An Efficient and Versatile Approach to Trust and Reputation using Hierarchical Bayesian Modelling

W. T. Luke Teacy[a], Michael Luck[b], Alex Rogers[a], Nicholas R. Jennings[a,c]

[a]*Electronics and Computer Science,*
University of Southampton, SO17 1BJ, UK.
{wtlt,acr,nrj}@ecs.soton.ac.uk
[b]*Department of Informatics,*
King's College London, WC2R 2LS, UK.
michael.luck@kcl.ac.uk
[c]*Department of Computing and Information Technology,*
King Abdulaziz University, Saudi Arabia

Abstract

In many dynamic open systems, autonomous agents must interact with one another to achieve their goals. Such agents may be self-interested and, when trusted to perform an action, may betray that trust by not performing the action as required. Due to the scale and dynamism of these systems, agents will often need to interact with other agents with which they have little or no past experience. Each agent must therefore be capable of assessing and identifying reliable interaction partners, even if it has no personal experience with them. To this end, we present HABIT, a Hierarchical And Bayesian Inferred Trust model for assessing how much an agent should trust its peers based on direct and third party information. This model is robust in environments in which third party information is malicious, noisy, or otherwise inaccurate. Although existing approaches claim to achieve this, most rely on heuristics with little theoretical foundation. In contrast, HABIT is based *exclusively* on principled statistical techniques: it can cope with multiple discrete or continuous aspects of trustee behaviour; it does not restrict agents to using a single shared representation of behaviour; it can improve assessment by using any observed correlation between the behaviour of similar trustees or information sources; and it provides a pragmatic solution to the *whitewasher* problem (in which unreliable agents assume a new identity to avoid bad reputation). In this paper, we describe the theoretical aspects of HABIT, and present experimental results that demonstrate its ability to predict agent behaviour in both a simulated environment, and one based on data from a real-world webserver domain. In particular, these experiments show that HABIT can predict trustee performance based on multiple representations of behaviour, and is up to twice as accurate as BLADE, an existing state-of-the-art trust model that is both statistically principled and has been previously shown to outperform a number of other probabilistic trust models.

Key words: trust, reputation, probabilistic trust, group-based trust

1. Introduction

In recent years, there has been an upsurge of interest in cross-enterprise service-oriented computing, which seeks to integrate computational resources seamlessly and dynamically across organisational boundaries. Perhaps the most prominent examples of such initiatives are the *semantic web* and *cloud computing*, but others include crowd-sourcing applications, peer-to-peer networks, sensor networks, and pervasive computing. A key attribute of all these systems is that individuals or organisations can offer data and software services that can be invoked remotely, with services from different organisations being used together toward some common goal. For example, companies such as Expedia[1] use web services from individual travel companies to find the best combination of travel and accommodation options to meet a customer's needs. Likewise, in a cloud computing scenario, a bioinformatician may process protein data

[1] http://www.expedia.com/

hosted by one institution with a remote application server for data mining run by another, and then store the result on a public cloud data service [36].

Now, when composing services in this way, there may often be a number of competing service providers that can fulfil a given requirement, each with different quality of service (QoS) characteristics and a different price. Thus to fulfil a user's requirements, decisions must be made about which providers to use, and these choices can affect both the quality and cost of the resulting solution. However, the multi-institutional nature of these systems means that there will invariably be uncertainty surrounding the capabilities and incentives of the individuals offering these services.

For instance, in the cloud computing example above, different data mining services may vary in terms of the accuracy of the machine learning algorithms they use, or the processor speed of the servers on which they run. However, it may not be in the best interests of a service provider to truthfully reveal its QoS attributes, or it may behave fraudulently to receive payment without providing the requested service. Moreover, the possible failure of some services should always be considered, so systems must adapt dynamically and automatically to meet changing circumstances, and be equipped with a means to identify the most reliable services. Similar QoS issues can also occur in crowd-sourcing applications, in which members of the public can collaborate to collect and share data about some common subject of interest [26]. For example, an online navigation service may provide real time traffic information using data collected from users with GPS enabled devices. However, depending on the type of device they use, the reliability of data from different users may vary. Moreover, due to their open nature, such systems are vulnerable to malicious users, who may try to disrupt the system by deliberately providing false data.

In light of this, it has been argued that such systems should be modelled as multi-agent systems (MAS) [28], since a core concern of MAS is the coordination of autonomous entities (agents), which must interact to achieve their goals despite potentially conflicting interests. However, in open and dynamic systems, the problem of trust among agents has been identified as a fundamental issue [50, 33]. That is, when an agent decides if and how to interact with its peers, it must assess the trustworthiness of those peers to ensure that it does not enter into suboptimal interactions with unreliable agents. In particular, through the use of trust, an agent can identify reliable agents with which to interact in pursuit of its goals.

In this context, trust can be viewed as the *subjective probability with which an agent (the truster) believes another agent (the trustee) will perform an action that has an influence on the truster's goals.*[2] To form this probability, the truster may need to draw on many sources of evidence to achieve a significant degree of accuracy. For example, in a large system, individual consumers may each only have a small number of prior experiences of a given provider, and thus need to pool their experiences to obtain sufficient observations to accurately assess the provider's reliability. For this reason, third party opinions, often referred to as a trustee's *reputation*, are a particularly important source of information in large open systems, where an agent may routinely come into contact with individuals with which it has little or no previous experience [63]. Unfortunately, reputation can be associated with many sources of inaccuracy not present in a truster's direct experiences. For instance, a reputation source may have biases toward or against a trustee, and so give either an unfairly high or low opinion. However, even if a reputation source does reveal its knowledge truthfully, it may not be a true reflection of how a trustee is likely to behave toward the truster. This may be either due to the trustee behaving differently toward different agents, or due to differences in the way that a truster and its reputation sources perceive and represent a trustee's behaviour.

In this context, there is a need to develop trust assessment models, or mechanisms, that can be used to *aid decision making by autonomous agents operating in large-scale open service-oriented environments.* In particular, such mechanisms are required to estimate the future behaviour of an agent's peers, so that it may decide how to interact with those peers, to minimise its risk and maximise its expected gain. Moreover, to achieve this goal, a trust model should make use of multiple information sources, such as reputation or direct experience, so that predictions are not sensitive to the absence or failure of any one source.

However, given the wide range of scenarios in which trust is important, we believe there can be no single trust assessment mechanism to best suit every possible domain. For example, the issues that influence how trust should be assessed include (but are not limited to) the following two factors:

Behaviour Representation: To adequately predict or compare the performance of different trustees, a truster must have some means to represent their behaviour in terms of a common set of attributes. However, precisely which

[2]This definition is adapted from [21].

attributes are appropriate for a given domain depends on the preferences of the truster that are relevant for comparison and the type of service being assessed. For example, when comparing Internet search engines, a truster's preferences are typically expressed in terms of the time taken to process a query and the number of relevant hits returned by the result. On the other hand, preferences about a provider of fresh fruit may be best expressed in terms of price and the condition of purchased fruit on delivery. This means that there is no single representation of behaviour that is correct for all domains, and so any generic trust model must be able to accommodate any choice of attributes that are appropriate for a target application.

Computational Resources: For most estimation or prediction problems, there is a wide range of statistical models that can potentially be used to automate their solution. However, not all are equivalent, and there is usually a trade-off between the generality and accuracy of a model, and the time or computational resources required to implement it. Trust assessment is no exception. For example, an application that requires time critical decisions deployed on a sensor network with limited computational resources will place much harder constraints on model complexity than one with access to a compute cluster.

From this analysis, it is clear that the design of a particular trust assessment model should be a domain-specific exercise. Nevertheless, the need to rely on multiple (potentially unreliable) information sources does apply to most trust assessment problems. Therefore, we believe that the way forward in this area is to provide an overarching framework that addresses these general issues but, at the same time, can be easily configured to meet the specific requirements of a given target domain.

For this reason, we propose a generic Bayesian trust model, known as HABIT (Hierarchical And Bayesian Inferred Trust), which can be easily configured to predict trustee behaviour in a wide range of scenarios. To achieve this, HABIT comprises a two-level hierarchical model in which the opinions of different reputation sources are modelled in the bottom level, and the correlation between these opinions and actual trustee behaviour is modelled in the top level. By allowing different representations of trustee behaviour and agent opinions to be used interchangeably in both levels, HABIT has the flexibility required to meet a wide range of application specific requirements. However, no matter how the model is instantiated, the details of HABIT's hierarchical structure (discussed in Section 3.3) still impose sufficient constraints to enable accurate predictions in general, while at the same time, remaining robust in the presence of inaccurate or intentionally misleading information.

In common with a growing number of other statistically principled trust models, HABIT's foundation in probability theory enables it make predictions that are both logically consistent, and take full account of the degree of uncertainty due to incomplete information. In contrast, non-statistical trust models either cannot quantify uncertainty in a trustee's behaviour [39, 47] or do so by relying on ad hoc heuristics [62, 77] — the correctness of which is hard to verify (see Section 2.1). However, in addition to the benefits inherited by all probabilistic models, HABIT exhibits the following key features, which together form a significant contribution to the state-of-the-art:

1. HABIT is the first mechanism, for assessing trust based on reputation, that is statistically principled, enables computationally tractable inference, and yet is not tied to any particular representation of trustee behaviour. This contrasts with existing models of trust (discussed in Section 2), which either lack the strong theoretical foundation required to assess trust in a principled way, or can only perform feasible inference on a limited set of behaviour representations. To show this, in this paper we discuss how HABIT can be instantiated to meet the needs of different applications with different behaviour representations (see Section 4); provide a generic Monte Carlo Algorithm that can be used for tractable inference of trust in a wide variety of circumstances (see Section 5); and give an example instance in which trust can be inferred analytically, using a simple closed form equation (see Section 6).
2. With HABIT, a truster can assess a trustee based on reputation from sources that do not share a common *representation* of trustee behaviour. In contrast, most existing trust models assume that all information shared between a truster and its reputation sources is based on a commonly agreed set of attributes (e.g. service price and delivery time), which may not be appropriate if different agents have different preferences about how a trustee behaves. The only other mechanism that allows information sources to use different representations is BLADE (see Section 2). However, while HABIT can be used with any behaviour representation, BLADE can only deal with discrete representations of behaviour (see Section 4.4).

3. Regardless of whether agents share a common representation of behaviour, HABIT is the only probabilistic trust model (apart from BLADE) that can extract useful information from reputation with different *semantics*, or from sources that deliberately try to mislead the truster in a consistent way. For example, if a reputation source always reports that a trustee fulfils its contract when in fact it breaks it, and reports that it breaks its contract when in fact it keeps it, then HABIT can still use this information by assuming that the opposite of the report is always (or at least usually) true. Similarly if, for example, a rater on Amazon[3] were to consistently give 3 stars to a supplier to which the truster would award 5, HABIT can learn to adjust the report accordingly. This is because HABIT does not assume any particular interpretation for reputation, but instead learns from any statistical dependence it discovers between its own direct experience and reputation, be that an inverse correlation or otherwise. In contrast, models that use discount factors to reduce the weight of evidence provided by reputation [32, 68, 74] typically assume that the truster and its reputation sources always say and mean the same thing by a given behaviour representation, and penalise reputation sources that appear to report otherwise. As such, models that use discount factors would discard reputation in cases such as those above, even though they may still provide useful information.
4. HABIT enables a truster to assess the behaviour of agents for which there is little or no previous experience, including reputation. 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 (such as organisation membership or the types of service they offer). This ability is particularly important for two reasons. First, upon initial entry to a system, a trustee will have no history of interactions with other agents. Thus, for any trustee, there will *always* be at least some point when it cannot be judged on its previous behaviour. As such, any complete trust and reputation system must have some means to assess a trustee that does not rely on prior experience. Second, even when there is a history of interactions with a trustee, any open system is still susceptible to the *whitewasher* problem [77], in which unreliable agents adopt a new identity, thereby absolving themselves from blame for any previous wrong doing. Early solutions to this problem typically suggest treating unknown agents as completely unreliable, but 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 (see Section 4.3).
5. As demonstrated in our experiments, HABIT outperforms BLADE when applied to discrete representations of behaviour which, *unlike* HABIT, is the only type of representation to which BLADE can be applied, and so is the only type with which these models can be compared. In particular, we found that HABIT was up to twice as accurate as BLADE (in terms of the mean absolute error) at predicting a truster's expected utility for interacting with a trustee (see Section 7). Here, we chose BLADE as a benchmark because, by being statistically principled and sharing some (but not all) of the beneficial properties described above, it can be viewed as the state-of-the-art among reputation-based trust models. In addition, BLADE has been previously shown to outperform a number of other probabilistic trust models [57].

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 discusses related work on trust assessment mechanisms; Sections 3 and 4 introduce the overarching HABIT model, and discuss how this can be applied to different applications; 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 general model; Section 6 specifies two ways in which HABIT can be applied to discrete representations of behaviour; Section 7 presents an empirical analysis, in which these instances of HABIT demonstrate better performance than BLADE; further to this, Section 8 presents additional experiments using data from a *real* webserver domain, showing that HABIT can make accurate predictions about *continuous* (as well as discrete) behaviour representations, and its performance is *robust* against the complex statistical properties of a real system; and, finally, Section 9 summarises the main properties of the model and discusses future work.

2. Related Work

The issue of trust in multi-agent systems is one that is widely recognised, and has been addressed by a number of different models and mechanisms. Generally, however, existing approaches can be classified as one of two types: (1)

[3] `http://www.amazon.com`

those that try to enforce good behaviour by incentives or penalties [6, 51, 34, 35]; and (2) those that try to predict how a given trustee will behave, so that a truster can choose to interact only with agents that it believes will behave beneficially toward it. While these two types of approach may be complementary, our focus in this paper is on the latter type, which tend to differ in how they represent an agent's behaviour and the information sources used to assess it. For example, with regard to information sources, we can use knowledge of the social rules and norms that apply in an environment [49, 52], knowledge of an agent's incentives [25, 6], or any relationships that are known to exist between agents [1]. However, although such evidence can play a significant role in some circumstances, it cannot be expected to be present in every domain. In contrast, perhaps the most widely available performance indicator across a number of domains is a trustee's past behaviour. Observations of past agent behaviour are thus widely recognised as an important and basic predictor in trust assessment, and so are adopted by the majority of trust models.

For this reason, this section reviews existing trust models, focusing on those that assess trust based on prior observations of agent behaviour. More specifically: Section 2.1 discusses the different types of approach that have been previously used to perform inference about trust; Section 2.2 discusses how trustee behaviour is represented in existing trust models, and how they cope with potentially unreliable reputation; and, finally, Section 2.3 reviews the different strategies that have been proposed to enable a truster to make decisions about trustees for which the truster has little or no direct or third party experience.

2.1. Approaches to Inference

With regard to inference, early mechanisms tend to adopt a heuristic approach, with improvised functions to account for different aspects of agent behaviour. For example, [62] introduces the REGRET system, which allows agents to evaluate each other's performance based on multiple domain dependent attributes (for example, service quality or price). Specifically, each time an agent interacts with a provider, it assigns a real value in the range $[-1, 1]$ to each attribute on which the provider is assessed, where negative values are interpreted as poor performance, and positive values as good performance. For each attribute, an agent's overall assessment of a provider is then calculated as a weighted average of individual interaction evaluations, which may be made directly by the truster, or reported from third party interactions with the provider. In each case, the weight assigned to each evaluation may be chosen as a function of a number of attributes. For instance, weights may be chosen as a function of time, with greater weight assigned to more recent observations (to allow for changing behaviour), or based on the source of the evaluation (direct or third party).

However, such techniques tend to have few theoretical properties to characterise how they should perform under different conditions, or to show how they compare to any theoretical optimal performance. For instance, in the case of REGRET, evaluation weights are assigned purely using intuition; no consideration is given to if or how the precise functions used could be modified to improve provider assessment. Similar arguments can also be made against the use of fuzzy logic [71, 70, 44], since it is difficult to define fuzzy sets in a way that is objective, or provably better than any other. More theoretically principled approaches include those based on formal logic, for example [47] and [39], but, while such work has its place, such as constructing formal arguments for negotiation between agents [48] and enacting security polices for access control [9, 38], the purely deductive style of reasoning performed cannot fully account for uncertainty about trustee behaviour.

In contrast, estimation problems involving uncertainty have been studied intensively within statistics, resulting in a large body of well established results and solutions (see [14] for an overview). These tend to have well defined goals and properties, with clear notions of what makes a solution optimal, and under what conditions optimality can be reached. Moreover, as argued by Jaynes [27], probability theory uses the only complete set of rules for reasoning about degrees of uncertainty that guarantees logically consistent inference. In other words, probability theory guarantees that the same conclusions will be reached from the same data, no matter how its rules are applied. This does not mean that probability theory is the only possible formalism for reasoning about uncertainty, but it does mean that any formalism that disagrees with probability theory cannot guarantee consistent inference.

Of course, even if inference is consistent, it is not necessarily correct. This is especially true in most real-world inference problems involving uncertainty, because simplifying assumptions always have to be made to enable tractable solutions. As such, any inference is only as good as the assumptions on which it is based, and so it is not impossible for a non-statistical model based on good assumptions to outperform a statistical one based on bad assumptions. Nevertheless, given a certain set of assumptions, it always makes sense to perform inference using a tried and tested

formalism that is guaranteed to produce logically consistent and often provably optimal results. Therefore, for problems that require reasoning under uncertainty, the obvious choice is probability theory and, consequently, a growing number of trust assessment mechanisms have been developed based on it. Moreover, the success of this approach has been demonstrated in the international ART testbed competition[4] [19, 46] in which, for two out of the three years the competition ran, the winning strategy was based on probability theory [67].

One early example of a probability based trust assessment mechanism is the Beta Reputation System [32], which represents a provider's performance in providing a particular service as a binary random variable: either the service meets the consumer's requirements, or it does not. The provider's general performance is then modelled by the probability that it will satisfy a consumer's requirements during a particular interaction. For example, if a provider provides a service ten times with a satisfaction probability of 0.4 then, on average, we would expect the consumer to be satisfied on four of those occasions, and unsatisfied on six. With this approach, each provider's satisfaction probability is assumed to be an intrinsic characteristic of its behaviour, which must be estimated based on available evidence. To do this, consumers cooperate by sharing reports of their frequency of satisfactory and unsatisfactory interactions with each provider, and based on this, apply Bayesian analysis to estimate the satisfaction probability for each provider. These estimated probabilities are then used to facilitate choices between providers, such that providers with high satisfaction probabilities are generally preferred to those with low satisfaction probabilities.

There are two main advantages to this approach. First, by pooling their experiences, each consumer essentially has multiple sources of information on which to assess a provider, so they can assess providers with which they have little or no direct experience. Second, by receiving frequencies of satisfactory and unsatisfactory interactions, a consumer can take proper account of the weight of evidence behind each reputation source's beliefs. This means that, when forming its own beliefs, a consumer will place more weight on an opinion based on a large number of interactions than one based on only a small number of interactions. In contrast, if each reputation source only reported an estimate of its satisfaction probability (for example 0.4 or 0.6) then a truster would not be able to aggregate reports optimally because the information about their certainty would be lost. By using reported frequencies rather than estimates, the Beta Reputation System avoids this problem, and so has been used as the basis for a number of other probabilistic trust models that share its binary representation of behaviour [74, 68, 75, 60].

