UNIVERSITY OF SOUTHAMPTON

An Investigation into Trust & Reputation for Agent-Based Virtual Organisations

by

W. T. Luke Teacy

Technical Report

Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

October 26, 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

by W. T. Luke Teacy

Trust is a prevalent concept in human society. In essence, it concerns our reliance on the actions of our peers, and the actions of other entities within our environment. For example, we may rely on our car starting in the morning to get to work on time, and on the actions of our fellow drivers, so that we may get there safely. For similar reasons, trust is becoming increasingly important in computing, as systems, such as the Grid, require computing resources to work together seamlessly, across organisational and geographical boundaries (Foster et al., 2001). In this context, the reliability of resources in one organisation cannot be assumed from the point of view of another. Moreover, certain resources may fail more often than others, and for this reason, we argue that software systems must be able to assess the reliability of different resources, so that they may choose which resources to rely upon.

With this in mind, our goal here is to develop a mechanism by which software entities can automatically assess the trustworthiness of a given entity (the trustee). In achieving this goal, we have developed a probabilistic framework for assessing trust based on observations of a trustee's past behaviour. Such observations may be accounted for either when they are made directly by the assessing party (the truster), or by a third party (reputation source). In the latter case, our mechanism can cope with the possibility that third party information is unreliable, either because the sender is lying, or because it has a different world view. In this document, we present our framework, and show how it can be applied to cases in which a trustee's actions are represented as binary events; for example, a trustee may cooperate with the truster, or it may defect. We place our work in context, by showing how it constitutes part of a system for managing coalitions of agents, operating in a grid computing environment. We then give an empirical evaluation of our method, which shows that it outperforms the most similar system in the literature, in many important scenarios.

Contents

List of Figures

List of Tables

Nomenclature

$\emptyset$	The empty set
$\mathcal{A}$	The set of agents
a_{tr}	The truster
a_{te}	The trustee
a_{rep}	A reputation source
$\mathcal{H}_{a_{tr},a_{rep}}$	The history of opinions given to a_{tr} by a_{rep}
$\rho_{a_{tr},a_{rep}}$	The estimated accuracy of a_{rep} according to a_{tr}
θ	The distribution parameter vector
ϕ	The hyperparameter vector
$\mathcal{R}^r$	The reputation source opinion under consideration.
$d(\theta_{a_{tr},a_{te}}\|\phi_r)$	The distribution that results from $\mathcal{R}^r$.
$d(\theta_{a_{tr},a_{te}}\|\phi_a)$	The adjusted distribution for $\mathcal{R}^r$.
$d(\theta_{a_{tr},a_{te}}\|\phi_o)$	The distribution of actual observed outcomes, for which a_{rep} give opinions similiar to $\mathcal{R}^r$.
$d(\theta_{a_{tr},a_{te}}\|\phi_c)$	The overall reputation distribution
$d(\theta_{a_{tr},a_{te}}\|\phi_{c-r})$	The overall reputation distribution, excluding $\mathcal{R}^r$

Acknowledgements

The work presented in this document is part of the CONOISE-G project, funded by the DTI and EPSRC through the Welsh e-Science Centre, in collaboration with the Office of the Chief Technologist of BT. Many thanks go to Nicholas R. Jennings and Michael Luck.

Chapter 1

Introduction

In this document, we present a model for reasoning about *trust* between intelligent software agents, operating in a large open distributed system, such as the Grid, pervasive computing and the semantic Web. In the present chapter, we introduce the concept of trust, describe its relevance to computing and outline the objectives of our work. Specifically, this chapter is structured as follows: Section 1.1 introduces our notion of trust; Sections 1.2 and 1.3 highlight the relevance of trust in computing and the specific set of problems we wish to address; Section 1.4 discusses why we are interested in agent systems in particular; Section 1.5 specifies the objectives of our work; and finally, Section 1.6 sets the scene for the material covered in the subsequent chapters.

