have a negative effect on performance, but can instead be used to provide a healthy scepticism in situations where inaccurate information is common.

6.3 Learning from Perfect Reputation Information

To compare the effect of reliable reputation on each truster's performance, we performed a set of experiments in which each truster received information from a perfect reputation source — one that, unknown to the trusters, provided observations that were as informative and identically distributed as each truster's direct observations. As before, direct experience was only available about the training trustees, forcing each truster to rely on external information to assess the test

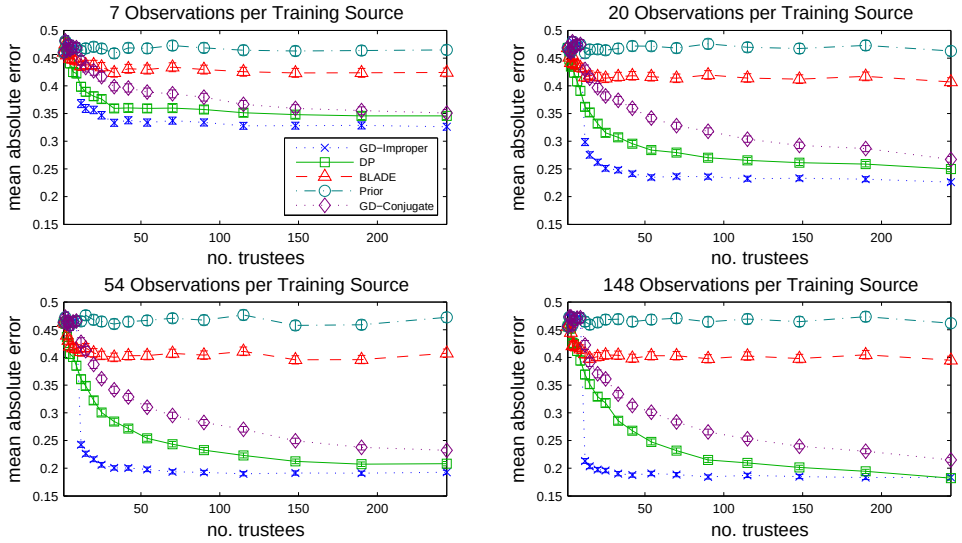

Fig. 3. Perfect Reputation

trustee. However, unlike the previous experiments, trustee behaviour parameters were always drawn from a uniform distribution, so that group behaviour could not provide any useful information over and above that provided by reputation. With these restrictions in place, the remaining variables that could impact performance are: (1) the number of direct observations available for each training trustee; (2) the number of observations reported by the reputation source about each training trustee; (3) the number of reported observations about the test trustee; and (4) the number of training trustees. Each of these variables was controlled directly with values for each ranging between 1 and 250. Figure 3 shows some of the results obtained when the number of direct and reported observations about each training trustee were kept equal at values of 7, 20, 54 and 148; the number of observations reported for the test trustee was 54; and the number of trustees varied between 1 and 250.

Unsurprisingly, these and other results show that all four control variables have a positive impact on performance as their values increase. However, although this is true for all the models evaluated, it is not true with equal measure. In particular, the same order observed in the previous section is maintained here, with GD-Conjugate requiring more information to overcome its prior than the other two instances of HABIT. However, the difference between DP and GD-Improper, which was insignificant before, is now strengthened in GD-Improper's favour. This is due to the way in which the Dirichlet Process is applied in the reputation model, which works best when significant numbers of trustees have been observed, relative to the number of observations of each agent. More significantly, however, all three instances of HABIT always perform at least as well as BLADE, and significantly outperform it as the amount of evidence increases. This highlights a problem with the strategy, used in BLADE, of trying to directly learn the correlation between a truster's direct observations and those reported by each reputation source. More specifically, for each reputation source, j, this approach attempts to learn the joint distribution of the observations $O_{tr \to te}$ and $O_{j \to te}$ as if they refer to the *same* interaction. However, only one of these can be observed for any particular interaction, because the under-