Unfortunately, the system exhibits two main limitations. First, by adopting a binary representation, the system cannot make predictions about more fine-grained attributes of trustee behaviour, which may have an important impact on a truster's decisions. For example, the utility of receiving a service may decrease in proportion to its delivery time or quality of service, neither of which can be represented by a binary variable. Second, there is no mechanism for dealing with malicious, or otherwise inaccurate, reputation. That is, one or more agents can decrease the accuracy of estimated satisfaction probabilities by introducing a bias into their reported experiences. This may be done intentionally, to manipulate a provider's reputation, or unintentionally, due to differences in the way different consumers assess performance. Although such effects may be reduced given sufficient reliable reports, the system is particularly vulnerable to malicious attacks if agents are allowed to report unlimited numbers of experiences unchecked. For example, if a provider wished to manipulate its reputation to increase its share of the market, it could (either anonymously or under a false identity) increase its estimated satisfaction probability by reporting an arbitrarily large number of satisfied transactions. For instance, if the total number of true experiences reported to the centre is 100, then by reporting 1,000,000 successful transactions, the provider could set its estimated satisfaction probability to greater than 0.9999.

2.2. Reputation Filtering and Behaviour Representation

To overcome the limitations of early trust assessment mechanisms, more recent work has developed statistical models that can assess trust based on a wider range of behaviour representations, and can filter out the effects of unreliable reputation. With respect to the former, this is achieved by replacing the binomial probability distribution used to model binary outcomes with one of a more expressive class of distributions capable of modelling richer behaviour representations. So far, this has typically been achieved using either Gaussian distributions, to model behaviour using continuous real-valued random variables [15, 66]; or multinomial distributions, to model behaviour using discrete random variables [20, 40, 31, 56, 55, 57].

[4]The ART (Agent Reputation and Trust) testbed provides a common platform on which to compare trust assessment mechanisms. From 2006 to 2008, it was used to hold international competitions annually at the International Joint Conference on Autonomous Agents and Multiagent Systems.

Together, these two types of distribution can capture a number of important aspects of trustee behaviour, such as the mean and variability in an agent's performance, or the relationship between different measures of performance. For example, by using a multivariate Gaussian distribution, continuous attributes, such as price and delivery time, may be modelled together allowing correlations between them to be discovered. For instance, we may find that a courier that offers a low mean price tends to have longer or more variable delivery times. Similar characteristics, based on one or more attributes, can also be modelled using multinomial distributions. However, with these, attributes that are continuous in nature, such as delivery time, must first be converted into a finite set of possible values. As such, when using multinomial distributions, predictions about continuous variables can only be made w.r.t. value ranges, such as 1–2 days for delivery, versus 3 or more days.

While these abilities show that Gaussian and multinomial distributions are much more expressive than the binomial distributions used by other statistical trust models, they are not the only possibilities, and others may be more appropriate in different cases. For example, delivery times can often be more accurately modelled using Poisson distributions [14], while some combinations of behaviour attributes may be best modelled using mixtures of discrete and continuous distributions. Unfortunately, neither of these possibilities, nor many others, are currently supported by current reputation models, which are all tied to one particular class of distribution.

The reason for this limitation is the difficulty in designing models of reputation that are both computationally efficient and effective for a range of situations and behaviour representations. This is particularly true in cases where reputation may be unreliable, due to the additional complexity in discovering potentially misleading opinions. Moreover, since unreliable reputation is to be expected in most real world applications, alleviating its effects is the key focus of most reputation models. In particular, three types of approach are prevalent in the literature, which we now describe, and refer to as the *majority rules*, *all-or-nothing* and *reputation function* approaches.

In the majority rules approach, the key intuition is that, when gathering opinions from multiple sources, the majority of sources are likely to be reliable. As such, identifying unreliable opinions boils down to identifying and discarding outlying opinions, which differ significantly from the norm. While this type of approach has so far only been applied to reputation using binary [76] and multinomial behaviour representations [40], there are a number of general techniques for outlier detection in statistics, which could be applied to more general cases [8]. However, even if this approach can be generalised, there are two important cases in which its adoption may have a detrimental effect. First, when a trustee deviates from its usual behaviour in a small number of cases, sources that happen to interact with the trustee in these cases will legitimately have opinions that differ from the norm. To give a balanced view of the trustee, these opinions should thus be taken into account, rather than being discarded as outliers. Second, when a trustee is a newcomer to a system with which no agent has significant experience, then *any* reported experience from any source must be fictitious. Thus, unless the true number of service transactions can be reliably monitored, this approach is still open to manipulation.

Due to these limitations, the majority of reputation filtering mechanisms attempt to assess the reliability of each reputation source individually, based on the perceived accuracy of its previous opinions. This is the basic intuition behind both the all-or-nothing and reputation function approaches, which work by comparing each source's past predictions with subsequent trustee behaviour observed directly by the truster: reputation sources that provide opinions that are generally uncorrelated with observed trustee behaviour are assumed to be unreliable, and so have less weight placed in their opinions.[5] However, the details of how this is achieved differ between the two approaches.

Of these two, the all-or-nothing approach is used to handle reputation in the majority of statistical models, including those described in [68, 72, 74]. Here, it is assumed that reports from a given reputation source are either completely unreliable or are equivalent to a truster's own direct experiences. If the former is *known* to be true, then a reputation source's reports are completely discarded, while if the latter is known, its opinions are given equal weight to the truster's own observations.[6] Typically however, neither of these are regarded as absolutely certain. Instead, a truster estimates the probability that a reputation source is reliable, which we refer to as the *probability of accuracy*, and uses this to weight the reputation source's observations. In practice, this generally means that reputation is neither completely discarded, nor given equal weight compared to a truster's own experiences, but is instead weighted

[5] As a result of this, perceived accuracy can only be assessed with respect to an individual truster's own direct observations. As such, the reliability of each reputation source may be perceived as different from the perspective of different trusters.

[6] In this respect, the majority rules approach could equally be described as an all-or-nothing approach but, for clarity, we only use the term here to refer to filtering mechanisms based on a source's past opinions.

according to its perceived level of accuracy. However, there are two main problems with this approach.

First, it is difficult to pin down exactly what it means for reputation to be accurate, and to calculate its probability. This is due to the large number of factors that usually influence opinion accuracy. For example, an opinion based on a small number interactions with a trustee cannot reasonably be expected to be as accurate as an opinion based on a large number of interactions, so it seems reasonable that a reputation source should be judged more on its confident opinions than on cases where it is not so confident. As result, many trust models of this kind, such as TRAVOS [68] and those proposed by Wang et al. [74], resort to heuristics to both assess a reputation source's accuracy, and to weight its opinions accordingly. Thus, even if these models are otherwise probabilistic, they lack the theoretical performance guarantees that are typically ensured by a more rigorous application of probability theory (see Section 2.1). In fact, the only model of this type to be derived using probability theory, without resorting to heuristics, is described by Vogiatzis et al. [72]. However, this model only remains computationally tractable by adopting a binary behaviour representation, and only allowing reputation sources to provide an estimate of behaviour, without any measure of confidence. It thus remains to be seen if the reputation equivalence approach can be extended to other behaviour representations, in a principled way, without becoming computationally intractable.

The second problem with this approach is that it does not deal well with opinions that are in any way subjective. This is because any deviation between a reputation source's opinion and a truster's own experience is normally viewed as evidence that the reputation source is unreliable. For example, if trustee behaviour is awarded marks out of five, but a truster consistently awards one less mark than a reputation source, then this could eventually lead to the reputation source being ignored completely, even though it still provides useful information. This problem is partially solved by building in some level of tolerance, so that small differences in opinion are accepted [68]. Alternatively, subjectivity may be removed completely by forcing reputation sources to describe their experiences of a trustee in purely objective terms. For example, in the POYRAZ model [11], an ontology is used to describe each transaction between a reputation source and a trustee in detail, so that a truster can then decide for itself how it would rate each transaction. Unfortunately, this may require a reputation source to provide more details about its experiences than it may be willing to volunteer, and in any case, eliminating all sources of subjectivity may not always be feasible.

Fortunately, this limitation is addressed by the reputation-function approach, in which reputation is modelled as an unknown stochastic function of trustee behaviour. By learning this function over time, a truster can extract useful information from reputation sources that adopt different behaviour representations and semantics, while at the same time, protect against unreliable reputation that bears little or no correlation to a trustee's actual behaviour. For example, in the case above, if a source consistently adds one to a truster's own score, then the truster could compensate for this by modelling it as a bias term in the reputation function, and then subtract this term from each opinion in order to extract useful information. In fact, useful information can even be extracted in more extreme cases, such as when a reputation source uses a completely different scoring system, or consistently reports that a trustee is bad when it is good and vice versa. This only requirement is that some statistical correlation can be observed between a reputation source's reports and truster's own experience, so that some useful information can be extracted. On the other hand, this approach is also robust against manipulation, since reports that show little or no correlation with a truster's own experience will be taken as evidence that the reputation source is unreliable.

So far, the only trust model to fully embrace this approach is BLADE [57]. In addition, the all-or-nothing approach can be viewed as a special case, in which a reputation function is either an identity function (i.e. reputation is equivalent to direct observations) or consists only of independent random noise, which provides no information about a trustee's actual behaviour. However, as we have seen, adopting this more simple approach loses most of the flexibility that the more general case has to offer. There is also a middle ground, occupied by TRAVOS-C [66], which models any deviation from the identity function as added noise. As such, this also penalises any difference between reputation and direct experience, but relatively small deviations can be modelled with small noise terms, permitting some information to be extracted.

Unfortunately, as with all other existing statistical trust models, both BLADE and TRAVOS-C are tied to their respective representations of behaviour. Moreover, since they adopt a more general approach, the computational complexity of extending their use to other behaviour representations is likely to be worse than for all-or-nothing models. The reasons for this are discussed in more detail in Section 5, in which we also show how, by adopting an entirely new approach, HABIT is able to retain the advantages of these earlier models, while remaining computationally tractable across a wide range of domains.

2.3. Group Behaviour and Context Modelling

In all of the statistical trust models discussed so far, the sole source of information about a provider's behaviour is direct or third party experience of its past behaviour. However, this does not deal adequately with cases in which no agent has significant experience of a provider, for example when a provider enters the system for the first time, or obtains a new identity by whitewashing. To deal with such cases, Zacharia et al. [77] propose that new entries to a market should always be assigned the lowest possible rating, removing any incentive to adopt a new identity. Although this approach effectively removes the whitewasher problem completely, in practice a significant proportion of new identities are likely to belong to legitimate newcomers. Thus, to avoid penalising such providers unfairly, a more pragmatic solution is to trust each newcomer according to the average performance of other similar agents already in the system. For example, agents may be assigned to groups based on attributes such as professional accreditation, or the types of services they offer. Providers for which there is little or no specific experience can then be assessed based on the behaviour of other members of their group.

This type of solution has been proposed by Sun et al. [65]; here trustees are assigned to different groups, and assigned a different initial trust value depending on the group they belong to. However, for this solution to work, the behaviour of group members generally needs to be correlated, so that the group's average performance gives a reasonable prediction for any given member. To this end, a number of techniques have been proposed to measure the similarity between agents, so that groups of agents with similar behaviour can be identified based on their shared attributes. For example, Liu et al. [42, 41] propose a technique whereby agent similarity is measured using a set of fuzzy logic rules, while Wang et al. [73] propose an alternative method, whereby agents that belong to social or business networks are assessed according to the behaviour of other agents with which they share links. However, both of these techniques have their limitations: in the former case, it is not clear how appropriate fuzzy logic rules can be chosen to reflect the similarity of agent behaviour in any given domain, while in the latter case, it is not always appropriate to assume that agents that share network connections will behave in a similar way.

To overcome these limitations, features that identify similar behaviour can be discovered automatically, rather than being specified in advance. For example, Burnett et al. [5] propose a method in which predicting an agent's trust value is treated as a regression problem using sets of observable binary feature variables as input. For instance, each feature may indicate whether the trustee holds a particular qualification, or belongs to a specific organisation.[7] Taken together, the set of all such features describing a trustee is then used to predict its trust value based on the observed trust value of others with shared features. The main advantage of this approach is that the relationship between observed features and trust does not need to be specified in advance, but can be learnt from observed data. Moreover, it can also be used in combination with any trust model (including many probabilistic models) that represents trust as a real-valued scalar.

Although Burnett et al.'s model is useful for providing an initial trust value for agents whose behaviour has not yet been observed, it does have two disadvantages. First, it does not provide a general mechanism for combining a trustee's initial value (based on other's behaviour) with evidence based on its own behaviour, such as a truster's direct experience with the trustee, or third party opinions.[8] Although a specific mechanism is suggested for combining evidence using the Beta Reputation System, it is not clear how this may be extended to other types of model, and in particular those that adopt a non-binary behaviour representation. This is important because, when no one source of evidence can provide a reliable prediction of behaviour, combining all the evidence in an appropriate way can often lead to better predictions. Second, when calculating a trustee's initial trust value, no account is made for the uncertainty in this value. This is because the model is trained based on a truster's opinions about any agent it has observed at least once, without regard to the amount of evidence or number of observations on which each opinion is based. When the amount of evidence is low, this may lead to predictions that are unduly biased toward certain types of behaviour, when in fact all types of behaviour are equally likely.

These problems are partially solved by Bayesian trust models, which provide not only an estimate of trustee behaviour, but also a measure of uncertainty based on the amount of evidence available. Using this additional information, evidence from different sources can be aggregated in a principled way by applying the rules of probability

[7] Other features that may be used to assess trust in this way are discussed in [4].

[8] Burnett et al. do introduce the notion of *stereotype* reputation, which allows trusters to share information for calculating initial trust values based on others' behaviour. Although useful, this is based on the simplifying assumption that reputation sources are truthful, and is not the same as combining evidence with reputation based on a trustee's *own* behaviour, which is achieved by HABIT.

of theory. In particular, this ability is exhibited by TRAVOS-C, which learns the distribution of the behaviours of a group of agents to form prior beliefs about the behaviour of an individual, and by the IHRTM model, proposed by Rettinger et al. [58, 59]. In common with Burnett et al.'s model, the latter can also account for how external factors and contextual information should influence trust in an agent. For example, such factors may include current market conditions, qualifications or awards from professional bodies, the geographic origin of a product or the asking price. However, this approach does not include a sophisticated model of reputation, and so it can only incorporate rudimentary reputation information, such as the five star seller ratings supplied by e-Bay,[9] merely to provide context.

Thus, with the exception of TRAVOS-C, there is no existing statistical trust model that can, *in a principled way*, combine evidence from direct experience, reputation and group behaviour. Of those that are statistically principled and can deal with potentially inaccurate reputation, all are tied to a specific (often coarse) representation of behaviour, and cannot be adapted to incorporate the properties of other statistical models in order to meet the specific requirements of a particular application.

3. The Generic HABIT Model

To overcome the limitations of existing trust models, we now introduce the HABIT model and discuss how it can be used to make rational decisions regarding the trust that an agent should place in its peers. To achieve this, the current section is divided into three parts: Section 3.1 introduces the basic notation used to define the HABIT model; Section 3.2 describes the role of HABIT in trust assessment by discussing how, in general, rational decisions involving trust can be made using decision theory; and finally, Section 3.3, defines the HABIT model at a generic level, independent of application-level considerations.

3.1. Basic Notation

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, 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, C, as O^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 $O_{2 \to 1} \in O^{apples}$ and $O_{1 \to 2} \in O^{money}$, 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 $O^{0:t'}_{tr \to te}$.

[9] `http://www.ebay.com/`

3.2. Making Decisions about Trust

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, our aim is to make a truster 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. In essence, we can achieve this in two parts.

First, we identify the possible outcomes of each course of action a truster may take, and identify how much the truster prefers each possible outcome. For example, if a truster chooses to purchase a movie service from a multimedia provider, possible outcomes include receiving a high quality video stream at low cost, receiving a low quality video at high cost, or paying for a contract that the provider decides not to honour at all. Clearly, in this case, a truster is more likely to prefer the first possibility over the latter two.

Second, for each action the truster may choose, we evaluate the likelihood of each possible outcome, based on all available evidence (including the truster's direct experiences and reputation). For example, it is reasonable to assume that a truster is more likely to receive a good quality service if it chooses to rely on a service provider that consistently provided good services in the past, rather than relying on one that has provided consistently poor quality of service. Based on this, a truster should then choose the course of action that is most likely to result in an outcome that it prefers, which may involve purchasing a service from a reliable provider, or opting out altogether.

To perform these two steps in practice, we turn to decision theory [2], because this provides the most principled foundation for choosing between actions with uncertain outcomes (such as deciding between competing trustees or service providers). To apply this, we first quantify a truster's preferences by defining a *utility* function that depends on the outcome of the truster's choice of action. This means that the preferences of any *tr* with regard to interacting with *te* are encoded by a utility function $U : \mathcal{O}^{\mathcal{C}} \to \mathbb{R}$, such that if *tr* prefers an outcome $x \in \mathcal{O}^{\mathcal{C}}$ over an outcome $y \in \mathcal{O}^{\mathcal{C}}$ then $U(x) > U(y)$, and if *tr* has equal preference for x and y, then $U(x) = U(y)$.

Second, we assess the value of *tr* interacting with *te* by calculating the *expected utility* (EU) of each choice of action, which depends on the *probability distribution* of the possible outcomes of that action. More specifically, if $p(O_{tr \to te})$ is a *probability measure* for possible outcomes of interactions between *tr* and *te*, then the expected utility for *tr* interacting with *te* is given by the Lebesgue integral[10] in Equation 1. Based on this, an agent can make the best possible decision in a given situation by choosing an action that maximises its expected utility, for example by choosing the best trustee.

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

The precise definitions of $U(O_{tr \to te})$ and $p(O_{tr \to te})$ depend on the particular application at hand, and so there is no specific solution that is appropriate to every domain. For instance, preferences about car insurance and the likely behaviour of a car insurance broker cannot be represented in exactly the same terms as the behaviour and preferences of Internet search engines. However, some aspects of $p(O_{tr \to te})$ can be discussed in more general terms, including the types of evidence used to assess *te* and how they are fused to form $p(O_{tr \to te})$. For example, as discussed in Section 2, we can use knowledge of the social rules and norms that apply in an environment or any relationships that are known to exist between agents. However, as such information may not be widely available across a variety of application domains, we choose to concentrate on observations of prior agent behaviour, and in particular the following three sources of such information:

1. the direct experience of the truster, gained through previous interactions with the trustee;
2. the reputation of the trustee, comprising all third party experiences reported to the truster; and
3. observations[11] of the behaviour of groups of agents that share characteristics with the trustee, such as organisational membership, the types of services they offer, or the length of time they have been present in a system.

In the next subsection, we introduce the generic HABIT model, which shows how trust can be modelled in general, using these three sources of information.

[10] When appropriate, we may use the language and notation of measure theory and Lebesgue integration to remain agnostic about the domain of integration. For discrete random variables, the Lebesgue integral may be replaced by a summation with a *probability function* taking the place of the measure. Likewise, for continuous random variables, the Lebesgue integral may be replaced by the more familiar Riemann integral, with the measure provided by a *probability density function* (p.d.f.) [61].

[11] These may be observed directly or reported by third party reputation sources.