1.1 The Meaning of Trust

In human society, we constantly rely on the actions of other people. Whether we're concerned about getting our post delivered on time, or buildings not collapsing around us, it is other people's actions that determine such things. Unfortunately, there is often a great deal of uncertainty surrounding the behaviour of our fellow humans. We cannot, in general, read minds, so we cannot be certain about intentions. Likewise, we cannot always tell whether others have the resources to fulfil their promises.

Managing this uncertainty is something we do almost subconsciously in our daily lives. For instance, in the workplace when we need to delegate a task, we normally choose a person who we believe is willing and able to do it (unless we have no option). Also, we may choose not to disclose information to someone if we believe they will use that information to our disadvantage. Both of these cases (and many more) involve assessing the future action of a person or other entity, and in such cases it is common to use the notion of *trust* (Misztal, 1996).

The concept of trust is thus prevalent in society and we use it in many contexts. As with many words in natural language, it is a term that is used frequently, understood implicitly, but not well defined. In this discussion, however, we are not concerned with finding a universally accepted definition. Instead, we adopt the definition given by Gambetta (1988), which captures the notion we are interested in:

"Trust is a particular level of the subjective probability with which an agent will perform a particular action, both before she can monitor such an action, and in a context in which it affects her own action."

There a number of points in this definition that warrant elaboration.

1. **Trust relates to a particular action** — Although sometimes we talk generally about our trust in an individual, a high level of trust in someone to perform one type of action does not imply a high level of trust in them to perform another. For example, just because we can trust a person to pick up a pen does not mean we trust them to run the country!

2. **Trust is a subjective probability** — Trust is subjective, because it is assessed from the unique perspective of the truster. It is dependent both on the individual set of evidence available to the truster and her relationship with the trustee.

3. **Trust is defined to exist before the respective action can be monitored** — Trust is a prior belief about an entity's actions. It is an assessment made in a context of uncertainty. Once the truster knows the outcome of an action, she no longer needs to assess trust in relation to that outcome. Consider the difference in the statements, 'I know you have brushed your teeth' and 'I trust that you have brushed your teeth'.

4. **Trust is situated in a context in which it affects the truster's own action** — By this, we mean that our interest is limited only to those actions of a trustee that have relevance to the truster. Specifically, we are interested in trustee actions that, if their outcomes are known, would usefully inform the truster's action decisions.

We believe that this is a strong definition because it captures both the purpose of trust, and its nature in a form that can be reasoned about. The purpose of trust is to aid an entity make decisions, when the goals of those decisions are affected by the behaviour of other entities (this is the subject of Point 4). Trust is by nature a probability of an entity performing a particular action. Defining trust as a probability allows us to reason about it analytically, much more readily than the loose concept that it is in natural language.

1.2 The Relevance of Trust in Computing

Issues of trust are becoming increasingly important in information technology, due to a predominant trend in modern computing: the shift towards large-scale open distributed systems. In such systems, users and software can interact with information services, computing resources and other users with whom they are unfamiliar or have no physical contact. Issues of trust arise from such interactions, just as they do in everyday life. For instance, we may ask if we trust an information service to provide us with accurate information, or a particular website to respect the privacy of credit card details. The participation of large numbers of entities with conflicting interests in a large open system means that these examples are not isolated. In particular, we identify three important (possibly overlapping) areas in computing that trust concerns:

Security — Broadly, computer system security can be viewed as an attempt to limit the actions that individuals or software can perform with a given computer system. We can view this problem as reverse trust, or equivalently fear (see Section 2.2.1): trust is concerned with a wish for an entity to perform an action, whereas fear is concerned with a wish for an entity *not* to perform an action. In this respect, we wish to avoid malicious actions, such as manipulation of important data or reading of trade secrets, and therefore attempt to limit the ability to perform such actions only to those who are unlikely to have incentive to act maliciously.

Traditionally, computer security has been concerned with lower level issues such as authentication, whereby the identity of a user is determined; authorisation, in which access to resources is granted; and data encryption (Gollmann, 1998; Pfleeger, 2002). Recently, however, some in this field have started to refer explicitly to issues of trust, though in some cases, the term has been used merely as a synonym for authorisation or authentication (Grandison and Sloman, 2000). Others refer to it as a richer concept; they see it as a prerequisite condition for authorisation. In this vein, Blaze et al. (1996) introduce the concept of *trust management*. Essentially, trust management is concerned with specifying and applying *security policies*, which state precisely what actions can be performed by a given entity.