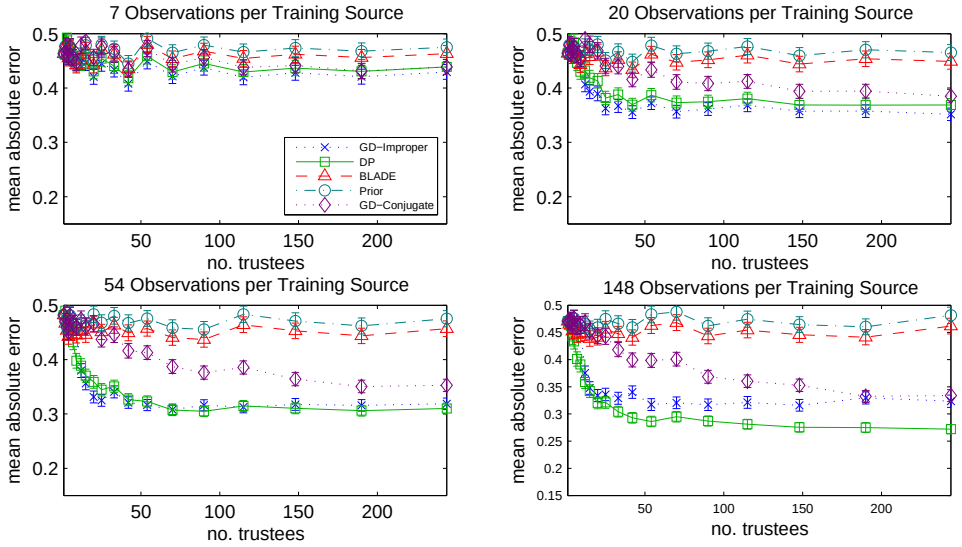

Fig. 4. 50% Noisy Reputation

lying assumption is that an interaction takes place privately between the trustee and a single observer, be that the truster itself or one of its reputation sources.

To overcome this, a truster must receive reports about multiple trustees. The mean behaviour of each trustee (direct and reported) then acts as a noisy observation of the joint value of $\langle O_{tr\to te}, O_{j\to te}\rangle$. If a trustee's behaviour is relatively consistent then this is almost as good as directly observing both values together. However, if a trustee's behaviour is relatively variable then the added uncertainty masks the correlation between the hidden outcome values. For discrete distributions, this problem reaches its peak for trustees that provide all possible outcomes with equal likelihood. From BLADE's perspective, this provides no information because it is impossible to distinguish between variance intrinsic to a reputation source's reports and the variance in the trustee's behaviour. HABIT takes a different approach: by looking for correlations between the *distributions* of reported outcomes, rather than the outcomes themselves, a report that accurately predicts a trustee's behaviour to be erratic is just as informative as one about a trustee that behaves consistently. This makes sense intuitively, and can explain the better performance exhibited by HABIT in these experiments.

6.4 Learning from Unreliable Reputation Information

Although the previous set of experiments show that all models can elicit useful information from good reputation, this benefit would be meaningless if they could not also deal with inaccurate reputation. In fact, the ability to cope with varying degrees of accuracy in reputation is precisely why we try to model its reliability in the first place. Thus, to evaluate this ability, we ran experiments under the same conditions outlined in the previous section, except that the reputation source reported independent random observations with a fixed probability. Specifically, with probability p, an observation reported by the reputation source was drawn from a uniform Dirichlet (independent of the trustee's behaviour), or with probability $1-p$, it was drawn

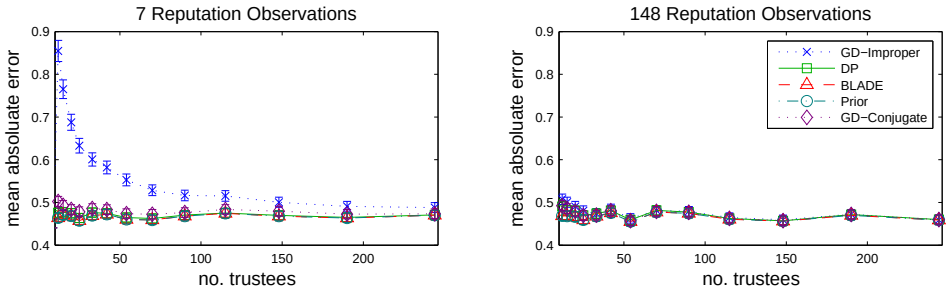

Fig. 5. Results when 148 observations were reported about the test trustee, all of which were independent of trustee behaviour.

from the trustee's behaviour distribution. In more detail, Figure 4 shows the results obtained, under equivalent conditions to Figure 3, when 50% of reported observations were independent of trustee behaviour. As one would expect, the performance of each model is similar to that obtained for perfect reputation, except that more evidence is required to reach equivalent levels of accuracy. Moreover, the lower bound on the average error is higher, due to the decrease in information provided by the reputation source. In particular, under these conditions, BLADE provides no significant gain over the prior. With respect to the evaluated instances of HABIT, these experiments show that, under some circumstances, the DP model can outperform both of the Gaussian based instances. This is due to the non-parametric nature of the Dirichlet Process, which theoretically places fewer constraints on the shape of joint parameter distributions, and so in some cases may be able to provide better results. However, more generally, this demonstrates that there is no model that performs best in every circumstance, and so it is useful to consider different models to meet the needs of specific applications.