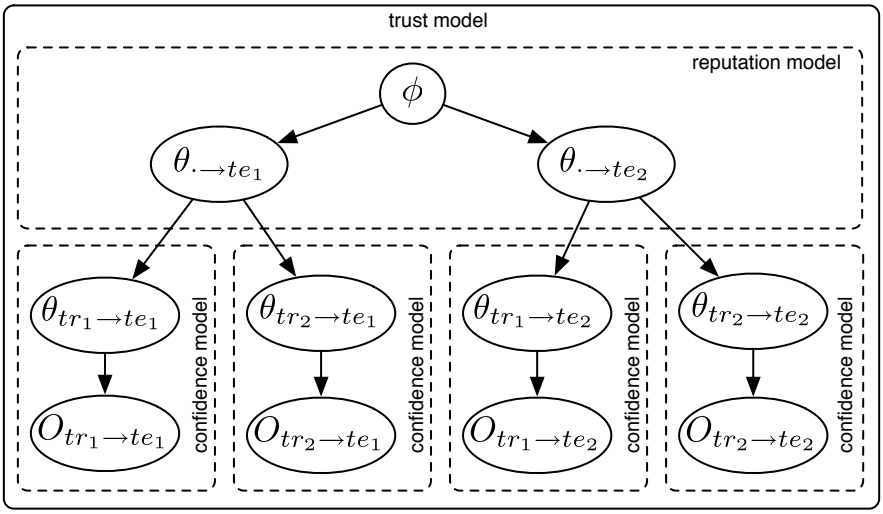

Figure 1: The Generic HABIT Model, illustrated as a Bayesian Network.

3.3. Model Architecture

When designing a generic trust model, such as HABIT, a trade-off must be sought between its generality, and the ease with which it may best be applied. That is, while an over-specified model may only be applicable to a few domains, an under-specified model may provide little insight about how to solve a given problem. Therefore, to strike a balance between these two extremes, trust must be modelled with sufficient detail to enable practical inference, while as far as possible remaining agnostic to those parts of the problem that are domain dependent.

With this in mind, 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). Both component types are generic, and so under reasonable restrictions, can be instantiated in different ways to meet different requirements.

Although this hierarchical approach has *not* been applied to probabilistic trust models before, we believe it strikes a good balance between generality and specificity for two reasons. First, in contrast to existing probabilistic trust models (see Section 2), HABIT places no major constraints on how an *individual* trustee's behaviour is modelled or represented. As we shall see, this domain dependent problem may be easily solved by incorporating standard statistical models. Second, without sacrificing computational efficiency, HABIT can model more complex relationships *between* trustee behaviour and reputation than is currently possible using any other model. Thus, HABIT's power and sophistication is not limited by its generality. Evidence for this is provided throughout the following sections. However, before we can discuss these benefits further, we must first introduce HABIT's structure in more detail.

To this end, the key components of HABIT are illustrated by the Bayesian network in Figure 1. In particular, for each truster, *tr*, and trustee, *te*, the role of the confidence model is to represent the probability distribution, $p(O_{tr\rightarrow te}|\theta_{tr\rightarrow te})$, of all observations $O_{tr\rightarrow te}$, where $\theta_{tr\rightarrow te}$ is a parameter vector[12] 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.

[12] 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.

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 confidence model. For instance, while one truster may represent behaviour only in terms of response time (which is reasonable if its utility function only depends 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 be modelled, possibly using a multivariate Gaussian distribution, or some more appropriate combination of conditional distributions.[13]

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 [22]. The difficultly 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 3.3). 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.[14] 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), and so satisfies the objective of modelling trust based on multiple information sources, as discussed in Section 1. 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*. It can then apply its knowledge of ϕ to make more informed inferences about *te* based on all available evidence.

[13] It is also possible that a specific truster may use different types of distribution to represent the behaviour of different trustees, or to model the behaviour of the same trustee in different contexts. However, as shall become clear later, the types of inference made possible through the reputation model can only provide a significant benefit if trusters observe the behaviour of multiple trustees using the same type of distribution (see Section 4.4).

[14] Strictly speaking, HABIT does not rely on the specification of a fixed parameter vector, ϕ, provided there is some way to model the joint distribution of the parameters, $\theta_{\cdot \to te}$ (for example, see Section 6.2). Nevertheless, it is notationally convenient to use ϕ to refer a specific characterisation of the joint distribution of $\theta_{\cdot \to te}$.

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

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	$\theta \cup \{\phi\}$
$\Phi_{\cdot\to j}$	$\theta_{\cdot\to j} \cup \{\phi\}$

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, different modelling assumptions and probability distributions are required to suit each target scenario. Fortunately, in its generic form, HABIT provides a clear and well-defined framework that may be easily instantiated for a given domain. Therefore, in this section, we outline the steps required to apply HABIT to a specific problem, and discuss the issues that should be considered when fulfilling these steps and how they can be addressed. More specifically, the rest of this section is structured as follows: Section 4.1 outlines the steps required to fully instantiate the generic version of the model introduced in the previous section; Section 4.2 discusses the trade-off between model sophistication (for making more informed decisions) and time complexity; Section 4.3 describes how HABIT can assess a previously unencountered trustee based on the behaviour of other agents, and discusses how this can be used to address the whitewasher problem; and finally, Section 4.4 describes how HABIT can incorporate information about the context of an interaction with a trustee, and use reputation from sources that adopt different representations of trustee behaviour.

4.1. Instantiating the Generic HABIT Model

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 $O^{0:t'}_{tr\to te}$, and all observations $O^{0:t'}_{i\to j}$ reported by an arbitrary observer, i, about an arbitrary trustee, j. More precisely, if $\mathcal{R}$ is the set of all agent pairs (i, j), including (tr, te), such that $O^{0:t'}_{i\to j}$ is observed or reported to tr, and $\mathcal{E} = \bigcup_{(i,j)\in\mathcal{R}} O^{0:t'}_{i\to j}$ is the set of all such evidence, then the goal is to estimate:

$$EU|\mathcal{E} = \int_{O^C} U(O_{tr\to te})\, dp(O_{tr\to te}|\mathcal{E}) \tag{2}$$

Here, the *predictive* distribution, $p(O_{tr\to te}|\mathcal{E})$, is derived by *marginalising out*[15] the unknown model parameters from the joint distribution, $p(O_{tr\to te}, \Phi|\mathcal{E})$, which is defined by the Bayesian network illustrated in Figure 1. As we discuss in later sections, this marginalisation is the key to HABIT's beneficial (and intuitive) properties, because it allows us to account for the different types of uncertainty that exist at each level of the model. For example, it seems reasonable to assume that a confident opinion, based on many reported interactions, will be more accurate than an uncertain opinion based on few interactions. As a consequence, when predicting a trustee's behaviour or assessing the reliability of a reputation source, it would seem sensible to pay more attention to confident opinions compared to uncertain opinions. This intuition is naturally captured by the marginalisation process, since uncertain opinions correspond to relatively flat probability distributions, which have limited influence on HABIT's parameter distributions, after marginalisation.

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 amount of time

[15] In Bayesian analysis, *marginalising out* a set of variables, X, from a joint probability distribution, $p(X, Y)$, refers to calculation of the probability of the remaining variables, $p(Y)$, by applying the sum rule of probability theory [27].

and 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 measures[16] (p.m.s) and parameters:

1. for each confidence model (i.e. each truster-trustee pair), the conditional distribution of interaction outcomes, $O_{i\to j}$, given a chosen parameter vector, $\theta_{i\to j}$, with p.m. $p(O_{i\to j}|\theta_{i\to j})$;
2. the prior distribution (that is, without knowledge of any observed outcomes) of each parameter vector $\theta_{i\to j}$, with p.m. $p(\theta_{i\to j})$;
3. the conditional distribution of all joint parameter vectors, $\theta_{\cdot\to j}$, given the hyperparameter vector ϕ, with p.m. $p(\theta_{\cdot\to j}|\phi)$; and
4. the prior distribution of the hyperparameter vector, ϕ, with p.m. $p(\phi)$.

Although having this number of unspecified components may seem 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. Nevertheless, 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. The result is then a full instance of HABIT tailored to a specific domain, which automatically inherits the beneficial properties of the generic version. In particular, this includes the ability to predict behaviour based on group behaviour as well as reputation; the ability to learn from reputation sources that use different representations or semantics; and as we shall see, the potential for computationally efficient inference. In contrast, models that do not adopt HABIT's hierarchical structure either lack its statistically principled grounding, are unable to extract useful information from reputation sources with different semantics, or are difficult to generalise to different types of behaviour representation. With this in mind, the following subsections identify the aspects that should be considered when instantiating the distributions above, along with the range of properties that may be achieved by doing so.

4.2. Model Sophistication and Time Complexity

As a general model, the purpose of HABIT is to provide a framework for reasoning about trust in a principled way across a wide range of problems. As such, we do not advocate the use of a single set of parameter models to model trust in every domain, but instead show how they can be applied in general. To a large extent, this means that the level of sophistication and the time complexity of inferences based on HABIT depend on the particular set of parameter models used to instantiate it.

For this reason, we refer to the general literature on Bayesian analysis for possible instances of the distributions outlined above. Specifically, any of these distributions could be instantiated by assuming they belong to any one of a number of parameterised families of distributions. As illustrated in Section 3, this includes Gaussian distributions, in which case the parameter vector may be defined as $\theta_{tr\to te} = \langle \mu, \sigma^2 \rangle$, where μ is the distribution's mean and σ^2 is its variance. Alternatively, other parameter models may be used, including: gamma, Poisson and multinomial distributions; multivariate generalisations of such distributions; or mixture models [22, 37, 45, 17].

For the most part, the choice of parameter models to use is down to how well the chosen distributions match the properties of the domain and to the computational resources and time available for a truster to make its decisions. Typically, there is a trade-off here because models that can approximate a wide range of phenomena, such as infinite mixture models [53, 3], are more computationally complex to perform inference with than their more basic counterparts.

However, there is one hard constraint imposed by HABIT on the choice of distributions: with respect to the outcome distributions, $p(O_{i\to j}|\theta_{i\to j})$, there must be some fixed length instance of $\theta_{i\to j}$ that can fully determine the shape of the distribution. This is to allow practical inference in the reputation model, which would require much more complex modelling if the length of the joint parameter vector, $\Phi_{\cdot\to j}$, was allowed to grow dynamically. This excludes $p(O_{i\to j}|\theta_{i\to j})$ from being instantiated using so called non-parametric models, such as Gaussian and Dirichlet

[16] Recall that, for discrete random variables, a probability measure may be provided by a probability function (p.f.), while for continuous random variables, it may be provided by a probability density function (p.d.f.) [61, 14].

processes [54, 24], that cannot be determined by a fixed number of parameters.[17] Nevertheless, this condition is not particularly restrictive, since the class of parametric distributions is large, and non-parametric models can still be used to instantiate the other required distributions listed above.

With respect to the other distributions, any parameter models that are convenient and in line with any known properties of the application domain may be chosen. However, in the interest of efficiency, we believe it is useful to follow the standard practice of choosing *conjugate* prior distributions. In general, these are classes of prior distributions that are assigned to model parameters, such that the posterior parameter distribution, given the evidence, also belongs to the same class of conjugate distributions [14]. The advantage of doing this is that the posterior distributions can be found using simple analytical equations that are easy to compute and so (in most cases) lead to efficient algorithms for inference.

In particular, it is useful to choose $p(\theta_{i\to j})$ and $p(\phi)$ as conjugate, so that the posterior distributions $p(\theta_{i\to j}|O_{i\to j}^{0:t'})$ and $p(\phi|\theta)$ are simple to calculate, which in turn simplifies the calculation of the predictive distribution. For example, if we assume that $O_{tr\to te}$ is a (univariate) Gaussian random variable, then $p(\theta_{tr\to te})$ is conjugate if it is a normal-inverse-gamma distribution, which generalises to become a normal-inverse-Wishart distribution when $O_{tr\to te}$ is multivariate [22]. [18] In this case, $p(\theta_{tr\to te}|O_{tr\to te}^{0:t'})$ will also be a normal-inverse-gamma distribution, obtained by simple equations involving the sample mean and sample variance of the observations $O_{tr\to te}^{0:t'}$.

4.3. Whitewashers and Group Behaviour

An important benefit of HABIT is its ability to predict a trustee's behaviour, based on the behaviour of groups of other agents. For example, suppose that *tr* repeatedly interacts with 20 trustees, which all tend to behave badly. If *tr* then encounters another trustee, *k*, then prior to observing its behaviour, it is impossible to know if *k*'s behaviour will follow this trend. Nevertheless, until *k* can prove otherwise, it is reasonable to predict that it will behave badly, because this is generally the case for all other encountered agents. On the other hand, if no such trend is observed, then it would be unreasonable to make such strong predictions. Instead, it may be better to assume that all possible behaviours are equally likely, or at most, tend to fall in a given range. As we now discuss, HABIT can account for all such possibilities, by making appropriate predictions based on any observed correlation in group behaviour. Moreover, unlike existing models, it can improve predictions by combining information from group behaviour, in a *principled way*, with information from a truster's direct experience of a trustee and its reputation.

The mechanism behind this ability can be illustrated by Bayesian model comparison. For example, suppose there are two alternative choices for ϕ, denoted ϕ_{good} and ϕ_{bad}, such that ϕ_{bad} predicts that all trustees behave badly and ϕ_{good} predicts that all trustees are good. By marginalising out all hidden confidence model parameters, $\theta_{tr\to j}$, we can obtain two alternative p.m.s for average trustee behaviour, $p(O_{tr\to\cdot}|\phi_{bad})$ and $p(O_{tr\to\cdot}|\phi_{good})$. Now, if a truster has observed 20 different agents, which are all bad, then the data likelihood of ϕ_{bad} will be higher than that for ϕ_{good}. HABIT would therefore place more weight in the prior belief that *k* will behave badly. However, if of the 20 previously observed trustees, 10 were bad and 10 were good, then neither hypothesis would dominate, and so *k* would be considered equally likely to be good or bad.

Of course, realistic settings are likely to be more complex than this simple example. In particular, there may be more than two (or possibly infinite) ways in which a trustee can behave, and due to differences in the number of direct observations and reputation, there may be varying degrees of uncertainty about the behaviour of each group member. Nevertheless, by applying Bayesian analysis, HABIT can automatically deal with these complexities, and so gain three main advantages, which are not found together in any existing trust model (see Section 2.3):

1. Since the degree of uncertainty about each trustee's behaviour is encoded in its confidence models, marginalising over its unknown behaviour parameters means that HABIT automatically accounts for this uncertainty. As such, when assessing a group based on its member's behaviour, those that are well known (due to frequent direct observations or reliable reputation) will influence this assessment more than members that are less well known.

[17] One exception to this rule is when non-parametric distributions are used to model dynamic behaviour (see Section 9).

[18] These distributions are derived by assigning a normal (i.e. Gaussian) distribution to the data mean and a gamma or Wishart distribution to the data distribution's variance or covariance matrix respectively. Alternatively, some authors assign a prior to the *precision* (the reciprocal of the variance or inverse covariance matrix), in which case the equivalent conjugate classes are referred to as normal-gamma or normal-Wishart.

2. Group behaviour will only affect predictions about an individual's behaviour to the extent warranted by the available evidence. Thus, as in the example above, if there is little evidence to suggest that one type of behaviour is more likely than any other, then predictions will not be unduly biased in its favour.
3. When making predictions about a trustee, HABIT will always combine information about group behaviour with other evidence in the most appropriate way, by taking into account the uncertainty in each type of evidence.

This ability to assess agents based on group behaviour can be applied in various ways. In particular, the simplest approach is for a truster to maintain a single reputation model for all agents, thereby enabling predictions about unknown agents by generalising from the behaviour of *all* other trustees. However, a more significant possibility is to first partition agents into non-overlapping groups containing agents with similar behaviour, and then maintain a *separate* reputation model for each group. Provided the behaviours of each group's members are more similar to each other than the population as a whole, predictions based on each group's reputation model will be more accurate than those based on a single shared reputation model for all agents.

For example, a pragmatic solution to the problem of whitewashing can be achieved by partitioning trustees into groups based on their total number of interactions.[19] Since whitewashers and newcomers will, by nature, have little or no recorded interactions with other agents, these will therefore be grouped together, separate from other better established agents. If whitewashing is a potential problem in a system, then a reputation model trained on a group of newcomers would learn to place less trust *a priori* in its members, based on their generally lower than average performance. In this way, HABIT can control whitewashing by collaborative filtering: the more whitewashing that is present in the system, the less newcomers will be trusted by the community as a whole, thereby reducing the incentive to whitewash in the first place. On the other hand, innocent newcomers will only be unfairly penalised in proportion to the number of whitewashers that are actually present in the system.

In addition to the number of recorded interactions, other domain dependent criteria may also be used to identify groups. For example, in many professions, service providers may be officially recognised by trading standard organisations, which verify qualifications, and enforce codes of conduct among their members. Membership of such bodies is therefore likely to be a good indicator of performance, so trustees that hold such memberships could be grouped together. Alternatively, groups of agents with similar behaviour may be discovered automatically, by applying any clustering algorithm to attributes relevant to the specific application domain. An in depth investigation into such methods is beyond the scope of this paper. However, as we demonstrate in Section 7.2, if the agents within a group do exhibit similar behaviour, then HABIT can harness this information to significantly improve prior predictions of performance. On the other hand, if there is little correlation in group behaviour (e.g. due to poor clustering or variations in trustee behaviour) then under normal circumstances,[20] HABIT's prediction accuracy will not be adversely affected.

4.4. Heterogeneous Reputation and Contexts

As mentioned previously, HABIT does not require different observers to use the same representation of behaviour. Instead, the only requirement is that each truster and reputation source models the distribution of a trustee's behaviour using some parameter vector, $\theta_{i \to j}$, so that a truster can model the joint distribution of the parameters used by different observers in its reputation model. This not only means that trusters can share information despite instantiating their confidence models using different parameter models (e.g. Gaussian or Poisson distributions), but it also means that trusters can share information about behaviour in different contexts.

For example, suppose two agents share information about search engines, despite one of them being exclusively interested in searching for text, while the other only searches for images. In effect, each agent is interested in assessing a search engine's ability in two different contexts. For a variety of reasons, a single search engine may be better at performing one of these types of service compared to another, but in general, it may be the case that search engines that are good at providing one service are also good at the other. In any case, HABIT can account for how much information one truster's observations provide about another, in effect learning the differences between contexts.

[19] For example, in many electronic marketplaces (such as online auction websites), it is possible for the total number transactions between a service provider and its clients to be recorded and published by a central authority.

[20] An adverse effect can only occur when the reputation model encodes a prior belief that trustees do exhibit similar behaviour, but no such correlation exists. However, in general, there is no reason to choose such a prior, unless there is reason to believe that such a correlation does exist.

Moreover, there is no reason why the same ability cannot be used for a truster to learn the correlations between its *own* observations of different contexts. To do so, a truster need only maintain different confidence models for each trustee, one for each context of interest. To use observations from one context to infer the trustee's behaviour in another, it simply needs to treat out-of-context observations as if they had been reported by a different observer, using the reputation model to learn the correlations between observations from different contexts.

In fact, the only restriction on different confidence models for different contexts and observers is that enough trustees are assessed using the same types of confidence model for the reputation model to learn the correlation between different sets of observations in the first place. This is because the reputation model works by assuming that the correlations that exist between confidence models are the same for all trustees or, more specifically, that all joint parameters $\theta_{\cdot \to j}$ are independent and identically distributed, according to ϕ. As with any parameter estimation problem, multiple observations are required in order to learn the parameters of a distribution. In the case of the reputation model, this means that learning ϕ requires information about multiple instances of $\theta_{\cdot \to j}$ or, in other words, multiple trustees are assessed using the same set of behaviour parameters.

5. Performing Inference with the HABIT Model

In the previous section, we described how HABIT can be instantiated in a variety of ways to meet different requirements. However, this ability to adapt the model would be meaningless if it were not possible to perform practical inference for a large class of instances. In this respect, HABIT has an advantage over existing statistical trust models because its structure maintains flexibility, while at the same time leads naturally to tractable inference with a variety of different behaviour representations.