Service Provision — In contrast to security, service provision concerns actions that a trustee is obliged to perform. Prime examples of this can be found in the semantic web (Berners-Lee et al., 2001), Pervasive Computing (Adelstein et al., 2004) and the Grid (Foster and Kesselman, 2004), in which certain tasks may be automatically delegated to systems that are outside the truster's direct control. In this context, there may be a number of competing systems that can fulfil a particular task, each providing a different quality of service. Obviously, it is in the best interest of the truster to delegate in a way that maximises the probability of the task being completed, with the highest possible quality of service.

Human derived Trust — To assess a trustee, a truster usually gathers evidence that supports one or more conclusions about the trustee's likely behaviour. In the preceding examples, this evidence gathering can, at least in part, be automated. For instance, automatic intrusion detection indicates that a particular user account is being used for malicious purposes, or that a service provider may be judged on the quality of service it has provided in the past. In other cases, such as online auction houses like e-bay[1], trust may depend on intangible qualities, only discernible from the subjective experience of a human user.

1.3 The Motivation for this Work

It is clear from the preceding section that trust affects a number of different areas in computing. Potentially, each of these would benefit from automated methods for addressing trust-related issues. However, in the case of the three areas that we have identified above, each has its own distinct set of requirements, and deserves its own specific treatment. Therefore, in this document, we have chosen to focus on just one of these areas — service provision[2]. In particular, we are interested in service provision in contexts in which human assessment of trust is impractical, either because too much data is required to make a judgement, or because too little time is available. To justify this choice, we shall expand on the importance of this area, by using the Grid as a motivating example where it is of prime importance. Specifically, we discuss the central role of service provision in grid systems, and the need to automate decisions involving service provision trust on the Grid.

In Foster et al. (2001), the aim of the Grid is stated as facilitating, *"coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations."* Specifically, the Grid is concerned with direct access to computers, software, data and other resources for multiple purposes that involve collaboration across geographical and institutional boundaries. In this context, a virtual organisation (VO) is the set of individuals or organisations that are involved in such a collaboration. It is envisaged that the resources available to a VO may offer varying degrees of reliability, and may leave and re-join the system at any time. In addition, the organisations that supply these resources could have different, and possibly conflicting, interests. Together, these properties imply an inherent unreliability in Grid resources; this is amplified further when we consider that the Grid is intended to take on global scale.

We can see from this description that a variety of trust issues arise from the Grid. In particular, resource failure should be taken as a rule, not an exception. Thus, we must

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

[2] This focus does not preclude us from making some wider contributions; rather, it concentrates our efforts on a more tractable set of problems.

attempt to minimise the risk of failure by choosing trustworthy resources, and compensate for component failure, when it does it occur, by finding replacements quickly and accurately. To do so, we must be able to accurately assess and compare the trustworthiness of competing providers. Both the requirement for accuracy and the requirement for speed exceed the practical limits of human ability for any sizable system; thus, automation is an essential attribute to any solution to this set of problems.

1.4 Trust as a Multi-Agent System Problem

At the beginning of this chapter, we briefly stated that we are interested in assessing trust between intelligent software agents. Although our motivation is to solve problems associated with systems such as the Grid and the semantic web, it can be useful to frame these systems as multi-agent systems. Here, we discuss the utility of this approach in our domain.

First, we need to be clear about what a software agent is, and what purpose it serves. We discuss the subject of agents in more detail in Section 2.1; for now, it suffices to note the definition of an agent given by Wooldridge and Jennings (1995) as, “a computer system that is situated in some environment, and that is capable of autonomous action in that environment in order to meet its design objectives”.