In terms of reputation reliability, a more extreme case is illustrated in Figure 5. Here, all observations reported by the reputation source were independent of trustee behaviour, the number of direct observations was 7 for each training trustee, the number of reputation observations about the test trustee was 148, and the number of reported observations for the training trustees was 7 (left) or 148 (right). This shows that when the amount of evidence concerning the reliability of a reputation source is low, but the number of reported observations — and hence the reported confidence of the reputation source — is high, the GD-Improper model can be led astray, expecting spurious correlations between the reputation and trustee behaviour. This is because GD-Improper has no strong prior belief to suggest that a highly confident report is inaccurate, and so (in the absence of any evidence to the contrary) takes the reputation source on its word. As shown in the figure, this disadvantage disappears given more observations about greater numbers of trustees. However, it demonstrates that the good performance of some prior beliefs in some circumstances may come at a cost in others. In this case, the initially sceptical prior used in the GD-Conjugate reputation model pays off, preventing it from performing worse than the prior. Again, this reinforces the belief that no single trust model will perform best in every circumstance, and the choice of model should be made by finding one that works well in the variety of circumstances exhibited by the target domain. Nevertheless, some models are more robust in a wider range of circumstances than others, and our results show that the DP instance can exhibit surprisingly good performance in a range of circumstances, given that it is analytically tractable and therefore efficient to compute precisely. However, the instances of

HABIT evaluated here are only examples of what is possible. The key advantage of HABIT is that it provides a common framework for developing computationally feasible and statistically principled models of trust, which have a number of performance advantages over the current state-of-the-art. By using it as a basis for more sophisticated instances, HABIT provides the potential to solve a wide range of trust and reputation problems with a high degree of accuracy.

7 Conclusions

In this paper, we have developed a generic Bayesian trust model, which facilitates decision making by autonomous agents in service-oriented environments. Although several such models have previously been proposed, HABIT exhibits four key advantages, which together make a significant contribution to the state-of-the-art: (1) It provides a statistically principled and tractable framework, which can be adapted to assess trust in a wide range of scenarios with different modelling requirements; (2) It can assess trust based on reputation, even if the agents that supply this information use different representations of trust or provide inaccurate or intentionally misleading; (3) Even when a truster has no previous experience or reputation with which to assess a trustee, HABIT can still provide statistically principled predictions of the trustee's behaviour by considering the behaviour of other agents; (4) Through empirical evaluation we have shown that, when applied to discrete representations of trustee behaviour, HABIT outperforms BLADE, which represents the current state-of-the-art in statistical trust modelling. Therefore, although HABIT is not limited to discrete representations, it performs favourably to existing statistical trust models, which typically are limited to such representations.

Bibliography

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

[2] Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems **43**(2) (2007) 618–644

[3] Rettinger, A., Nickles, M., Tresp, V.: A statistical relational model for trust learning. In: Proc of AAMAS'08. (2008) 763–770

[4] Teacy, W.T.L., Chalkiadakis, G., Rogers, A., Jennings, N.R.: Sequential decision making with untrustworthy service providers. In: Proc. of AAMAS'08. (2008) 755–762

[5] Teacy, W.T.L., Patel, J., Jennings, N.R., Luck, M.: Travos: Trust and reputation in the context of inaccurate information sources. JAAMAS **12**(2) (2006) 183–198

[6] Regan, K., Poupart, P., Cohen, R.: Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In: Proceedings of the 21st National Conference on Artificial Intelligence. (2006) 206–212

[7] Zacharia, G., Moukas, A., Maes, P.: Collaborative reputation mechanisms in electronic marketplaces. In: Proceedings of 32nd Hawaii International Conference on System Sciences, Maui, Hawaii, IEEE Computer Society Press (1999)

[8] Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer (1993)

[9] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. 2 edn. Chapman & Hall/CRC (2003)

[10] Rasmussen, C.E.: The infinite gaussian mixture model. Advances in Neural Information Processing Systems **12** (2000) 554–560

[11] Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer (2005)

[12] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)

[13] Mackay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press (2003)

[14] DeGroot, M., Schervish, M.: Probability & Statistics. 3rd edn. Addison-Wesley (2002)

[15] Ghosh, J.K., Ramarmoorthi, R.V.: Bayesian Nonparameterics. Springer (2003)

[16] Cohen, P.R.: Empirical Methods for Artificial Intelligence. M.I.T. Press (1995)