The main reason for this is the way in which HABIT models reputation, which accounts for most of the model's complexity. Therefore, to understand this advantage, we must compare how reputation is modelled in HABIT to the three existing approaches identified in Section 2.2. Of these, the *majority-rules* approach is the easiest to generalise to different behaviour representations. This is because it works by identifying outliers among all the opinions about a trustee, and general methods for identifying outliers are already available in the literature [8]. However, as also discussed in Section 2.2, the assumption that the majority of opinions are correct does not always hold, and in some cases, outliers may represent real and important anomalies in a trustee's behaviour.

In contrast, the other two approaches (*all-or-nothing* and *reputation-function*) are robust in cases where even the majority of opinions are unreliable, but are also harder to generalise. This is because both model the perceived accuracy of a reputation source's opinions, based on the (often complex) statistical relationship between reputation and a truster's own direct experiences. The reason for this complexity is illustrated in Figure 2, which shows the structure of the Bayesian network used in HABIT in comparison to that used in both reputation-function and all-or-nothing models (which can be viewed as a special case).[21] Specifically, this common structure (illustrated on the left of the figure) is characterised by two sets of hidden parameters: one set representing the intrinsic behaviour of each trustee encountered, and another that encodes the reputation function for each reputation source.

The problem here is that each reported observation sets up a possible dependency between the parameters representing the observed trustee's intrinsic behaviour and the parameters representing the reputation function of the observer. In the figure, this is represented by the pair of directed edges that link each reported outcome to a matching pair of trustee behaviour and reputation function parameters [30]. Since these dependencies are transitive, they can propagate through the network each time an outcome for a specific pair of agents is reported, until evidentially all the hidden parameters become dependent. For example, if we condition on the bottom-centre variable (as illustrated on the left of Figure 2) the direction of the edges breaks one possible path to dependence between the top two nodes. However, the existence of the other observed variables means that the top two nodes are only conditionally independent when all three variables on the bottom row are observed. Generally, the more such dependencies exist, the harder it is to perform inference, unless these dependencies can be isolated using conditional independence relations. This applies even to approximate inference techniques, such as Monte Carlo algorithms or variational methods [43], which

[21] Strictly speaking, heuristic all-or-nothing models, including TRAVOS [68] and those proposed by Wang et al. [74], do not employ a Bayesian network like this, because they do not filter reputation using strict Bayesian analysis. However, the intuitation behind these models can still be analysed in this way, and may help explain why these models are so far limited to binary representations.

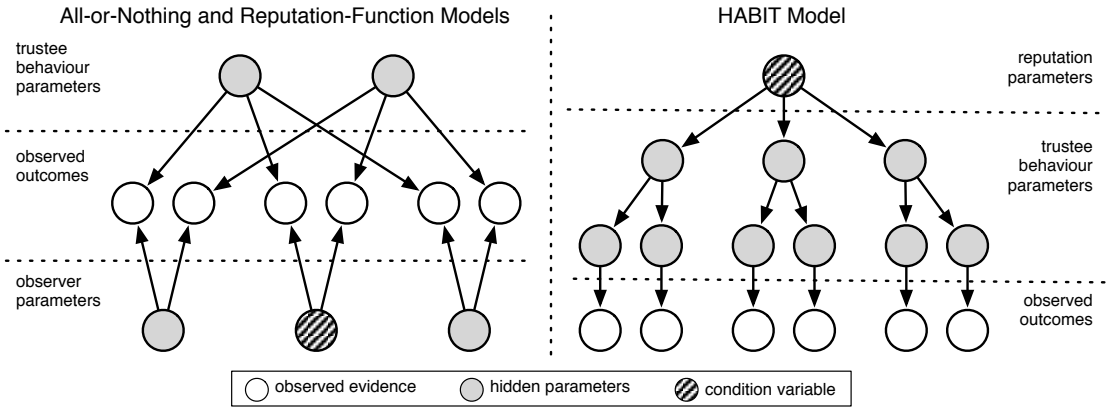

Figure 2: Comparison of HABIT to reputation function and all-or-nothing models, such as BLADE & TRAVOS.

generally require more sophistication and more samples to handle multiple highly dependent random variables (see below).

In contrast, HABIT's hierarchical structure means that dependencies between variables can be isolated by conditioning on a smaller set of hidden variables. For example, as illustrated, conditioning on the root variable makes each branch independent of the other. In general, this simplifies inference, but to what degree depends on how the model is instantiated. However, in many cases (such as described in Section 6) it can make the difference between: a model in which all or most reasoning steps can be performed analytically; or a model that requires reasoning about high dimensional joint parameter distributions (such as in TRAVOS-C) in which it can become exponentially hard to find a solution (within a given margin of error) as the amount of correlation between different agents' behaviour increases.

Nevertheless, as with most nontrivial Bayesian models, performing all inference analytically with HABIT is infeasible in general.[22] Instead, tractable algorithms must be sought that can approximate the optimal Bayesian solution within a reasonable amount of time, which is made easier by HABIT's hierarchical structure. 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:

$$EU|\mathcal{E} \approx \sum_{i=1}^{n} U(O_i) \tag{3}$$

with the accuracy of the estimate increasing as n becomes large [43]. 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 3.3); the generated values for Φ are simply discarded because they are not required. Second, we express this

[22] However, under certain circumstances, analytical solutions for this model are possible; for example, see Section 6.2.

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 $(i, j) \in \{(k, l) \mid \theta_{k \to l} \in \theta \setminus \theta_{tr \to te}\}$ **do**
    $\theta_{i \to j} \Leftarrow$ sample from $p(\theta_{i \to j} | O_{i \to j}^{0:t'})$
  end for
  $\phi \Leftarrow$ sample from $p(\phi | \theta \setminus \theta_{tr \to te})$
  $\theta_{tr \to te} \Leftarrow$ sample from $p(\theta_{tr \to te} | \theta_{. \to te} \setminus \theta_{tr \to te}, \phi, O_{tr \to te}^{0:t'})$
  $O_{tr \to te} \Leftarrow$ 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.}

joint distribution in terms of simpler conditional distributions as follows:

$$
\begin{aligned}
p(O_{tr \to te}, \Phi | \mathcal{E}) &= p(O_{tr \to te} | \Phi, \mathcal{E}) p(\Phi | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te}, \Phi \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi \setminus \theta_{tr \to te}, \mathcal{E}) p(\Phi \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi_{. \to te} \setminus \theta_{tr \to te}, O_{tr \to te}^{0:t'}) p(\Phi \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi_{. \to te} \setminus \theta_{tr \to te}, O_{tr \to te}^{0:t'}) p(\phi, \theta \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi_{. \to te} \setminus \theta_{tr \to te}, O_{tr \to te}^{0:t'}) \\
&\quad \times p(\phi | \theta \setminus \theta_{tr \to te}, \mathcal{E}) p(\theta \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi_{. \to te} \setminus \theta_{tr \to te}, O_{tr \to te}^{0:t'}) \\
&\quad \times p(\phi | \theta \setminus \theta_{tr \to te}) p(\theta \setminus \theta_{tr \to te} | \mathcal{E}) \\
&= p(O_{tr \to te} | \theta_{tr \to te}) p(\theta_{tr \to te} | \Phi_{. \to te} \setminus \theta_{tr \to te}, O_{tr \to te}^{0:t'}) \\
&\quad \times p(\phi | \theta \setminus \theta_{tr \to te}) \prod_{\theta_{i \to j} \in \theta \setminus \theta_{tr \to te}} p(\theta_{i \to j} | O_{i \to j}^{0:t'})
\end{aligned} \tag{4}
$$

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 4, and using the generated samples from the rightmost p.m.s in the equation to satisfy the conditional variables for the p.m.s to the left. This process is summarised in Algorithm 1.[23] 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.[24] Of these, $p(O_{tr \to te} | \theta_{tr \to te})$ can be chosen directly to suit the target application, while the other three distributions should be derived according to

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

[24] Significantly, this includes methods that can only be simulated using Markov Chain Monte Carlo or variational algorithms.

Equations 5 to 7, where $p(\theta_{i\to j})$ and $p(\phi)$ are suitable prior distributions.[25]

$$p(\theta_{i\to j}|O_{i\to j}^{0:t'}) \propto p(\theta_{i\to j}) \prod_{O\in O_{i\to j}^{0:t'}} p(O|\theta_{i\to j}) \tag{5}$$

$$p(\phi|\theta \setminus \theta_{tr\to te}) \propto p(\phi) \prod_{\vartheta\in\theta\setminus\theta_{tr\to te}} p(\vartheta|\phi) \tag{6}$$

$$\begin{aligned} p(\theta_{tr\to te}|\Phi_{\cdot\to te} \setminus \theta_{tr\to te}, O_{tr\to te}^{0:t'}) &\propto p(\theta_{tr\to te}|\theta_{\cdot\to te} \setminus \theta_{tr\to te}, \phi) \\ &\times \prod_{O\in O_{tr\to te}^{0:t'}} p(O|\theta_{tr\to te}) \end{aligned} \tag{7}$$

Ideally, $p(O_{tr\to te}|\theta_{tr\to te})$ and the distributions derived by Equations 5 to 7 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 the desired distribution. If this is possible, the number of samples required to accurately estimate the expected utility can be very low, and it is straightforward to calculate the estimation error (w.r.t. the utility) using the standard deviation of the generated samples.[26]

Ensuring that i.i.d. sampling is possible for $p(O_{tr\to te}|\theta_{tr\to te})$ can be achieved by simply choosing an existing parameter model for which this is possible. Similarly, choosing conjugate priors (see Section 4.2) to instantiate $p(O_{i\to j})$ and ϕ will, in many cases, mean that i.i.d. sampling is also possible for Equations 5 and 6. For example, the normal-inverse-gamma distribution, which is conjugate for Gaussian distributions, can be sampled from directly using algorithms for generating gamma and normal random variables [23, 64].

However, efficient i.i.d. sampling is less likely to be possible for Equation 7, because apart from the trivial case where $O_{tr\to te}^{0:t'} = \emptyset$ (i.e. a truster has no direct experience with a trustee), it is difficult to ensure that $p(\theta_{tr\to te}|\Phi_{\cdot\to te}\setminus\theta_{tr\to te})$ is conjugate with respect to $O_{tr\to te}^{0:t'}$. Moreover, some of the more sophisticated statistical models (such as infinite mixture models) and even some simple models (such as the conjugate class for gamma distributions [13, 16]) cannot be sampled from directly. In such cases, there are two existing types of solution to choose from.

Markov Chain Monte Carlo Methods: Commonly abbreviated to MCMC, these are a class of algorithms for generating a sequence of samples, where each sample depends on the previous sample in the sequence [43]. More specifically, MCMC methods produce a sequence of values $\mathbf{x}_1, \ldots, \mathbf{x}_k$ with domain $\mathcal{X}$, where each $\mathbf{x}_i$ is generated from a distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1})$, which depends on the previous sample $\mathbf{x}_{i-1}$. Here, $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ is known as the *transition* distribution, and $p(\mathbf{x}_i) = p(\mathbf{x}_{i-1}) = \int_{\mathcal{X}} p(\mathbf{x}_i|\mathbf{x}_{i-1})\, dp(\mathbf{x}_{i-1})$ is known as the *stationary* or invariant distribution. The rationale behind these methods is that, provided the stationary distribution is the distribution we wish to estimate, then the sample mean will converge to the expected value of the distribution as k becomes large, as is the case with independent sampling. The advantage is that, for many distributions, MCMC methods exist that are simple to implement, even if independent sampling is infeasible. However, the number of samples required to reach a specific level of accuracy is usually much greater than with independent sampling.

Variational Methods: This class of algorithms is used to estimate complicated probability distributions using one of a number of simpler types of distribution. Typically, this is achieved by minimising some measure of the difference between the target and approximate distributions, such as their Kullback–Leibler divergence. The result is a simpler distribution that can be used in place of the original for sampling purposes, and in analytical equations. These methods can often achieve a reasonable level of accuracy more efficiently than MCMC methods. However, unlike MCMC methods, the simplicity of the approximate distribution places an upper bound on the level of accuracy, which cannot be surpassed by simply generating more samples.

Such solutions are readily available for many useful parameter models. For example, both variational methods and MCMC techniques have been intensely studied for infinite mixture models [3, 53]. Similarly, for cases that involve non-conjugate priors, such as Equation 7, there are existing MCMC algorithms that are generally applicable [7, 12].

[25] The expression $x \propto y$ (read x is proportional to y) means that $x = cy$, where c is some constant. In the case of $p(x) = cf(x)$, where $p(x)$ is a valid p.m., then c is the unique value that ensures that $p(x)$ integrates to 1 over the domain of x.

[26] Typically, 30 to 100 (i.i.d.) samples will be sufficient for most applications [43].

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 $p(\phi|\theta \setminus \theta_{tr \to te})$ by generating a sequence of values labeled $\phi_1, \ldots, \phi_k$, such that, for each $i > 1$, $\phi_i \sim p(\phi_i|\theta \setminus \theta_{tr \to te}, \phi_{i-1})$. It is perfectly fine to use these in Line 6 of Algorithm 1, in place of independent samples from $p(\phi|\theta \setminus \theta_{tr \to te})$, with each ϕ_i being generated using different samples for $\theta \setminus \theta_{tr \to te}$ (generated by Lines 2 to 5 in the algorithm). Overall, the stationary distribution for the generated ϕ values will still be $p(\phi|\theta \setminus \theta_{tr \to te})$, and so the expected utility estimate will still converge to its true value. Similarly, any other of the required distributions can be sampled from using MCMC; convergence will still be guaranteed, albeit more slowly in terms of the total number of samples [18, 43].

To summarise, although it may be necessary to resort to variational or MCMC methods for some instances of HABIT, this is not a requirement of the general model. Moreover, even if such methods are used, the conditional independence relationships implied by HABIT's structure can still be used to isolate different sets of variables, allowing good estimates to be generated relatively quickly. In contrast, the structure of all-or-nothing and reputation function models (illustrated in Figure 2) introduces a high degree of dependence between the model parameters. For all but the simplest types of behaviour representation, this would necessitate the use of variational or MCMC methods in practice, and makes it difficult for such methods to converge to good solutions quickly. As such, even if all-or-nothing and reputation function models can be extended beyond the simple behaviour representations that they are currently limited to, we would reasonably expect their computational overhead to be much greater than an equivalent instance of HABIT. In fact, as we demonstrate in the next section, certain choices of reputation model allow HABIT to produce analytical equations for a large class of behaviour representations, enabling exact solutions to be generated efficiently *without* the need for approximation.

6. Instantiated Models for 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. For this purpose, this section introduces two instances of the generic HABIT model that are used for the empirical evaluation presented in Section 7, and are then adapted for the real-world experiments presented in Section 8.

These instances differ from each other in terms of their sophistication and complexity (which are discussed in Section 4.2). However, for evaluation purposes, both 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, which we use in Section 7 as a benchmark because it is representative of the state-of-the-art among statistically principled trust models. However, as HABIT can be instantiated in many different ways, it is important to emphasise that using a discrete representation is a choice made for evaluation and illustration purposes only, rather than a limitation of the generic HABIT model. To the contrary, adaption of these instances to continuous domains is straightforward, and was performed as part of the experiments we describe in Section 8.

To distinguish between these two instances, we refer to them by the main classes of distribution used to instantiate their confidence and reputation models (see Figure 1). Specifically, the first instance, referred to as the *DP-Dirichlet* model, instantiates its confidence models using Dirichlet distributions, and instantiates its reputation model using the non-parametric Dirichlet process model (DP). Similarly, the second instance, referred to as the *Gaussian-Dirichlet* model, also uses Dirichlet distributions in its confidence models, but uses a multivariate Gaussian distribution to instantiate its reputation model.

In both cases, Dirichlet distributions are used in the confidence models because these provide the standard Bayesian model for reasoning about a discrete random variable (in this case, trustee behaviour) given direct samples from the variable's distribution. In this respect, both instances are not only equivalent to each other, but also equivalent to many existing models of trust, including BLADE and the work presented by [31] and [55]. 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, these two models differ (both from each other, and existing trust models) in how they achieve the more complex problem of assessing a trustee's behaviour based on reputation and experience of other agents. In the

case of the DP-Dirichlet model, this is achieved by the DP model, which allows the expected utility of interacting with a trustee to be calculated analytically. Thus, inference according to the DP-Dirichlet model can be performed efficiently and exactly, which is the main theoretical advantage of this approach compared to other possible instances. In contrast, the Gaussian-Dirichlet model does not lead to analytical inference, and so more computationally expensive Monte Carlo sampling is required to approximate the expected utility of an interaction. Nevertheless, as we shall demonstrate in Section 7, the Gaussian-Dirichlet model can provide better predictions than the DP-Dirichlet model in certain circumstances, and demonstrates that more complex models are sometimes warranted to provide more accurate estimates of utility.

The following three subsections describe the theoretical aspects of the DP-Dirichlet and Gaussian-Dirichlet models in more detail. More specifically, Section 6.1 describes how Dirichlet distributions are used to instantiate the confidence models, which are common to both instances. This corresponds to the first two steps of the instantiation process (outlined in Section 4.1), which specify each of the distributions $p(O_{i\to j}|\theta_{i\to j})$ and $p(\theta_{i\to j})$. Finally, Sections 6.2 and 6.3 complete the last two steps of the instantiation process for each instance, and discuss how each model can be used for inference.