The concept of an agent provides us with a metaphor for designing complex software systems. We call a system composed of a number of interacting agents (each with potentially different goals and capabilities) a *multi-agent system.*

Agents often need to form themselves into resource-sharing collectives that act in a coordinated manner, effectively these are virtual organisations. Thus, agents and grid computing share some common ground in that they both involve groups of software entities that can operate together as whole. Where the two fields differ is in their focus. Grid computing has traditionally concentrated on underlying infrastructure: tools, protocols, and middleware, which enable secure coordinated resource sharing. Agent-based computing involves algorithms and methodologies for building autonomous problem solvers. Grid systems have, for the most part, been inflexible in terms of the way their composite resources can interact, and agents can provide a means to alleviate this situation (Foster et al., 2004).

From this discussion, the relevance of agency as a metaphor in grid computing is clear. However, this relevance also extends to pervasive computing environments and the semantic web. Much pervasive computing research already makes use of agent technologies; examples are given in Ramchurn et al. (2004). There is also a current trend towards convergence, or at least integration, between semantic web and grid technologies; for instance, recent work on the Grid, such as Tuecke et al. (2003), makes use of standards developed by the semantic web community.

Returning to our own perspective, by considering trust in the context of agents, we can draw upon existing concepts and technologies that have been developed in that field. Moreover, the idea of an agent as an entity with individual goals and capabilities is an appropriate metaphor for exploring issues of trust, particularly when placed in a social context.

1.5 Research Requirements

In Section 1.3, we stated that the focus of our research is service-provision trust, specifically in contexts in which human assessment of trust is not practical, and discussed the importance of this type of trust in large open distributed systems, such as the Grid. Against this background, the overall aim of our research is *to develop a computational model for assessing the trustworthiness of an agent, to provide a particular service.* In this section, we detail a set of requirements that are relevant to this goal, which we refer to subsequently when assessing our own work and other computational trust models.

Essentially, assessing the trust in an agent can be viewed as an on-line learning problem, in which the future behaviour of that agent must be predicted given the available evidence. Depending on the situation, there may be a variety of predictive information sources that can be used to perform this task. These sources may have varying degrees of predictive power, and may or may not be available in a given situation. For instance, if a truster has previously interacted with a trustee before, then the truster can use past observations of the trustee's behaviour to estimate the outcome of a future interaction. However, if a truster has not previously interacted with a trustee, it must rely on other information, such as the behaviour of other similar trustees it has interacted with.

In light of this, the challenge for trust assessment is to identify sources of evidence, fuse this information in a way that respects the relative predictive value of each source, and find mechanisms to cope when any particular source is unavailable. Furthermore, this must be done such that the opportunity for a trustee to outwit the learning process, by behaving in a certain way, is minimised. In the following subsections we expand on this by identifying three key sets of requirements: general requirements, that must be addressed by any solution to this problem; additional requirements, imposed on solutions aimed at large-scale distributed systems; and *reputation* requirements, which apply when information about a trustee is gathered via a third party.

In the last set of requirements, we use the term *reputation* to refer to the overall opinion of a community with regard to a trustee. We differentiate this from the concept of *trust*, which we consider to exist between two entities — a truster and a trustee. In a large open system, the likelihood of a truster having information about a trustee may be much lower than the likelihood of *some* other agent possessing such information. The opinions of other agents about a trustee, which constitute the trustee's reputation, are therefore an important source of evidence for the truster. However, third party opinions are inherently unreliable compared to directly observed evidence, for several reasons that we discuss in Section 1.5.3.

1.5.1 General Requirements

1.1 **Varying degrees of Evidence** As a group of entities interacts over time, the number and type of interactions that occur between group members may change. A trust model should make use of this information, on average increasing the accuracy of its results as the frequency of interactions in the system increases. The system should not, however, be dependent on such information. In particular, the system should be able to operate in the following situations:

1.1.1 The truster has direct experience of the trustee.

1.1.2 The trustee is not known directly by the truster, but is known by other entities within the system.

1.1.3 The trustee is not known to any entity in the system.

1.1.4 No entity in the system is previously aware of any other entity.

Rationale: We should endeavour to make as much use of the information provided by the environment as possible, but should also be able to provide reasonable results when certain sources are not available. This is particularly important, for instance, when a large system has just been initialised and no interactions have yet taken place.

1.2 **Context Dependence** The decision of how much trust to place in an entity should depend on the context in which that decision is being made. In particular, it should be dependent on the following:

1.2.1 The truster's individual preferences.

Rationale: Different agents may have different priorities in relation to service provision, so what one agent considers a good service will not necessarily be the same as another.

1.2.2 The truster's relationship with trustee.

Rationale: An agent's behaviour towards its peers may differ depending on the relationship that holds between them; for example, behaving more favourably to a friend than a foe.

1.2.3 The type of action the trustee is entrusted to do, or refrain from.

Rationale: Our definition of trust relates to a particular action; trust to perform one task does not imply trust to do something different.

1.2.4 Time of assessment

Rationale: In general we cannot assume that an agent's behaviour will remain constant over time. Consequently, the more time that has passed since an observed interaction, the less predictive value it will have for future predictions.

1.3 **Incentive Compatibility** If the proper functioning of a system relies on an agent behaving in expected way, then it should not be in an agent's interest to violate that expectation. This requirement is referred to in the literature as *incentive compatibility* (Jurca and Faltings, 2003).

Rationale: A rational agent will always attempt to further its goals by whatever means possible, even if this means breaking the rules, or otherwise behaving in an anti-social way. If we intend to rely on certain types of behaviour being prevalent in a group of agents, we must provide or identify an explicit reason for which agents will choose to act in the desired way.

1.4 **Identity Changing** If participants in a system are capable of changing their *identity* in an environment in an attempt to increase their reputation, the average utility of other members should not decrease due to such a change. Note that this is a special case of Requirement 1.3.

Rationale: One objective behind assessing the future action of an agent is to identify agents that behave in an unfavourable way. It is unlikely to be in an agent's interest to have such a poor reputation so, if by changing its identity it can start again in a better position, it is likely to do so. Clearly this is not desirable, as it allows deceitful agents to continually take advantage of other agents in a system, without being penalised.

Some, for instance Friedman and Resnick (2001), argue that because of this, the cost of starting again with a new identity should be made so great that it should outweigh any benefit of doing so. In some cases, however, this constraint may be too harsh as it could imply that valuable agents need to belong to a system for significant period before being identified. Delaying recognition of good agents will incur its own costs to the community as a whole.

The underlying metric we wish to maximise is the average utility of participants in a system. Keeping in mind the situation described above, we believe this is a more appropriate requirement, which allows for more pragmatic approaches to the identity changing problem.

1.5 **Trust Representation**

The way in which a trust model represents trust between agents should satisfy the following requirements:

1.5.1 **Representation of Uncertainty** A representation of trust should quantify the uncertainty surrounding a trustee's behaviour. Moreover, it should distinguish between two types of uncertainty: *intrinsic* uncertainty, due to variability in a trustee's behaviour; and *evidential* uncertainty, due to a lack of evidence on the part of the truster, concerning the trustee's behaviour.

Rationale: As stated in Section 1.1, trust is primarily concerned with the uncertainty surrounding a trustee's action, and how this uncertainty affects the decisions made by the truster. It is therefore essential that a trust model somehow captures this uncertainty. Furthermore, it is useful to distinguish between uncertainty that is inherent in a trustee's behaviour and uncertainty that is due to a lack of evidence. Although both may have an impact on how a truster behaves toward a trustee, the latter may prompt the truster to seek more evidence before making further decisions.

1.5.2 **Trust Grounding** Any model of trust should include a clear mechanism through which a truster can derive a trust value for a trustee from relevant information it obtains from its environment; for example, if a truster assesses trust based on observations of a trustee's past behaviour, there should be a clear mechanism to derive a trust value based on these observations.

Rationale: This requirement is a clear prerequisite to our aim of fully automating trust assessment. If such a mechanism did not exist, the appropriate trust value for a trustee could not be determined.

1.5.3 **Minimisation of Arbitrary Model Parameters** The specification of a trust model can be simplified by introducing a set of parameters that should be adapted to environmental conditions. In some cases the optimal values for such parameters may be clear, but in others estimating optimal values may be an intractable problem. Any model of trust should attempt to minimise parameters that fall into the latter category.