6.1. Learning from Direct Experience

In the generic HABIT model, the distribution of an interaction outcome $O_{i\to j}$ for each pair of agents, (i, j), is determined by a parameter vector $\theta_{i\to j}$. When $O_{i\to j}$ belongs to a finite and discrete domain, the most appropriate definition of $\theta_{i\to j}$ is generally a vector, $\langle\theta^{(1)}_{i\to j}, \ldots, \theta^{(k)}_{i\to j}\rangle$, where each $\theta^{(l)}_{i\to j}$ specifies the probability that $O_{i\to j}$ will be one of k possible values. Thus, the probability (given $\theta_{i\to j}$) that $O_{i\to j}$ takes on its lth possible value is defined as:

$$p(O_{i\to j} = l|\theta_{i\to j}) = \theta^{(l)}_{i\to j} \tag{8}$$

As described previously, a standard practice when performing Bayesian inference about an unknown parameter, such as $\theta_{i\to j}$, is to assign it a conjugate prior distribution. In the case of discrete distributions parameterised in this way, this usually means assigning a Dirichlet distribution, which has the following p.d.f:

$$p(\theta_{i\to j}) = \frac{\prod_{l=1}^{k}\left(\theta^{(l)}_{i\to j}\right)^{\alpha_l}}{Beta(\alpha)} \tag{9}$$

Here, $Beta(\cdot)$ is the multivariate Beta function, which acts as a normalising constant for the p.d.f., and is defined in terms of the gamma function as:

$$Beta(\alpha) = \frac{\prod_{l=1}^{k}\Gamma(\alpha_l)}{\Gamma(\sum_{l=1}^{k}\alpha_l)} \tag{10}$$

The shape of the distribution is specified by a hyperparameter vector, $\alpha = \langle\alpha_1, \ldots, \alpha_k\rangle$, of which high values correspond to high certainty about the true value of $\theta_{i\to j}$. Thus, it is usual to assign low initial values to α to represent an initial state of uncertainty about $\theta_{i\to j}$. Subsequently, once tr has obtained direct experience with te, the hyperparameters are updated, in line with Bayes rule, to represent the change in the truster's beliefs about the trustee in light of the new evidence. For Dirichlet distribution priors, this is equivalent to incrementing each α_l by the number of times, n_l, that $O_{i\to j}$ was observed to have its lth possible value:

$$p(\theta_{i\to j}|O^{0:t'}_{i\to j}) = \frac{p(O^{0:t'}_{i\to j}|\theta_{i\to j})p(\theta_{i\to j})}{p(O^{0:t'}_{i\to j})} = \frac{\prod_{l=1}^{k}\left(\theta^{(l)}_{i\to j}\right)^{\alpha_l+n_l}}{Beta(\alpha+n)} \tag{11}$$

where $n = \langle n_1, \ldots, n_k\rangle$. For example, from left to right, the plots in Figure 3 show one possible progression, for $k = 3$, from a prior belief that all possible values of $\theta_{i\to j}$ are equally likely (encoded by $\alpha_l = \langle 1, 1, 1\rangle$), to a certain belief that $\theta_{i\to j}$ is centred around $\langle 0.02, 0.82, 0.16\rangle$. Using this model, it is possible for a truster to calculate, analytically, the expected utility of interacting with te based on its direct experience. Given that the expected value of the Dirichlet

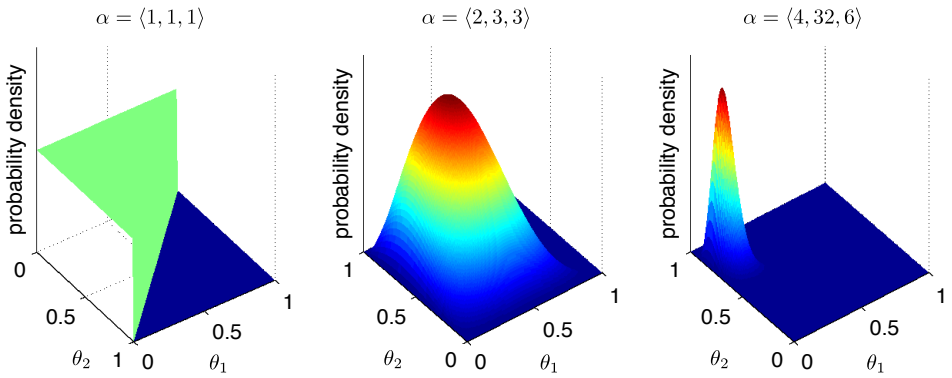

Figure 3: Example Dirichlet Distributions.

distribution is defined as $E[\theta_{i\to j}] = \alpha / \sum_{l=1}^{k} \alpha_l$, this is calculated as follows:

$$\begin{aligned} E[U(O_{tr\to te})|O_{tr\to te}^{0:t'}] &= \sum_{l=1}^{k} U(O_{tr\to te} = l) p(O_{tr\to te} = l|O_{tr\to te}^{0:t'}) \\ &= \sum_{l=1}^{k} U(O_{tr\to te} = l) E[\theta_{tr\to te}^{(l)}|O_{tr\to te}^{0:t'}] \\ &= \sum_{l=1}^{k} U(O_{tr\to te} = l) \frac{\alpha_l}{\sum_{i=1}^{k} \alpha_i} \end{aligned} \tag{12}$$

More interesting, however, is how the same can be achieved while taking into account an agent's reputation and observations of other services. This can be achieved by combining confidence models (instantiated using Dirichlet distributions) with one of the possible reputation models described in the following subsections.

6.2. The Dirichlet Process Reputation Model

Although, in general, inference in HABIT using both direct experience and reputation cannot be performed analytically, one way to instantiate the reputation model that can lead easily to analytical solutions is to use Dirichlet processes. In particular, when we adopt confidence models based on Dirichlet distributions with discrete behaviour representations, this leads to simple closed form equations to calculate the expected utility of interacting with any trustee, which can be calculated quickly and exactly without resorting to Monte Carlo or variational methods.

Not to be confused with Dirichlet distributions, a Dirichlet process is a non-parametric model for random values that have countably infinite (rather than finite) domains. First described by [69], this can be understood intuitively in terms of a generalised Pólya's urn sampling scheme. That is, suppose that we choose a ball (at random) from an urn containing balls of different colours. On observing the ball's colour we then return the ball to the urn along with a new ball that either has the same colour or a new colour, not previously contained in the urn. Moreover, the probability of the new ball having the same colour is proportional to the number of existing balls in the urn of that colour, and the probability of returning a ball of a different colour is proportional to a constant, $c_0 \geq 0$. This scheme has a reinforcing effect: colours that have been observed in the past are more likely to be chosen in the future because observing a colour increases the number of balls of that colour with positive probability.

Applied to our scenario, the coloured balls are analogous to trustees drawn from the population of agents. Moreover, any new (previously unobserved) trustee will have a specific value for $\theta_{\cdot\to te}$ with probability proportional to the number of previously encountered trustees with the same value. Alternatively, with probability proportional to c_0, a new trustee may have a value of $\theta_{\cdot\to te}$ different from any other that has been observed. In the following subsections, we describe what this means in terms of the probability distributions that make up the reputation model, show how these can be used to perform inference, and discuss the general theoretical properties of this reputation model.

6.2.1. The Reputation Parameters and their Distributions

In the generic HABIT model, we specify the reputation model in terms of the parameter vector, ϕ, its prior distribution, $p(\phi)$, and each of the conditional distributions, $p(\theta_{\cdot\to te}|\phi)$. However, strictly speaking, there is no fixed equivalent to ϕ that can fully specify a Dirichlet process precisely because it is a non-parametric model. Instead, a Dirichlet process is specified in terms of all previously observed evidence which, as more trustees are observed, allows the Dirichlet process to approximate their parameter distribution with arbitrarily high precision. More specifically, using our notation, a Dirichlet process is fully specified by:

- the chosen constant, c_0, which specifies the probability of encountering a trustee with $\theta_{\cdot\to te}$ different from any previously encountered;
- a prior distribution, $p_0(\theta_{\cdot\to te})$, from which hitherto unencountered values of $\theta_{\cdot\to te}$ are drawn; and
- the parameter vectors, $\{\theta_{\cdot\to j}\}_{j=1}^{n}$, of all previously encountered trustees, $1,\ldots,n$ (excluding the current trustee, te).

In this way, the role of ϕ is played by the combination of c_0 and all previously observed $\theta_{\cdot\to j}$, and the role of its prior, $p(\phi)$, is played by $p_0(\theta_{\cdot\to te})$. Similarly, rather than specifying $p(\theta_{\cdot\to te}|\phi)$, the posterior distribution of $\theta_{\cdot\to te}$ is specified by $p(\theta_{\cdot\to te}|\{\theta_{\cdot\to j}\}_{j=1}^{n})$, which is defined as follows:

$$p(\theta_{\cdot\to te}|\{\theta_{\cdot\to j}\}_{j=1}^{n}) = \frac{c_0}{c_0+n}\cdot p_0(\theta_{\cdot\to te}) + \frac{1}{c_0+n}\sum_{j=1}^{n}\delta_j(\theta_{\cdot\to te}) \tag{13}$$

where $\delta_j(\cdot)$ is a Kronecker delta function, defined as

$$\delta_j(\theta_{\cdot\to te}) = \begin{cases} 1 & \text{if } \theta_{\cdot\to te} = \theta_{\cdot\to j} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

In general, $p_0(\cdot)$ may be any suitable prior, depending on the how $\theta_{\cdot\to te}$ is instantiated. However, to be consistent with the analysis of the previous section, it is appropriate to define $p_0(\theta_{\cdot\to te})$ as a combination of the prior distributions used in the confidence models for each $\theta_{i\to te}$. More specifically, suppose that k_i is the number of possible outcomes that the i observer[27] may observe during an interaction with a trustee,[28] and that the prior distribution used in each of the confidence models associated with i is a Dirichlet distribution with hyperparameter vector $\alpha_i = \langle\alpha_{i,1},\ldots,\alpha_{i,k}\rangle$. The most consistent definition of $p_0(\cdot)$ is then:

$$p_0(\theta_{\cdot\to te}) = \prod_{i=1}^{m} p(\theta_{i\to te}) = \prod_{i=1}^{m}\frac{\prod_{l=1}^{k_i}\left(\theta_{i\to te}^{(l)}\right)^{\alpha_{i,l}}}{Beta(\alpha_i)} \tag{15}$$

where m is the number of observers. For example, for trinary outcome domains the posterior for each $\theta_{i\to te}$ in $\theta_{\cdot\to te}$ will have the form illustrated in Figure 4 (Part A), with point distributions added to the prior at the locations of each observed $\theta_{\cdot\to j}$.

These point distributions are what allows the Dirichlet process to approximate the distribution of $\theta_{\cdot\to te}$ with arbitrarily high precision: as more trustees are observed, more point distributions are added, giving a better reflection of the underlying distribution. However, on its own, the value of this flexibility in the model's form is limited because, for many choices of $\theta_{\cdot\to te} \setminus \theta_{tr\to te}$, the conditional distribution of $\theta_{tr\to te}$ will bypass the point distributions and so be equivalent to the prior, $p_0(\cdot)$. For example, in Figure 4 (Part A), every conditional distribution is a two dimensional slice through the joint distribution. Since the domain of the parameters is continuous, the majority of these slices will not include any of the spikes at observed parameter locations, and so will not be affected by them. Consequently, predictions about $\theta_{tr\to te}$ would effectively ignore the evidence provided by reputation.

[27] Here, an observer may be the truster itself or one of its reputation sources.

[28] Different observers may have different numbers of possible outcomes because they do not necessarily represent trustee behaviour in the same way (see Section 3).

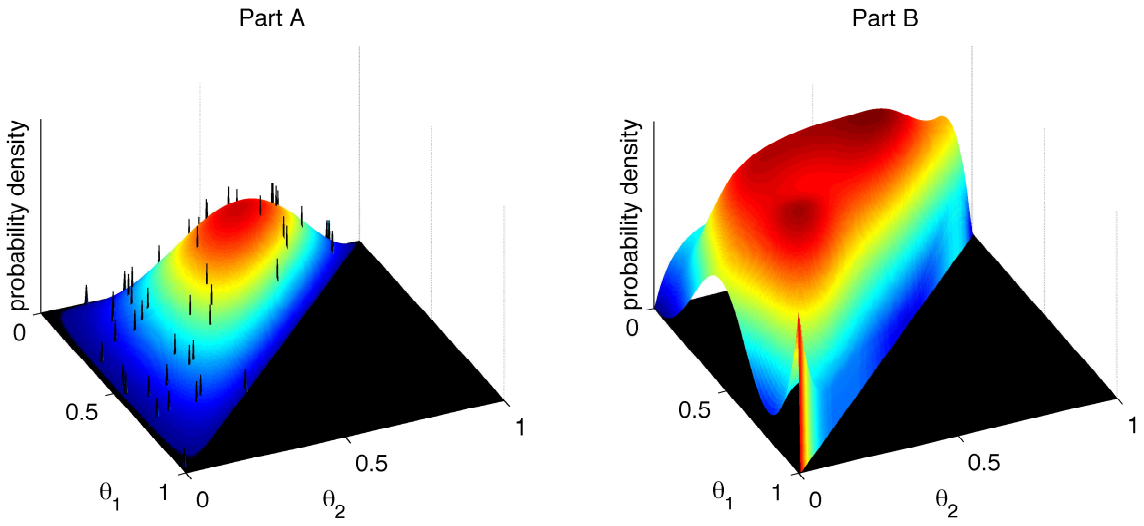

Figure 4: Dirichlet Process Predictive Distributions

However, in practice, a truster will never observe $\theta_{\cdot \to te}$ directly, and so will only ever have incomplete information about its value, inferred through all interaction outcomes, $O_{i \to te}$. The resulting distribution of $\theta_{\cdot \to te}$ is a mixture model, similar to that illustrated in Figure 4 (Part B), which is composed of the prior, $p_0(\theta_{\cdot \to te})$, and the confidence models for each previously observed trustee.[29] This extra uncertainty actually acts to the model's advantage by interpolating between the expected values of each $\theta_{\cdot \to te}$, smoothing out the p.d.f. More specifically, the posterior distribution of $\theta_{\cdot \to te}$ is obtained by marginalising out the unknown parameter values as follows:

$$
\begin{aligned}
p(\theta_{\cdot \to te} | \{O^{0:t'}_{\cdot \to j}\}_{j=1}^{n}) &= \frac{c_0 \cdot p_0(\theta_{\cdot \to te})}{c_0 + n} + \frac{1}{c_0 + n} \sum_{j=1}^{n} \int_O p(\theta_{\cdot \to j} | O^{0:t'}_{\cdot \to j}) \, d\delta_j(\theta_{\cdot \to te}) \\
&= \frac{c_0 \cdot p_0(\theta_{\cdot \to te})}{c_0 + n} + \frac{1}{c_0 + n} \sum_{j=1}^{n} p(\theta_{\cdot \to te} | O^{0:t'}_{\cdot \to j}) \\
&= \frac{c_0 \cdot p_0(\theta_{\cdot \to te})}{c_0 + n} + \frac{1}{c_0 + n} \sum_{j=1}^{n} \prod_{i=1}^{m} \mathcal{D}(\theta_{i \to te} | O^{0:t'}_{i \to j}) \qquad (16)
\end{aligned}
$$

where each $\mathcal{D}(\theta_{i \to te} | O^{0:t'}_{i \to j})$ is the posterior Dirichlet distribution, given the outcomes of all interactions between the ith observer and the jth trustee.

6.2.2. Performing Inference

Together with the confidence model defined in Section 6.1, the Dirichlet process model defined above can be used to perform sampling, as shown in Algorithm 1. As is generally the case for any reputation model (see Section 5), these samples could then be used to approximate a truster's expected utility for interacting with a trustee. However, although this may be the only available option for many reputation models, such sampling only approximates the true expected utility predicted by the model. In contrast, the expected utilities predicted by the Dirichlet process reputation model can be calculated exactly and analytically. This makes sampling unnecessary, and so we do not discuss it here in detail. Instead, the expected utility for a truster, tr, interacting with a trustee, te, can be derived as follows. First, if

[29] The plots in Figure 4 were generated using a Dirichlet distribution with $\alpha =< 2, 3, 2 >$ for $p_0(\theta_{\cdot \to te})$, 50 trustees with uniformly selected $\theta_{\cdot \to te}$, and 10 observations of each trustee used to form the mixture model in Part B.

$p_0(\theta_{\cdot\to te})$ is defined according to Equation 15 then Equation 16 is proportional to:

$$\begin{aligned} p(\theta_{\cdot\to te}|\{O^{0:t'}_{\cdot\to j}\}^n_{j=1}) &\propto c_0 \cdot p_0(\theta_{\cdot\to te}) + \sum_{j=1}^{n}\prod_{i=1}^{m} p(\theta_{i\to te}|O^{0:t'}_{i\to j}) \\ &\propto c_0 \cdot \prod_{i=1}^{m} \mathcal{D}(\theta_{i\to te}|\emptyset) + \sum_{j=1}^{n}\prod_{i=1}^{m} \mathcal{D}(\theta_{i\to te}|O^{0:t'}_{i\to j}) \end{aligned} \tag{17}$$

where $\mathcal{D}(\theta_{i\to te}|\emptyset)$ is the prior Dirichlet distribution of $\theta_{i\to te}$, given the empty observation set, $\emptyset$. Using this formulation, the predictive distribution given both direct experience and reputation is therefore

$$\begin{aligned} p(\theta_{tr\to te}|\mathcal{E}) &\propto p(\theta_{\cdot\to te}|\{O^{0:t'}_{\cdot\to j}\}^n_{j=1}, O^{0:t'}_{\cdot\to te}) \\ &\propto p(\theta_{\cdot\to te}|\{O^{0:t'}_{\cdot\to j}\}^n_{j=1}) \prod_{i=1}^{m} p(O^{0:t'}_{i\to te}|\theta_{i\to te}) \\ &\propto c_0 \cdot \prod_{i=1}^{m} \frac{Beta(O^{0:t'}_{i\to te})\mathcal{D}(\theta_{i\to te}|O^{0:t'}_{i\to te})}{Beta(\emptyset)} \\ &\quad + \sum_{j=1}^{n}\prod_{i=1}^{m} \frac{Beta(O^{0:t'}_{i\to j} \cup O^{0:t'}_{i\to te})\mathcal{D}(\theta_{i\to te}|O^{0:t'}_{i\to j} \cup O^{0:t'}_{i\to te})}{Beta(O^{0:t'}_{i\to j})} \end{aligned} \tag{18}$$

where $Beta(O)$ is the normalising constant for the posterior Dirichlet, $\mathcal{D}(\theta|O)$, defined for a parameter vector, θ, and conditioned on a set of observations, O. By grouping together all factors not involving $\theta_{tr\to te}$ to form a set of weights $\{w_j\}^n_{j=0}$, we find that this has the general form:

$$p(\theta_{tr\to te}|\mathcal{E}) \propto c_0 w_0 \cdot \mathcal{D}(\theta_{tr\to te}|O^{0:t'}_{tr\to te}) + \sum_{j=1}^{n} w_j \cdot \mathcal{D}(\theta_{tr\to te}|O^{0:t'}_{tr\to j} \cup O^{0:t'}_{tr\to te}) \tag{19}$$

This means that the conditional distribution of $\theta_{tr\to te}$ (given all available evidence) is a mixture of Dirichlet distributions comprising the distribution of $\theta_{tr\to te}$ given direct experience only, $\mathcal{D}(\theta_{tr\to te}|O^{0:t'}_{tr\to te})$, and the confidence models for all other encountered trustees, updated to account for the truster's direct experience of the trustee. Moreover, for each previously observed trustee, j, the corresponding Dirichlet is weighted according to how similar j's direct observations and reputation are to that of te. From this, the predictive distribution, $p(O_{tr\to te}|\mathcal{E})$, is given by the expected value of the Dirichlet mixture. The expected utility for tr interacting with te can thus be calculated analytically and precisely, by substituting $p(O_{tr\to te}|\mathcal{E})$ into Equation 12:

$$E[U(O_{tr\to te})|O^{0:t'}_{tr\to te}, \{O^{0:t'}_{\cdot\to j}\}^n_{j=1}] = \sum_{O_{tr\to te}} U(O_{tr\to te}) p(O_{tr\to te}|\mathcal{E}) \tag{20}$$

According to decision theory, a truster can then choose rationally which agents to interact with (taking into account all available evidence) by choosing those agents that maximise Equation 20.

6.2.3. Properties of the Dirichlet Process Reputation Model

Overall, there are two key advantages to the Dirichlet process reputation model. First, as a non-parametric model, it can approximate the joint distribution of trustee behaviour and reputation with arbitrarily high precision, given observations from a sufficient number of trustees. This contrasts with parametric models, which are constrained by the space of distributions that can be represented by their parameters. Second, as described previously, it provides analytical tractable predictions of expected utility that do not require approximation through sampling.

This not only enables solutions that are exact (according to the model), but also makes their calculation efficient. Specifically, from Equation 18 it is clear that the computational complexity of calculating the predictive distribution is $O(n \cdot m)$, where n is the number of observed trustees, and m is the number of observers (including the truster and its reputation sources). Moreover, although this equation is specific to the case of Dirichlet confidence models and multinomial behaviour representations, equivalent closed form equations can also be derived for other choices of

confidence model, enabling exact and efficient solutions for other types of behaviour representation. For example, in the experiments presented in Section 8, the *DP* truster makes predictions about a continuous behaviour representation using similar equations with the same complexity.

However, this approach does have one limitation: it relies on uncertainty about the trustees' parameters to smooth their joint distribution, which leads to useful predictions. Thus, 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.