Rationale: In many cases, a poor parameter setting may severely affects performance. If estimating good parameter settings becomes a hard problem, it may limit the usefuless of the model.

1.6 **Exploration of Trustee Behaviour** Rather than relying on its knowledge, a truster should, in some circumstances, interact with a little-known agent, so that it may learn about its behaviour.

Rationale: If a truster knows that certain interaction partners provide an acceptable level of service, they might never choose to interact with any other agent that they know less about. Viewed at a system level, this attitude may mean that new agents never get a foothold in the environment, even if they offer a better service than other established agents. In reinforcement learning (Sutton and Barto, 1998), this problem is known as the exploration exploitation trade-off.

1.5.2 Requirements for Large Scale Distribution Systems

2.1 **Group Assessment** The system should be able to assess the trustworthiness of a coalition, based on information about the coalition members; this should include cases in which only a subset of the members are known to the truster.

Rationale: One reason why a virtual organisation may be formed is to pool resources from a number of agents to provide a service which could not otherwise be provided. In such cases, all group members will impact on the quality on the service provided. Since the members of a VO are likely to have existed before the VO was formed, previous interactions with the members are obviously an important source of information.

2.2 **Scalability** The model should be scalable; that is, its use should be practical and useful regardless of the size of the system it is applied to.

Rationale: We envisage that our trust system could be used to enable trust-aware decision-making on the Grid. The current vision of the Grid is of large, geographically distributed system that will grow from the combination of many smaller systems. The number of entities interacting in our target environment may therefore vary by several orders of magnitude.

2.3 **Robustness** The model should be robust in the face of the failure of system components.

Rationale: If the system is to operate successfully in a large distributed environment, we must assume that elements of the system may fail on a regular basis, and take steps to minimise the effect of such failures on the performance of the system as a whole.

1.5.3 Reputation Requirements

3.1 **Assessment of Source Accuracy** When accepting an opinion about a trustee from a reputation source, the truster should judge how likely the opinion is to be accurate, based on its origin. This should determine the impact of the opinion on the truster's overall assessment of the trustee.

Rationale: Sources of reputation will not always be reliable; for example, a close colleague of a trustee is likely to have a strong incentive to exaggerate the trustee's credentials, and hence provide unreliable information.

3.2 **Heterogeneous Preferences** The reputation mechanism should not depend on the assumption that a truster's preferences with regard to service provision are similar to that of its reputation sources.

Rationale: Consider as an example a rock music fan and a classical music fan. Assuming their musical preferences are mutually exclusive, if one asks the other what they think of a particular composition, and the reply given is, "very good", this will not be consistent with the result if the roles were reversed. This is because the reputation source's evaluation is semantically dependent on its own preferences. Any reputation mechanism should account for this problem.

3.3 **Correlated Evidence Problem** Consider two or more reputation sources that have all witnessed the same set of interactions with a trustee, and then update their opinions of the trustee in light of this evidence. Later, an agent wishing to assess the trustee requests the opinion of all the reputation sources and, due to their shared experience, all return similar results. Clearly, the combined information from the reputation providers is smaller than if they had all based the their opinions on separate evidence of similar magnitude. Failure to distinguish between these cases is known as the *correlated evidence problem* (Pearl, 1988) A reputation-sharing mechanism should provide a means to avoid this problem.

3.4 **Witness Propagation** Consider three agents α, β, and γ; α requests reputation information from β and subsequently updates its beliefs. Later, γ requests reputation from α, who gives a reply based on the information it received from β. If the original information given by β was misleading, then the information given to γ by α will be misleading also. This is known as the *false witness propagation* problem, itself an instance of the correlated evidence problem. A reputation system should ensure that this does not occur.

Rationale: If information is allowed to propagate like this through the system, it becomes very hard to control its impact and can result in highly correlated evidence.

1.6 Document Structure

The remainder of the document is structured as follows: Chapter 2 gives a review of relevant literature, and draws comparisons between existing models and our requirements; consequently, we identify shortfalls that warrant our attention. Chapter 3 defines a framework for reasoning about trust; Chapter 4 outlines our current trust model, and provides an empirical evaluation. Finally, Chapter 5 draws our conclusions and details our plans for further work.

Chapter 2

Literature Review

In this chapter, we present a review of the literature relevant to understanding and evaluating our work. We divide our discussion into three parts. First, Section 2.1 puts our motivations in context, by giving an overview of agent systems. Second, Section 2.2 describes some background theoretical work on trust, and evaluates existing models that attempt to solve some of the requirements we set out in Section 1.5. Finally, we conclude the chapter in Section 2.3, summarising the related work to date, and identifying key outstanding issues.

2.1 Multi-Agent Systems

In this section, we revisit the topic of multi-agent systems for two reasons: (1) we need to introduce terminology that will be used subsequently and (2) we need to understand the role that trust plays within a multi-agent system. We begin by expanding on what we mean by the terms, ‘agent’ and ‘multi-agent system’, and then introduce key issues in multi-agent systems and introduce the role of trust.

The term, agent, has been used by many in computing over the years, but there has never been a universally accepted definition. In (Smith et al., 1994) the term is defined as, “A persistent software entity dedicated to a specific task”. In contrast, Russell and Norvig (2003a) emphasise an agent’s awareness of its environment. They take an agent to be, “anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.” Despite this lack of consensus, the view of Wooldridge and Jennings (1995) has been widely adopted. They state the following minimal set of properties necessary for something to be classed as an agent:

autonomy — agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state

social ability — agents interact with other agents (and possibly humans) via some kind of agent-communication language

reactivity — agents perceive their environment (which may be the physical world, a user via a graphical user interface, a collection of other agents, the Internet, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it

pro-activeness — agents do not simply act in response to their environment, but instead are able to exhibit goal-directed behaviour by taking the initiative

Agents provide us with an abstraction metaphor for designing and building complex software systems. In this way, they can be compared to software objects, the main difference being that agents have control over their behaviour, whereas objects, in general, do not. Typically, the objects in a system share the same goal, which is not necessarily the case with agents. A multi-agent system is one that contains more than one agent, and in which agents interact to achieve their goals. The combination of agents into a collective raises a number of questions that are not present when a single agent is considered in isolation — trust can be considered as one of these issues.

2.2 Computation Models of Trust

As stated previously, there are a number of areas for which trust is relevant. Generally, these fall into three major categories: security issues, user-to-user trust and service provision. In the work described here, we are concerned primarily with service provision, and so in this section we focus our attention only on related work that is relevant to this set of problems. Specifically, we divide our discussion into four areas. First, Section 2.2.1 identifies important issues and puts the proceeding discussion in context by reflecting on the cognitive aspects of trust. Second, Section 2.2.2 surveys the major approaches for forming trust based on information directly available to a truster. Third, Section 2.2.3 addresses the problems that arise when trust is based on third party opinions. Finally, Section 2.2.4 outlines an alternative approach to trust assessment: it discusses mechanisms that attempt to enforce trustworthy behaviour by making it in a trustee's best interest to be trustworthy.

2.2.1 Cognitive Viewpoint

If we wish to assess issues of trust, it is important that we are clear about precisely what we are trying measure. We have already made some progress toward this in Chapter 1, by formally defining trust as probability. However, we need to investigate deeper to find

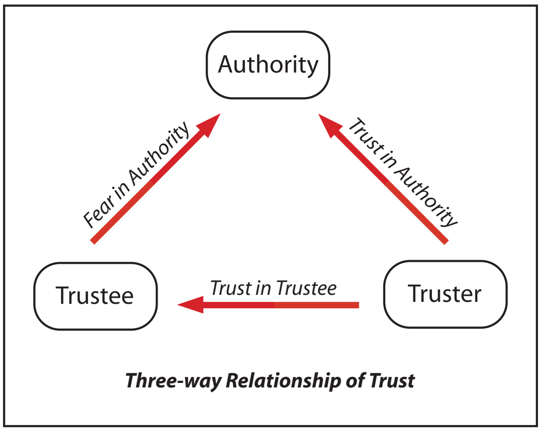

Figure 2.1: The three way relationship between trust, fear and authority

the conditions that must be present for a state of trust to exist between a truster and trustee.