6.3. The Gaussian Reputation Model

To overcome the potential problems of the Dirichlet process when trustee behaviour is well known, the alternative is to use a different reputation model, in which we assume that observing a trustee with parameter vector $\theta_{\cdot\rightarrow te}$ not only increases the likelihood of observing other trustees with $\theta_{\cdot\rightarrow j} = \theta_{\cdot\rightarrow te}$, but also increases the likelihood of observing $\theta_{\cdot\rightarrow j}$ within some neighbourhood of $\theta_{\cdot\rightarrow te}$. In this section, we show how this can be achieved using Gaussian distributions, since these are one of the simplest and widely used parameter models. However, the concept is the same for many other parameter models, including gamma distributions, Dirichlet distributions and various mixture models.

6.3.1. The Reputation Parameters and their Distributions

The principle here is to instantiate the reputation model by assuming that each joint parameter, $\theta_{\cdot\rightarrow te}$, is drawn from a multivariate Gaussian distribution with an unknown mean vector, μ and covariance matrix Σ. As this is a parametric model, the mapping between this instantiation and the generic model is more obvious than with the Dirichlet process. Specifically, in terms of the steps outlined in Section 4.1, we define ϕ as a vector comprising all elements of μ and Σ, which together specify how trustees behave in general and how informative a truster's reputation sources tend to be. In particular, μ specifies the average parameter values for all trustees, while the Σ matrix specifies how much the behaviour and reputation of individual trustees tend to deviate from this mean on average, and also how informative each reputation source's opinion is for determining a trustee's behaviour parameters, $\theta_{tr\rightarrow te}$. For each trustee, j, the conditional distribution $p(\theta_{\cdot\rightarrow j}|\phi)$ is therefore given by the standard Gaussian p.d.f:

$$p(\theta_{\cdot\rightarrow j}|\phi) = \frac{1}{\sqrt{2\pi^k|\Sigma|}} \exp\left[-\frac{1}{2}(\theta_{\cdot\rightarrow j} - \mu)^T\Sigma^{-1}(\theta_{\cdot\rightarrow j} - \mu)\right] \tag{21}$$

where k is the total number of scalar parameters in $\theta_{\cdot\rightarrow j}$. Likewise, the prior distribution $p(\phi)$ can be derived by placing a conjugate prior on the joint distribution of μ and Σ. Since these are the parameters of a Gaussian distribution, the chosen prior is thus a normal-inverse-Wishart distribution [22], which can be defined by the following p.d.f:

$$p(\phi) = p(\mu, \Sigma^{-1}) = p(\mu|\Sigma^{-1})p(\Sigma^{-1}) \tag{22}$$

where:

$$p(\mu|\Sigma^{-1}) = \frac{\sqrt{v}}{\sqrt{2\pi^k|\Sigma|}} \exp\left[-\frac{v}{2}(\mu - \mathbf{m})^T\Sigma^{-1}(\mu - \mathbf{m})\right] \tag{23}$$

$$p(\Sigma^{-1}) = \frac{|\Sigma^{-1}|^{(\alpha-(k+1))/2}}{\Gamma_k(\alpha)|B|^{\alpha/2}} \exp\left[-tr\left(B^{-1}\Sigma^{-1}\right)/2\right] \tag{24}$$

$$\Gamma_k(\alpha) = 2^{\alpha k/2}\pi^{k(k-1)/4}\prod_{i=1}^{k}\Gamma\left(\frac{\alpha + 1 - i}{2}\right) \tag{25}$$

Here, $2\alpha > k - 1$, B is non-singular, and $\mathbf{m}$, v, B, α are the hyperparameters defining the shape of the distribution. Given such a prior, the posterior distribution, $p(\phi|\theta \setminus \theta_{tr\rightarrow te})$, can be calculated by updating the hyperparameters as follows:

$$\alpha' = \alpha + n \tag{26}$$

$$v' = v + n \tag{27}$$

$$B' = B + S + \frac{vn}{v+n}(\bar{\theta} - \mathbf{m})(\bar{\theta} - \mathbf{m})^T \tag{28}$$

$$\mathbf{m}' = \frac{v\mathbf{m} + n\bar{\theta}}{v+n} \tag{29}$$

Algorithm 2 Monte Carlo Algorithm for Gaussian Reputation Model with no Direct Experience.

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 $(i, j) \in \{(k, l) \mid \theta_{k \to l} \in \theta \setminus \theta_{tr \to te}\}$ **do**
    $\theta_{i \to j} \Leftarrow$ sample from $p(\theta_{i \to j} | O_{i \to j}^{0:t'})$
  end for
  $\phi \Leftarrow$ sample from $p(\phi | \theta \setminus \theta_{tr \to te})$
  $\theta_{tr \to te} \Leftarrow E[\theta_{tr \to te} | \theta_{\cdot \to te} \setminus \theta_{tr \to te}, \phi]$
  $O_{tr \to te} \Leftarrow E[O_{tr \to te} | \theta_{tr \to te}]$
  $EU \Leftarrow EU + U(O_{tr \to te})/n$
end for

{EU is now an estimate of tr's expected utility for interacting with te.}

where n is the number of previously encountered trustees (excluding te), $\bar{\theta}$ is the sample mean and S is the sum of squares:

$$\bar{\theta} = \frac{1}{n} \sum_{j=1}^{n} \theta_{\cdot \to j} \tag{30}$$

$$S = \sum_{j=1}^{n} (\theta_{\cdot \to j} - \bar{\theta})(\theta_{\cdot \to j} - \bar{\theta})^T \tag{31}$$

6.3.2. Performing Inference

Together with the confidence model in Section 6.1, this fully defines the distributions required for sampling in Algorithm 1. Specifically, each $p(\theta_{i \to j} | O_{i \to j}^{0:t'})$ is the posterior Dirichlet distribution taken from each confidence model, $p(\phi | \theta \setminus \theta_{tr \to te})$ is the normal-inverse-Wishart distribution defined above, and $p(O_{tr \to te} | \theta_{tr \to te})$ is the discrete distribution with probabilities given by $\theta_{tr \to te}$. All of these distributions can be sampled from simply and independently with widely available algorithms (Section 5). This leaves only $p(\theta_{tr \to te} | \theta_{\cdot \to te} \setminus \theta_{tr \to te}, \phi, O_{tr \to te}^{0:t'})$, which can be sampled from using the MCMC method proposed by [12]. Moreover, in the special case where there is no direct experience of te, independent expected utility samples can be generated using a simpler procedure, provided by Algorithm 2. This works because the expected utility can be calculated analytically given $\theta \setminus \theta_{tr \to te}$, and so only these parameters need to be sampled.[30]

The only issue with using Gaussian distributions in this way is that their support is the entire real line. In contrast, the Dirichlet parameters, $\theta_{i \to j}$, over which the Gaussian is defined, must always sum to 1. Their domain is therefore much smaller than can be represented by a Gaussian, making it inconsistent with what we know. However, in practice this causes few problems because we never need to sample from the Gaussian directly (which could produce invalid parameter values that do not sum to 1) and, by choosing an appropriate prior, it is easy to ensure that the expected $\theta_{i \to j}$ value is always in the correct domain. Also, since the parameters must sum to 1, it is only necessary to model the distribution of the first $k - 1$ elements of each $\theta_{i \to j}$, for domains with k possible outcomes. The missing parameters can be deduced later, which enables more efficient inference by reducing the dimensions of the reputation model, and helps to avoid the numerical problems that may arise when calculating covariance matrices with highly correlated values.

6.3.3. Properties of the Gaussian Reputation Model

The main advantage of the Gaussian reputation model is that, unlike the Dirichlet process reputation model, it does not depend on uncertainty about trustee behaviour to smooth the joint parameter distribution. In theory, this means

[30] In all of the experiments discussed in Section 7, no direct experience was made available to the trustees in question. For this reason, all of these experiments used Algorithm 2.

that, in cases where a truster has only encountered a small number of trustees, but has interacted with them a large number of times, the Gaussian reputation model should provide more reliable predictions than the Dirichlet process model. However, inference using the Gaussian reputation model cannot be performed analytically, and so must be performed using Monte Carlo sampling in Algorithm 2. Although, like the Dirichlet process reputation model, the computational complexity is linear in the number of trustees and observers, multiple samples are now needed for each trustee-observer pair. This is generally more computationally expensive, particularly if MCMC methods are required to combine evidence from direct experience and reputation, in which case hundreds of samples may be required to produce an accurate estimate. Moreover, Gaussians are a relatively simple class of distributions, which can only be used to model linear dependencies between random variables (such as the parameters in the reputation model). Nevertheless, the principle of how they can be applied in the reputation model is the same as many other parameter models, including gamma distributions, Dirichlet distributions and various mixture models. They therefore provide a useful demonstration of how HABIT can be used in practice.

7. Simulated Experiments

Having defined two concrete examples of HABIT, we now use these instances to demonstrate its empirical performance compared to BLADE, since as a *reputation function* model, this can be viewed as the state-of-the-art among reputation-based trust models (see Section 2.2). More specifically, by instantiating the generic model as described in Section 4.1, the experiments presented here show that HABIT's general properties can translate into real performance benefits. In particular, the results in this section show that HABIT is up to twice as accurate as BLADE (in terms of mean absolute error) at predicting trustee behaviour. In the following subsections, we outline the methodology used in these experiments, and discuss the performance of the models when used to perform inference based on group behaviour (see Section 7.2) and reputation (see Sections 7.3 to 7.5). A similar set of experiments is also presented in Section 8, which additionally show that HABIT can accurately predict performance from real-world data using a continuous rather than discrete representation of behaviour.

However, in both sections, the effect of direct experience with a trustee in estimating its behaviour is not evaluated, because this is solely determined by the choice of probability distributions used to model direct experience, and does not depend on HABIT's unique structure. In particular, to enable a fair comparison with BLADE, we instantiate HABIT using Dirichlet distributions, which are the same family of probability distributions used to handle direct experience in BLADE. In this case, HABIT's ability to make predictions based on direct experience is therefore equivalent to BLADE's. Thus, to evaluate HABIT's performance, it is sufficient to focus on predictions based on reputation and group behaviour, since it is the way that HABIT handles these sources of evidence that is unique among trust models.

7.1. Experimental Design

To determine performance, all experiments presented in this section were conducted in a simulated environment in which five trusters (labeled *DP, GD-Conjugate, GD-Improper, Prior* and *BLADE*) 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:

- The DP agent assessed trustees using the DP-Dirichlet instance of HABIT, described in Section 6.
- The GD-Improper agent used the Gaussian-Dirichlet model, introduced in Section 6, with an *improper* prior distribution [14] placed on the reputation hyperparameters, ϕ. Effectively, this meant that no prior assumptions were made by this agent about reputation, or about the similarity between different trustees' behaviour. Instead, inferences in the reputation model relied on the observed evidence alone.
- The GD-Conjugate agent also used the Gaussian-Dirichlet model, but instead placed a conjugate prior distribution on ϕ, representing the prior belief that reputation provides no useful information about a trustee's behaviour.
- The BLADE agent, as its name suggests, performed all inference using the BLADE trust model, to provide a benchmark with which to compare the three HABIT-based agents.

- To provide a more basic benchmark, the Prior agent used only its prior beliefs to assess trustee behaviour, ignoring all other available evidence. Specifically, the Prior agent assumed that, for each trustee, all possible behaviours where equally likely. An equivalent assumption was made by the other four agents, *a priori*, which in their case could be revised in light of observed evidence. As such, the Prior agent represented how each of the other agents would perform if they ignored all observations of trustee behaviour.

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, which were all assumed to belong to the same group as the test trustee. This provided a basis on which trusters could (potentially) learn the average behaviour of a group of agents, and the reliability of a single source from which reputation was obtained. 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 solely on reputation and group behaviour.

Here, multiple groups were not considered, because HABIT's predictions about a trustee depend only on the observed behaviour of other members of its group. Thus, to evaluate HABIT's ability to make predictions based on group behaviour, it is sufficient to consider how varying the behaviour and number of other agents within a trustee's group affects the accuracy of HABIT's predictions.[31] Similarly, multiple concurrent reputation sources were not considered, because both HABIT and BLADE assume that the reliability of different sources is independent, and so each source is judged on its own merit. As such, the weight placed on a given reputation source is not affected by the weight placed on any other source, and so the effect of each source on prediction accuracy is also independent. To understand HABIT's performance, it is therefore important to measure its ability to evaluate individual sources with varying degrees of accuracy, and this is best achieved by considering opinions from a single source at a time.

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.[32] 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 [10].

7.2. 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, dictating 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.

[31] Since BLADE does not use group behaviour to make predictions, its performance is unaffected.

[32] 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.

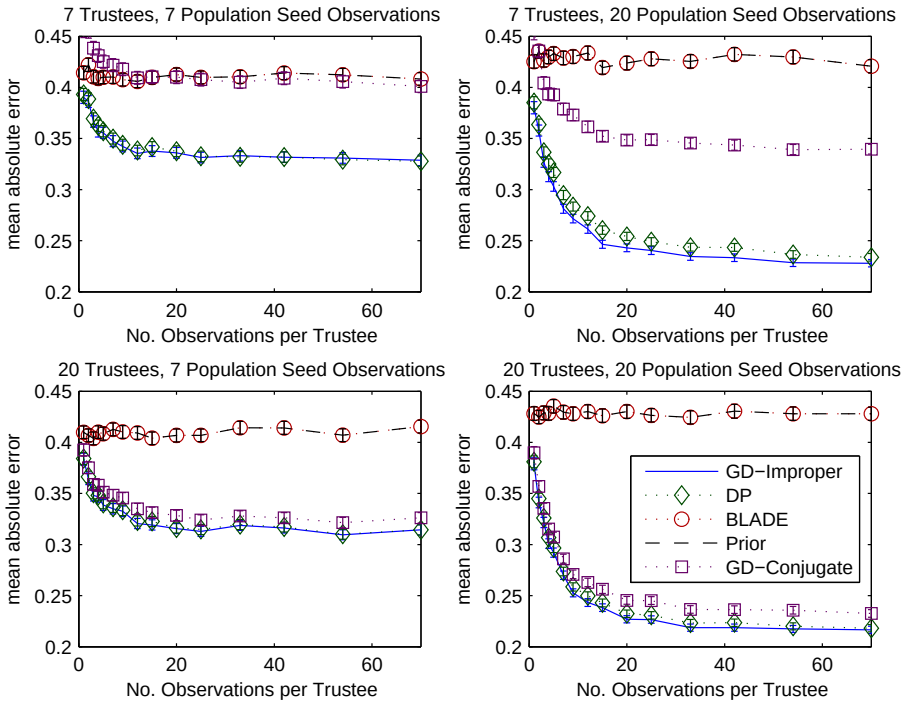

Figure 5: Group Behaviour

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 Dirichlet distributions, using the magnitude of the Dirichlet hyperparameters as a proxy for the similarity between agents. More specifically, we allowed the mean of the Dirichlet distribution to vary randomly by choosing it from a uniform distribution[33] 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 result in trustee parameters that deviate less from the mean. Therefore, by increasing the magnitude, we increase the amount of correlation between behaviour of the training trustees and the behaviour of the unknown test trustee.

The ability to decipher this information is demonstrated in Figure 5, 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 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, the fact that BLADE does not allow for possible dependencies between trustees' behaviour means that, in these experiments, it performed no better than the prior (which was shared by all trusters, including BLADE). 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 belief 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.

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

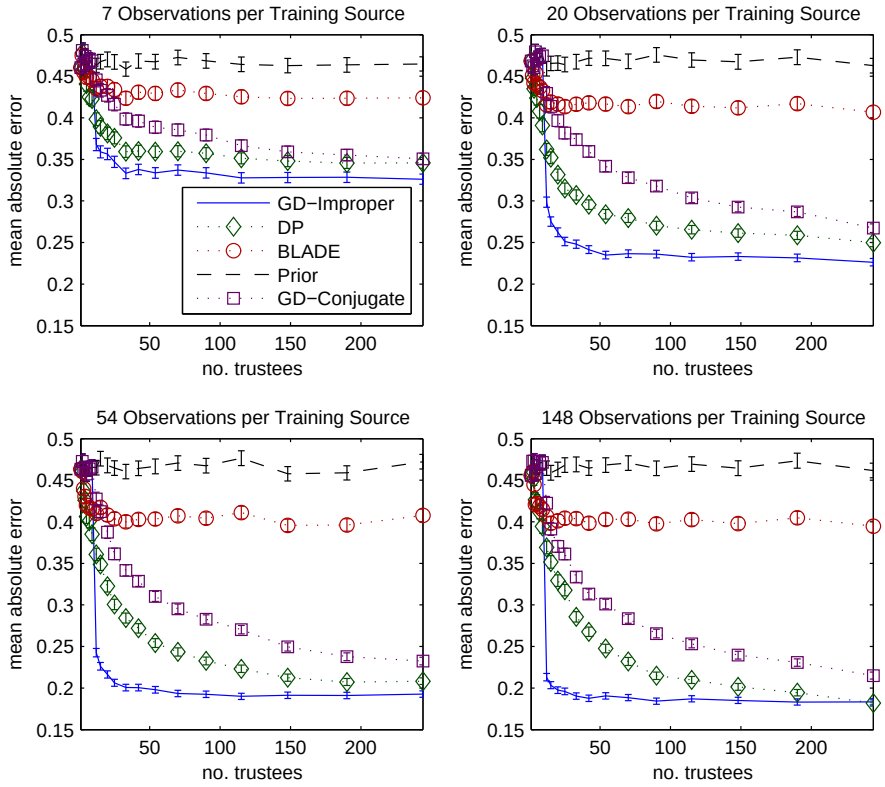

Figure 6: Perfect Reputation

7.3. 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 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 6 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 can be explained by the discussion in Section 6.3, in which we highlight the importance of having observed significant numbers of

trustees, relative to the certainty about their parameters. 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 and TRAVOS-C, 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 underlying 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.

In contrast, 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. However, by marginalising over HABIT's latent confidence model parameters (see Section 4.1), a report based on just a few observations is still judged to be less informative than one based on many. HABIT is therefore able to distinguish between more different types of uncertainty than BLADE, which explains HABIT's better performance in these experiments.

7.4. Learning from Unreliable Reputation Information

Although the previous set of experiments shows 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 distribution (independent of the trustee's behaviour), or with probability $1 - p$, it was drawn from the trustee's behaviour distribution.

In more detail, Figure 7 shows the results obtained, under equivalent conditions to Figure 6, 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 may be due to the non-parametric nature of the Dirichlet Process, which theoretically places fewer constraints on the shape of joint parameter distributions which, in some cases, may provide better results. However, more generally, this demonstrates that no model 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 8. 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 there is little evidence about the reliability of a reputation source, but the number of reported observations (and hence the reported confidence) 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

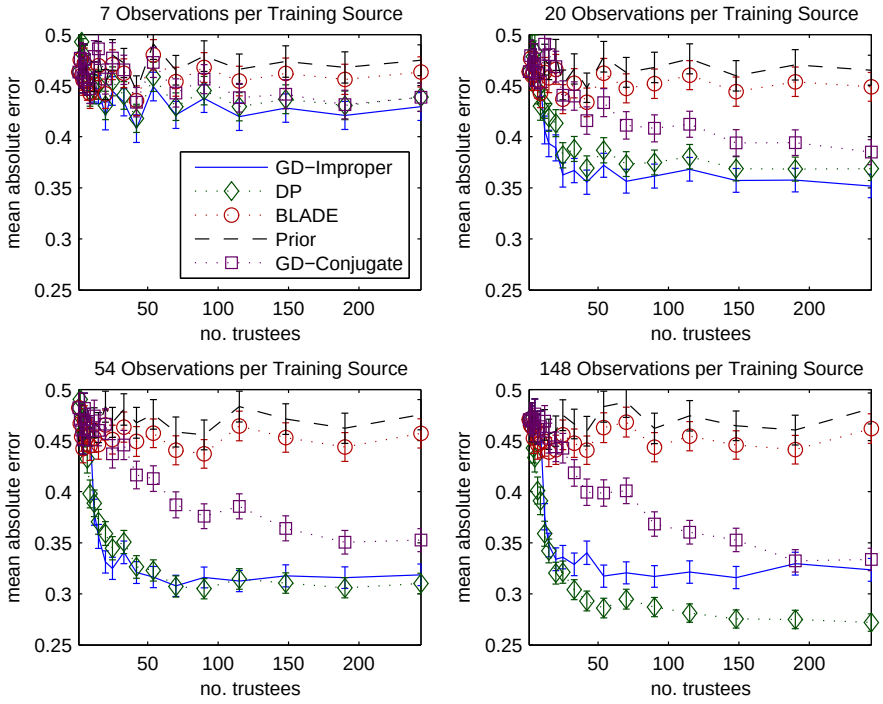

Figure 7: 50% Noisy Reputation

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 among the simplest 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.

7.5. Learning Rate

From the above experiments, it is clear that HABIT's learning rate varies significantly across the different instances and conditions evaluated. While this is to be expected, it is important to understand the reasons behind these differences, so that we can determine what to expect from HABIT's performance under difference circumstances. More specifically, our objective in the above experiments (and in trust modelling in general) is to estimate the expected utility for interacting with a trustee. By applying Bayesian analysis, all of the agents evaluated above (including BLADE) are able to achieve this optimally, given their respective modelling assumptions and the data available (see Section 2.1). Any differences in performance must therefore be due to differences in these assumptions, but since HABIT is a general model, these vary depending on the types of probability distribution used in each instance. Nevertheless, we can expect the following two properties to hold in general.

First, there is a generally a tradeoff between a model's learning rate and the limit of its accuracy. This is because simple models that restrict the space of possible probability distributions tend to learn faster than more flexible models, but often limit the achievable accuracy. This is demonstrated in our experiments when we compare the more flexible

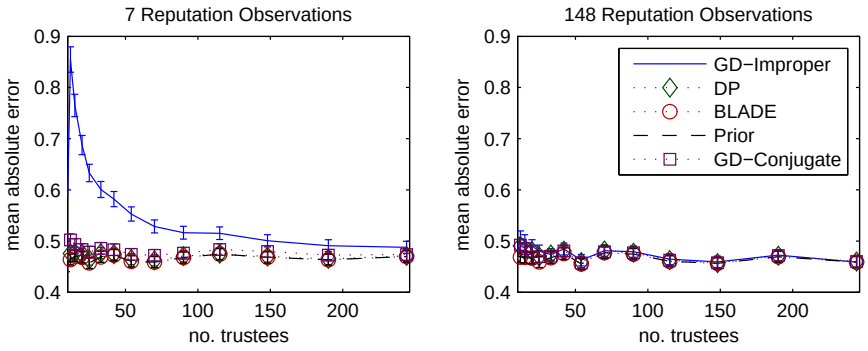

Figure 8: 100% Noisy Reputation.

Dirichlet Process reputation model, to the more constrained Gaussian reputation model. Given sufficient data, we find that the Dirichlet Process model always matches or outperforms the Gaussian model in terms of accuracy. However, this extra flexibility also comes at a cost, since the learning rate for the Gaussian model is faster in general, allowing it to outperform the Dirichlet Process when less data is available.

Second, since HABIT's reputation model is more flexible than any existing statistical trust model, we would generally expect it to have a better limit of accuracy than other models, even though it may sometimes require more data to reach this limit. In our experiments, this is demonstrated by BLADE, which generally requires less data than HABIT to reach the limit of its accuracy. Significantly, however, HABIT is still able to outperform BLADE in most cases, because it quickly surpasses the limit of BLADE's accuracy and carries on improving by extracting more information from more data.

Moreover, although none of the other reputation filtering mechanisms discussed in Section 2.2 can be applied to the multinomial behaviour representation used here, we can reasonably expect a similar comparison against these models. This is because, by applying the commonly used all-or-nothing approach, any reputation that shows any significant deviation from the truster's direct experience would be quickly discarded. While this is appropriate in the case of a completely unreliable reputation source, all reputation would eventually be discarded in the 50% noisy case as well. This would result in a loss of useful information, and a limiting accuracy equal to the prior. The only case where we might expect this approach to outperform HABIT is when reputation is perfect and identical to a truster's direct experiences. In this case, the all-or-nothing approach may learn to value reputation given less data than HABIT. However, this would require the approach to be extended to non-binary domains, and if reputation were in any way subjective, this would again result in most (if not all) reputation being discarded.

8. Web Polling Experiments

In the previous section, we showed that, when instantiated and applied to discrete domains, HABIT consistently outperforms BLADE when predicting trustee behaviour on the basis of group behaviour and third party opinions. However, although BLADE can only be directly applied to such domains, there are many situations in which trustee behaviour is more naturally represented by continuous random variables. In contrast, HABIT is a general framework, which, in principle, may be applied to any such representation. In this section, we demonstrate this ability by evaluating HABIT's performance in a web provision domain, in which different web servers (acting as trustees) were asked to respond to HTTP requests and assessed on their response time. Unlike the previous set of experiments, this data was not simulated, but was collected by polling a set of news websites hosted by a variety of different webservers from across the globe. The rationale for this approach is that, although news websites represent only one kind of service available on the Internet, they are easily accessible for the purpose of experimentation, and in terms of response times at least, provide a reasonable proxy for other types of data provision that one may expect to find in service-oriented domains, such as the semantic web or cloud computing. Moreover, while using real data does not allow us to evaluate the same range of controlled conditions that may be generated under simulation, it does allow us to test whether

HABIT's performance is robust when faced with continuous data from a real-world system. With this in mind, we begin in Section 8.1 by describing the process by which we collected data for these experiments, and how we then instantiated HABIT to model this data appropriately. In Section 8.2 we then describe the methodology and results of these experiments, which demonstrate that HABIT can indeed be applied to such domains, and be used to accurately predict trustee behaviour.

8.1. Data Collection and Modelling

As with the previous set of experiments, our goal here is to evaluate HABIT's ability to predict trustee behaviour, when presented with different amounts of information from sources with varying degrees of reliability. In particular, and in line with the previous experiments, we wish to demonstrate the following.

- When a truster has no direct experience with a trustee, it can decrease its mean estimation error by using reputation from third party sources.
- In the absence of direct experience with a trustee or relevant reputation, HABIT can decrease its mean estimation error by drawing on past experience of the group behaviour of other trustees.
- When compared to its prior estimate, HABIT's mean estimation error is never increased by taking into account either group behaviour or reputation.

Unlike our previous experiments, however, we wish to show that these hypotheses hold when applied to real rather than simulated trustee behaviour, represented by a continuous rather than discrete random variable. As a consequence, we must first sample the behaviour of a real system and choose appropriate instances of HABIT that can model this behaviour in a reasonable way. To this end, we began by randomly selecting a set of 50 globally distributed webservers to represent the trustees in our system, all belonging to the same group. As discussed in Section 7.1, HABIT's predictions about a trustee only depend on the behaviour of other members of its group. It is therefore sufficient to consider the behaviour of a single group, containing the trustee, rather than generating other groups, which do not affect predictions. To measure their behaviour, we then sent HTTP requests to each webserver, and recorded the time taken for each to respond. This was done from four different client PCs:

- one located on the University of Southampton network, which provided a source for the truster's own direct experience;
- two other PCs on the same network, which represented reliable reputation sources; and
- a fourth client, connected via a rural ADSL connection in Ireland, which represented a more remote, and thus less reliable reputation source.

In each case, data was collected by polling all 50 webservers at 1 minute intervals over the same 3–4 day period, resulting in 5186 data points per webserver-client pair with missing values removed.[34]

Having collected this data, we then had to decide how best to instantiate HABIT to model webserver behaviour. From this perspective, two main characteristics influenced our decision: (1) since behaviour was characterised by response times, negative values could never occur; and (2) for any given server, the distribution of response times tended to have a significant number of outliers with values more than 3 standard deviations above the mean. For example, this is illustrated in Figure 9, which shows a box plot of the response time distributions for a selection of webservers, as observed by the four client PCs. Each box is drawn with errorbars at approximately ± 2.7 standard deviations, a central line at the median, and edges at the 25th and 75th percentiles respectively. Outliers are marked by crosses and, for clarity, those beyond 300 milliseconds are compressed into a band, which is displayed at the top of the figure and is delineated by the horizontal dashed line.

Although various techniques could be used to model this behaviour, one simple approach is to fit a log normal distribution to the response time distribution for each server. From our perspective, this has three main advantages:

[34] In most cases, missing values were caused by client downtime.

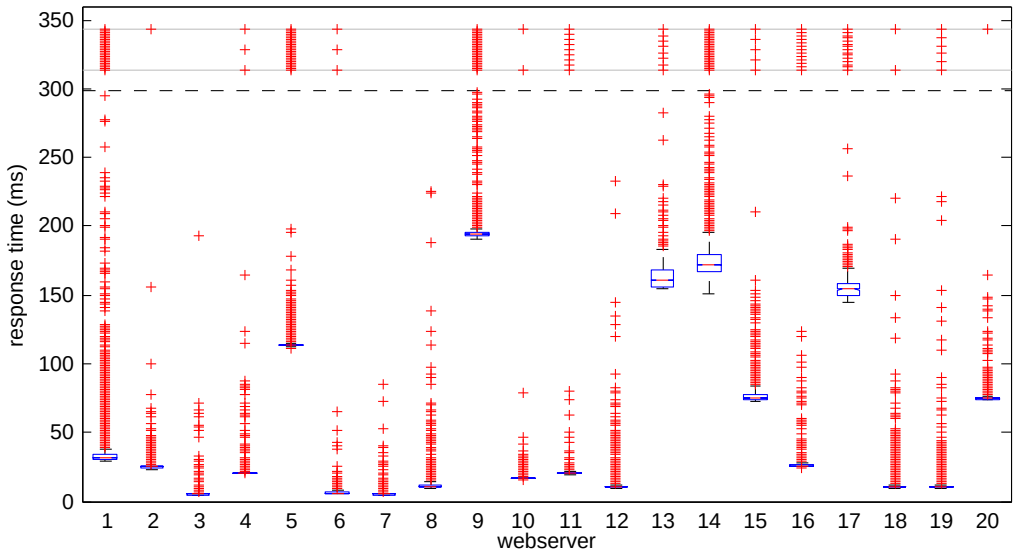

Figure 9: Box plot of response times for selected webservers.

1. the support of a log-normal distribution is $[0, \infty)$, meaning that negative response times are always assigned zero probability;
2. log-normal distributions feature a heavy tail toward larger values, and thus can (to some extent)[35] capture the range of outliers with long response times; and
3. they can be implemented by fitting a normal distribution to the log data, which allowed us to reuse much of our existing code from the previous experiments (see below).

Based on this analysis, we implemented two new instances of HABIT, by taking the Dirichlet Process and Gaussian Reputation Models described in Sections 6.2 to 6.3, and combining them with confidence models appropriate for modelling log-normal behaviour distributions. This was achieved by replacing the Dirichlet distributions used for the previous experiments with normal-inverse-gamma distributions, which are a special case of the normal-inverse-Wishart parameter models already used in the Gaussian reputation model. Although these are normally used to provide a conjugate prior for Gaussian distributions, they can also be used to perform inference about log-normal distributions by learning the corresponding normal distribution for the log response times. The required log-normal distribution can thus be recovered by a straightforward and well known transformation [29].

8.2. Evaluation

Having implemented these new instances of HABIT, we repeated experiments similar to those in Section 7, in which a set of trusters was presented with a controlled number of direct observations and reputation reports about a set of *training* trustees (see Section 7.1). As before, this information formed the basis from which HABIT could learn the predictive value of both group behaviour and reputation, by comparing the directly observed behaviour of each training trustee with their reputation reports and the behaviour of other trustees.[36] This was then used to predict the response time of a randomly selected *test* trustee with which the truster had no direct experience. From this, the mean absolute prediction error was recorded over multiple independent runs as a measure of each truster's performance.

To achieve this, the behaviour observed for each trustee was sampled from the previously acquired collection of webserver response times. More specifically, at the beginning of each run, both the training trustees and the test

[35] Even in the log scale, the distribution of observed response times still tended to feature heavy tails toward large values. Nevertheless, as the results in Section 8.2 show, the instances of HABIT we tested were still capable of good performance despite this characteristic of the data.

[36] As in Section 7, reported and direct observations were the only sources used by HABIT to assess the predictive value of reputation and group behaviour. In particular, the location of trustees and reputation sources was not revealed to the truster, nor was it told if they were remote or local.

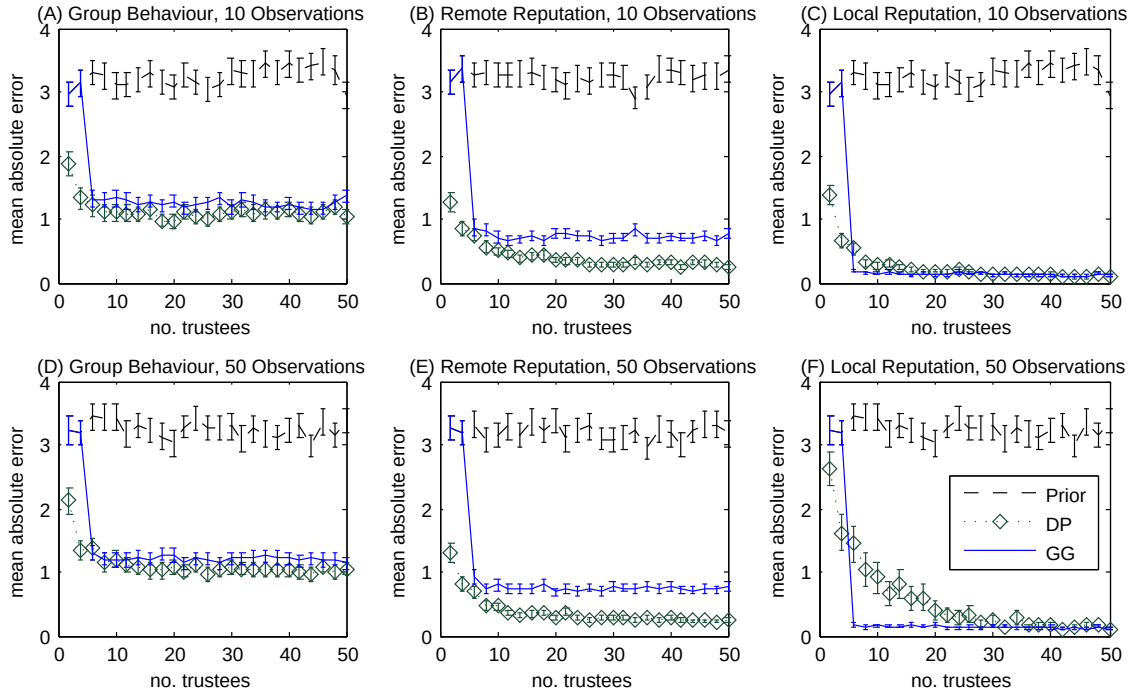

Figure 10: Web Polling Results.

trustee were randomly selected from the available set of 50 webservers. An equal number of direct observations for each training trustee was then sampled without replacement from the observations gathered by the truster's designated PC. Similarly, if reputation was made available, this was based on observations made by one of the other client PCs, with equal numbers of observations sampled for both the training trustees and the test trustee.

Using this methodology, we evaluated the performance of three trusters: *DP* and *GG*, which used the Dirichlet Process and Gaussian reputation models respectively, and *Prior* which, as before, made predictions based on the confidence model prior, ignoring all observations and reputation. Since we have already investigated the use of different reputation model priors in Section 7, here we used only improper priors in both the reputation and confidence models (which is why we only have one GG truster rather than GG-Improper and GG-Conjugate). Moreover, as explained previously, no truster based on BLADE was tested since this could not be directly applied without some arbitrary discretisation of the data. Given that all three trusters shared the same confidence model prior, all would return the same prior estimate, and so achieve equivalent results if no observations or reputation are provided. Ideally, (as was mostly the case in the simulated experiments) this means that the Prior estimate should represent an upper bound on mean prediction error, since any increase would mean that the truster was misled by the data.

Representative results from these experiments are plotted in Figure 10, which shows the mean absolute error for log response times[37] achieved by each truster when the number of training trustees ranged from 1 to 50, and the number of observations per training trustee was changed from 10 (top row) to 50 (bottom row). In particular, the first column shows the prediction performance when no reputation was provided (hence the trusters had to rely on group behaviour only), the second column plots the results when reputation was received from the remote client PC in Ireland, and the final column shows the results for reputation from one of the local clients based at Southampton. Results for reputation from the fourth client, also at Southampton, were similar to those in the final column, and so are not shown.

As might be expected, the results here show similar trends to those from the simulated experiments. This demon-

[37]Here, absolute errors are plotted w.r.t. log rather than normal response times in the interest of clarity only. Changing to the normal scale has no significant impact on the conclusions or analysis.

strates that HABIT can indeed provide relatively low estimation errors for both discrete and continuous representations of behaviour, and can be applied successfully to data from a real system. In particular, referring back to the hypotheses in Section 8.1, we found that all three hold for all of the conditions that we tested for this set of experiments. Both instances tested were therefore able to improve on their prior estimate based on both group behaviour and reputation. Perhaps the only unexpected result here is that the minimum error achieved is as low as it is: close to zero when local reputation is used. This can be attributed to the typically low variance in response times for individual webservers in this domain, and the high correlation between response times observed by different PCs on the Southampton network. As such, these results should only be interpreted in relative terms. With higher variance in response times, we would reasonably expect the mean error to increase. However, this effect would be experienced by any trust model, not just HABIT, since the variance in trustee behaviour places a limit on how well it can be predicted.

Focusing on the GG truster specifically, we can observe an obvious step trend in which predictions based on reputation are more accurate than those based on group behaviour alone, and are particularly so when the reputation is provided by a local rather than remote PC. There are two reasons for this. First, while group behaviour can only predict how trustees behave in general, reputation is specific to the trustee in question, and so can be more informative. Second, since the Southampton clients all shared the same network infrastructure, we were able to observe a higher correlation between their observed response times relative to the more remote Irish PC. As such, the GG truster was able to learn that reputation from a local source was almost as informative as its own direct experience. The only anomaly with the GG truster's performance is that, due to the improper prior adopted in its reputation model, it requires observations from at least five trustees to form a valid distribution from which to make predictions. With less than that, the best it can do is return an estimate based on the confidence model prior, which is why its accuracy matches that of the Prior truster in these cases. While this is likely to be only a minor issue for the majority of domains,[38] it may be overcome by providing a more informative prior.

Finally, while the GG and DP trusters did achieve similar performance for predictions based on group behaviour, there are clear differences between their performance when reputation is available. Once again, this can be attributed to the different nature of the reputation models. That is, since the DP model can represent a larger and more flexible set of distributions, it can generally learn more complex relationships between reputation and direct experience. As a result, it has the potential to outperform the Gaussian reputation model, as it does here when using remote reputation. However, as discussed in Section 6.2.1, the DP model actually requires confidence model uncertainty in order to generalise well. In the case of local reputation, we observed that the standard deviation in response times observed by the local Southampton PCs was generally less than that for the remote client. Together with the high correlation between observations made by the Southampton clients, this would naturally lead to sparse mixture models more similar to Figure 4 (Part A), than the more smooth and informative distribution in Figure 4 (Part B). As in the previous experiments, this problem can be overcome, provided the number of observed trustees is large relative to the number of observations per trustee. This effect can be seen clearly in Figure 10 (Parts C and F), in which the performance of the DP truster approaches that of the GG truster as the number of training trustees increases.

9. Conclusions

In open environments, software and hardware services can be integrated dynamically across organisational and geographical boundaries, to fulfil user requirements. Unfortunately, from a user's perspective, there may be significant uncertainty surrounding the incentives and capabilities of providers that offer such services, and the potential failure of certain resources must be considered normal. Thus, such systems must be able to adapt dynamically and automatically to changing circumstances by allocating resources to meet both new requirements and existing requirements when resources fail. In light of this, it has been argued that such systems should be modelled as multi-agent systems, in which autonomous software agents can trade resources in an open market, despite potentially conflicting interests. In particular, such agents must be able to assess the trustworthiness of their peers, so that they only choose suppliers of reliable, high-quality services, to minimise the chance of service failure and maximise their gain.

To address this need, 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 five key advantages, which together make a significant contribution to the state-of-the-art:

[38] In applications where reputation is useful, we would expect an agent to interact with many more than 5 trustees during its lifetime.

1. HABIT can assess trust based on reputation, even if the reputation sources that supply this information use different representations or semantics for trustee behaviour. Moreover, HABIT is robust in cases where such information is inaccurate or intentionally misleading. This is important because, in an open and dynamic environment, agents are likely to encounter reputation sources that are malicious or assess trustee behaviour according to different criteria.
2. 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. In systems that are susceptible to the *whitewasher* problem or frequently acquire new members, this enables reasonable decisions to be made about previously unencountered agents, for which there is little or no specific information.
3. 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.
4. HABIT provides a statistically principled and tractable framework, which can be adapted to assess trust in a wide range of scenarios with different modelling requirements. This is important because there can be no single trust model that best suits every possible application. Instead, HABIT provides a strong theoretical basis from which to create more specialised models to meet the needs of any given application. In particular, by adopting the hierarchical structure of the generic HABIT model, any instance of HABIT naturally inherits the beneficial properties of this framework, including the abilities described above and (by exploiting conditional independence between confidence models) the potential for computational efficiency. As demonstrated in Section 7, these properties *do not* follow from the use of Bayesian statistics alone, which is why HABIT is able to outperform BLADE as a state-of-the-art probabilistic trust model.
5. Finally, while performance benefits may be achieved by fine-tuning HABIT to a particular application, good performance may still be achieved be selecting an off-the-shelf instance of HABIT designed for use in a number of domains. For example, the instances described in Section 6 may be applied (without modification) to any domain that uses a discrete behaviour representation. Thus, without creating instances beyond those described here, HABIT can already be applied to a range of problems as large as any targeted by existing probabilistic trust models. However, unlike existing models, instances of HABIT can be designed for other domains without difficulty. Thus, it would be straightforward to create a library of instances suitable for a large range of applications.

Although HABIT has a number of advantages, there are some areas in which further work is needed, and in particular, we identify the following four.

Comparison with Other Trust Models As argued earlier, we chose BLADE as a benchmark in our experiments because it is representative of the state-of-the-art, and as a *reputation-function* model, is capable of extracting useful information from reputation in more cases than any other existing trust model. Nevertheless, it would be useful to compare HABIT against other trust models, and identify any cases in which they may provide better performance. In particular, by comparing against *all-or-nothing* models, such as those described in [72] and [74], we may discover cases in which it is best to assume that reputation is either equivalent to direct experience, or completely unreliable. Moreover, if such cases do exist, it may be worth investigating instances of HABIT that incorporate this assumption.[39] However, as current all-or-nothing models are limited to binary representations, a fair comparison would require trustee behaviour to be limited to binary outcomes, and so would require different experimental conditions than those used in this paper.

Evaluation of Additional Instances Although the instances of HABIT presented in this paper can be used to model trust and reputation in wide range of scenarios, other instances may be appropriate in some cases. It would therefore be advantageous to expand HABIT's repertoire, by investigating more sophisticated instances than those presented here. For example, these could include the use of infinite mixture models [53] in the reputation

[39] This may be easily achieved, by restricting each $\theta_{k \to te}$ to be either independent of $\theta_{tr \to te}$, or equal to it. See Section 3 for details.

model to overcome the narrow peaks sometimes possible in the current application of the Dirichlet process (see Section 6.2). Although this would come at the cost of its analytical tractability, it could potentially combine the flexibility of the Dirichlet process with better performance when there is a high degree of certainty about a small number of trustees. In particular, if mixtures of conjugate distributions were possible, then integrating evidence from direct experience with evidence from group behaviour and reputation would be more straightforward.

Modelling Dynamic Behaviour Throughout this paper, we have made the implicit assumption that the distribution of an agent's behaviour is static, and so not time dependent. This choice was made in the interest of clarity, allowing us to focus on the group and reputation modelling features of HABIT, which form its main contributions. Nevertheless, in real world problems, both trustee and reputation source behaviour is likely to change with time, so an ability to model dynamic behaviour is important in many realistic settings. Fortunately, just as HABIT can be instantiated in different ways to handle different types of behaviour distribution, it can also easily incorporate existing models of dynamic behaviour. For example, this may be achieved by instantiating HABIT's confidence models using standard techniques for modelling dynamic phenomena, such as Gaussian Processes,[40] or Hidden Markov Models. Incorporating such techniques into the reputation model is also possible, and may be useful to identify correlations or discrepancies in a reputation source's reliability over time. However, in many cases, we believe that the extra complexity of a dynamic reputation model may not be necessary. Instead, it is sufficient for the reputation model to refer to the *current* beliefs encoded in each agent's confidence model, which may nevertheless be based on past experiences. Intuitively, this means that reputation sources will be judged only on their latest opinions about each trustee, which is exactly the approach taken by many other trust models, including BLADE and Vogiatzis et al.'s approach [72].

Modelling Contextual Information Finally, another significant possibility is the use of more detailed contextual information. Currently, in HABIT, group behaviour can only be accounted for by assigning agents to a single group, or to fixed groups based on external factors. However, Rettinger et al. [58, 59] have shown how hierarchical Bayesian models can be used to assign more weight to prior experience of interactions with trustees that has taken place in a similar context to the current decision being made. For example, the advertised cost of a product, the time of year, or the point of origin are all factors that could potentially affect trust and can be accounted for using their model. Although this work does not account for reputation, it has significant synergy with the approach taken in HABIT, and so there is the potential to combine the two approaches to bring to bear a wide range of information to predict a trustee's performance.

References

[1] Ashri, R., Ramchurn, S. D., Sabater, J., Luck, M., Jennings, N. R., 2005. Trust evaluation through relationship analysis. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems. ACM Press, Utrecht, the Netherlands, pp. 1005–1011.

[2] Berger, J. O., 1993. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag.

[3] Blei, D. M., Jordan, M. I., 2006. Variational inference for dirichlet process mixtures. Bayesian Analysis 1 (1), 121–114.

[4] Burnett, C., Norman, T., Sycara, K., 2011. Sources of stereotypical trust in multi-agent systems. In: Proceedings of the 14th International Workshop on Trust in Agent Societies. p. 25.

[5] Burnett, C., Norman, T. J., Sycara, K., 2010. Bootstrapping trust evaluations through stereotypes. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. pp. 241–248.

[6] Burnett, C., Norman, T. J., Sycara, K., 2011. Trust decision-making in multiagent systems. In: Proceedings of the Twenty Second International Joint Conference on Artificial Intelligence. pp. 115–120.

[7] Carlin, B. P., Gelfand, A. E., 1991. An iterative monte carlo method for nonconjugate bayesian analysis. Statistics and Computing 1 (2), 119–128.

[8] Chandola, V., Banerjee, A., Kumar, V., Jul. 2009. Anomaly detection: A survey. ACM Computing Surveys 41 (3), 15:1–15:58.

[9] Chapin, P. C., Skalka, C., Wang, X. S., 2008. Authorization in trust management: Features and foundations. ACM Computing Surveys 40 (3), article 9.

[10] Cohen, P. R., 1995. Empirical Methods for Artificial Intelligence. M.I.T. Press.

[11] Şensoy, M., Zhang, J., Yolum, P., Cohen, R., 2009. Poyraz: Context-aware service selection under deception. Computational Intelligence 25 (4), 335–366.

[40] Modelling dynamic behaviour is therefore an exception to the rule excluding non-parametric models (such as Gaussian Processes) from instantiating the outcome distributions (see Section 4.2): practical inference with the reputation model can be achieved with non-parametric confidence models, provided a fixed size parameter vector can be found to represent behaviour in the current time-step.

[12] Damien, P., Wakefield, J., Walker, S., 1999. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxillary variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61 (2), 331–344.
[13] Damsleth, E., 1975. Conjugate classes for gamma distributions. Scandinavian Journal of Statistics 2 (2), 80–84.
[14] DeGroot, M., Schervish, M., 2011. Probability & Statistics, 4th Edition. Pearson Education.
[15] Despotovic, Z., Aberer, K., 2005. Probabilistic prediction of peers' performance in p2p networks. Engineering Applications of Artificial Intelligence 18 (7), 771–780.
[16] Diebolt, J., El-Aroui, M., Garrido, M., Girard, S., 2005. Quasi-conjugate bayes estimates for gpd parameters and application to heavy tails modelling. Extremes 8 (1), 57–78.
[17] Escobar, M. D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90 (430), 577–588.
[18] Evans, M., Swartz, T., 1999. Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford University Press.
[19] Fullam, K., Klos, T., Muller, G., Sabater, J., Schlosser, A., Topol, Z., Barber, K., Rosenschein, J., Vercouter, L., Voss, M., 2005. A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies. In: Proceedings of 4th International Joint Conference on Autonomous Agents and MultiAgent Systems. ACM Press, Utrecht, the Netherlands, pp. 512–518.
[20] Fung, C. J., Zhang, J., Aib, I., Boutaba, R., 2011. Dirichlet-based trust management for effective collaborative intrusion detection networks. IEEE Transactions on Network and Service Management 8 (2), 79–91.
[21] Gambetta, D., 1988. Can we trust trust? In: Gambetta, D. (Ed.), Trust: Making and Breaking Cooperative Relations. Basil Blackwell, Ch. 13, pp. 213–237, reprinted in electronic edition from Department of Sociology, University of Oxford.
[22] Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2003. Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC.
[23] Gentle, J. E., 1998. Random number generation and Monte Carlo methods. Springer-Verlag.
[24] Ghosh, J. K., Ramarmoorthi, R. V., 2003. Bayesian Nonparameterics. Springer-Verlag.
[25] Hazard, C. J., Singh, M. P., 2011. Intertemporal discount factors as a measure of trustworthiness in electronic commerce. IEEE Transactions on Knowledge and Data Engineering 23 (5), 699–712.
[26] Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing 65, 550–557.
[27] Jaynes, E. T., 2003. Probability Theory: The Logic of Science. Cambridge University Press.
[28] Jennings, N. R., 2001. An agent-based approach for building complex software systems. Communications of the ACM 44 (4), 35–41.
[29] Johnson, N. L., Kotz, S., Balakrishnan, N., 1994. Continuous univariate distributions, 2nd Edition. Vol. 1. Wiley.
[30] Jordan, M. I., 1999. Learning in Graphical Models. M.I.T. Press.
[31] Jøsang, A., Haller, J., 2007. Dirichlet reputation systems. In: Proceedings of the 2nd International Conference on Availability, Reliability and Security. Vienna, Austria, pp. 112–119.
[32] Jøsang, A., Ismail, R., 2002. The beta reputation system. In: Proceedings of the 15th Bled Conference on Electronic Commerce. Bled, Slovenia, pp. 324–337.
[33] Jøsang, A., Ismail, R., Boyd, C., 2007. A survey of trust and reputation systems for online service provision. Decision Support Systems 43 (2), 618–644.
[34] Jurca, R., Faltings, B., 2003. Towards incentive-compatiable reputation management. In: Falcone, R., Barber, S., Korba, L., Singh, M. (Eds.), Trust, Reputation and Security: Theories and Practice. Vol. 2631 of Lecture Notes in Artificial Intelligence. Springer-Verlag, pp. 138–147.
[35] Jurca, R., Faltings, B., 2007. Obtaining reliable feedback for sanctioning reputation mechanisms. Journal of Artificial Intelligence Research 29, 391–419.
[36] Khan, K. M., Malluhi, Q., 2010. Establishing trust in cloud computing. IT Professional 12 (5), 20–27.
[37] Lee, P. M., 2004. Bayesian Statistics: An Introduction, 3rd Edition. Hodder Arnold.
[38] Li, N., Mitchell, J. C., Winsborough, W. H., 2005. Beyond proof-of-compliance: Security analysis in trust management. Journal of the ACM 52 (3), 474–514.
[39] Liau, C. J., 2003. Belief, information acquisition and trust in multiagent systems — a model logic approach. Artificial Intelligence 149 (1), 31–60.
[40] Liu, S., Zhang, J., Miao, C., Theng, Y., Kot, A. C., 2011. iclub: An integrated clustering-based approach to improve the robustness of reputation systems (extended abstract). In: Proceedings of the 10th International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 1151–1152.
[41] Liu, X., Datta, A., Rzadca, K., 2011. Trust beyond reputation: A computational trust model based on stereotypes. Tech. Rep. abs/1103.2215, Computing Research Respository (CoRR).
[42] Liu, X., Datta, A., Rzadca, K., Lim, E., 2009. Stereotrust: A group based personalized trust model. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM Press, pp. 7–16.
[43] Mackay, D. J. C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.
[44] Mahapatra, A., Tarasia, N., Sahoo, M., Paikray, H. K., Das, S., 2011. A fuzzy approach for reputation management in online communities for bittorrent p2p network. International Journal of Computer Science and Information Technologies 2 (4), 1564–1568.
[45] Miller, R. B., 1980. Bayesian analysis of the two-parameter gamma distribution. Technometrics 22 (1), 65–69.
[46] Murillo, J., Munoz, V., López, B., Busquets, D., 2010. Developing strategies for the art domain. In: Current Topics in Artificial Intelligence. Vol. 5988 of Lecture Notes in Computer Science. Springer-Verlag, pp. 171–180.
[47] Osman, N., Robertson, D., 2007. Dynamic verification of trust in distribued open systems. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, pp. 1440–1445.
[48] Parsons, S., Tang, Y., Sklar, E., McBurney, P., Cai, K., 2011. Argumentation-based reasoning in agents with varying degrees of trust. In: Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems. pp. 879–886.
[49] Pinyol, I., Centeno, R., Hermoso, R., Torres da Silva, V., Sabater-Mir, J., 2010. Norms evaluation through reputation mechanisms for bdi agents. In: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence. pp. 9–18.
[50] Ramchurn, S. D., Huynh, D., Jennings, N. R., 2004. Trust in multi-agent systems. The Knowledge Engineering Review 19 (1), 1–25.
[51] Ramchurn, S. D., Mezzetti, C., Giovannucci, A., Rodríguez-Aguilar, J. A., Dash, R. K., Jennings, N. R., 2009. Trust-based mechanisms for

robust and efficient task allocation in the presence of execution uncertainty. Journal of Artificial Intelligence Research 35, 119–159.
[52] Ramchurn, S. D., Sierra, C., Godo, L., Jennings, N. R., 2004. Devising a trust model for multi-agent interactions using confidence and reputation. International Journal of Applied Artificial Intelligence 18 (9-10), 883–852.
[53] Rasmussen, C. E., 2000. The infinite gaussian mixture model. Advances in Neural Information Processing Systems 12, 554–560.
[54] Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. MIT Press.
[55] Reece, S., Roberts, S., Rogers, A., Jennings, N. R., 2007. A multi-dimensional trust model for heterogeneous contract observations. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence. AAAI Press, Vancouver, British Columbia, Canada, pp. 128–135.
[56] Reece, S., Rogers, A., Roberts, S., Jennings, N. R., 2007. Rumours and reputation: Evaluating multi-dimensional trust within a decentralised reputation system. In: Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems. Honolulu, Hawaii, USA, pp. 1063–1070.
[57] Regan, K., Poupart, P., Cohen, R., 2006. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In: Proceedings of the 21st National Conference on Artificial Intelligence. Boston, USA, pp. 206–212.
[58] Rettinger, A., Nickles, M., Tresp, V., 2008. A statistical relational model for trust learning. In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems. Estoril, Portugal, pp. 763–770.
[59] Rettinger, A., Nickles, M., Tresp, V., 2011. Statistical relational learning of trust. Machine Learning 82 (2), 191–209.
[60] Ries, S., Heinemann, A., 2008. Analyzing the robustness of certaintrust. In: Trust Management II. Vol. 263 of IFIP Advances in Information and Communication Technology. Springer Boston, pp. 51–67.
[61] Roussas, G. G., 2004. An Introduction to Measure-theoretic Probability. Elsevier Academic Press.
[62] Sabater, J., Sierra, C., 2002. Social regret, a reputation model based on social relations. SIGecom Exchanges 3 (1), 44–56.
[63] Sen, S., 2002. Believing others: Pros and cons. Artificial Intelligence 142 (2), 179–203.
[64] Smith, W. B., Hocking, R. R., 1972. Wishart variate generator. Applied Statistics 21, 341–345.
[65] Sun, L., Jiao, L., Wang, Y., Cheng, S., Wang, W., 2005. An adaptive group-based reputation system in peer-to-peer networks. In: Proceedings of the 1st International Workshop on Internet and Network Economics. Vol. 3828 of Lecture Notes in Computer Science. Springer-Verlag, Hong Kong, China, pp. 651–659.
[66] Teacy, W. T. L., 2006. Agent-based trust and reputation in the context of inaccurate information sources. Ph.D. thesis, School of Electronics and Computer Science, University of Southampton.
[67] Teacy, W. T. L., Huynh, T. D., Dash, R. K., Jennings, N. R., Luck, M., Patel, J., 2007. The art of iam: The winning strategy for the 2006 competition. In: The 10th International Workshop on Trust in Agent Societies. pp. 102–111.
[68] Teacy, W. T. L., Patel, J., Jennings, N. R., Luck, M., 2006. Travos: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems 12 (2), 183–198.
[69] Thomas, F. S., 1973. A bayesian analysis of some nonparametric problems. The Annals of Statistics 1 (2), 209–230.
[70] Venanzi, M., Piunti, M., Falcone, R., Castelfranchi, C., 2011. Facing openness with socio cognitive trust and categories. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. pp. 400–405.
[71] Venanzi, M., Piunti, M., Falcone, R., Castelfranchi, C., 2011. Reasoning with categories for trusting strangers: a cognitive architecture. In: Proceedings of the 14th International Workshop on Trust in Agent Societies. pp. 109–124.
[72] Vogiatzis, G., MacGillivray, I., Chli, M., 2010. A probabilistic model for trust and reputation. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. pp. 225–232.
[73] Wang, X., Maghami, M., Sukthankar, G., 2011. A robust collective classification approach to trust evaluation. In: Proceedings of the International Workshop on Trust in Agent Societies. pp. 125–139.
[74] Wang, Y., Hang, C., Singh, M. P., 2011. A probabilistic approach for maintaining trust based on evidence. Journal of Artificial Intelligence Research 40, 221–267.
[75] Wang, Y., Singh, M. P., 2007. Formal trust model for multiagent systems. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India, pp. 1551–1556.
[76] Whitby, A., Jøsang, A., Indulska, J., 2004. Filtering out unfair ratings in bayesian reputation systems. In: Proceedings of the 7th International Workshop on Trust in Agent Societies. New York, USA, pp. 106–117.
[77] Zacharia, G., Moukas, A., Maes, P., 2000. Collaborative reputation mechanisms for electronic marketplaces. Decision Support Systems 29 (4), 371–388.