In this respect, we consider the work of Castelfranchi and Falcone (2001), who adopt the same basic definition of trust as us (i.e. a subjective probability of a trustee's action). Building upon this, they make two things explicit: first, they identify beliefs that a truster must hold before it can rationally believe a trustee will carry out a given action; second, they identify a three-way relationship that exists between the concepts of *trust*, *fear* and *authority*.

The core beliefs that are prerequisite to a belief of trust are as follows:

- the truster must believe that the trustee is *willing* to carry out the action;
- the truster must believe that the trustee is *capable* of carrying out the action.

In turn, these beliefs may be conditioned on a number of other beliefs that, for the most part, are domain dependent. In general, however, we can distinguish between two different sets of beliefs: internal beliefs, which relate to the trustee's mental state, and external beliefs, which concern environmental conditions. To illustrate the impact of the latter, consider an entity, 'Captain Joe', who is capable of sailing a particular boat, the 'Jolly Roger'. If something was to happen to the Jolly Roger so as to cause it to sink, then Captain Joe will no longer be able to sail the boat, despite his skills as a sailor.

Figure 2.1 illustrates the relationship between trust, fear and authority mentioned earlier. Fear can be said to be negative trust; it is trust in an entity to carry out an action that has a negative effect on the truster. Like trust, fear in an entity requires the conditions of willingness and capability to be present. When we introduce an authority, which is capable of punishing unsolicited behaviour, an interesting dynamic is set up between the authority, truster, and trustee. The fear a potential wrong-doer has in an authority decreases its likelihood of behaving illegally. On the other hand, if a victim trusts an authority to protect its rights, and it can assume potential criminals hold similar beliefs about the authority, then its trust in potential criminals can be increased.

The influence of authority on trust relationships is also acknowledged in Dasgupta (1988), who argues that, if a rational agent is put in a position where it can choose to benefit at the expense of others, it will always choose to benefit, unless it has reason to fear retribution.

With these factors identified, the question is, how can we use them to develop an automated method for reasoning about trust. An attempt to do this has been made in Falcone et al. (2003), in which they use a fuzzy logic approach (Zadeh, 1975, 1965) to build the beliefs of trust, willingness and capability from other beliefs that are largely domain dependent.

Although identifying the composite beliefs that make up trust gives us a better understanding of what we our attempting to measure, one major question remains unanswered: how can a truster determine its core beliefs based on observations of its environment? Clearly, factors such as trustee willingness and capability are not directly observable in general; they must be estimated from observable evidence. Moreover, depending on evidence that is observable, attempting to estimate separate beliefs about such factors may not be practical at all. For example, consider a scenario in which all we can observe is an agent's external behaviour in the absence of any other environmental data. In this case, the best we can do is quantify the uncertainty in the trustee's behaviour directly; we cannot possibly distinguish between the trustee defaulting on its obligations because it wants to, or because it cannot do otherwise.

2.2.2 Learning from Direct Observations

In this section we turn our attention to methods of representing trust, and how to ground such representations in evidence directly observable to a truster. We differentiate direct evidence from evidence as reported by other agents, the latter of which raises a separate group of problems that we address in Section 2.2.3.

Generally, existing trust models represent trust in one of three ways: (1) they adopt an improvised representation, based on intuitive assumptions about the meaning of trust; (2) they can apply probability theory; or (3) they apply Dempster-Shafer theory. For the purposes of clarity, we separate our discussion according to this categorisation, and discuss each in turn in the subsequent subsections.

Value	*Meaning*	*Description*
-1	Distrust	Completely untrustworthy
0	Ignorance	Cannot make trust-related judgement about entity
1	Minimal	Lowest possible trust
2	Average	Moderate trustworthiness
3	Good	More trustworthy than most entities
4	Complete	Completely trust this entity

Table 2.1: Trust Value Semantics used by Abdul-Rahman *et al.*

2.2.2.1 Improvised Models of Trust

As mentioned earlier, although the concept of trust is prevalent in society, there is some disagreement and confusion about its precise definition. Perhaps partially as a result, a range of different representations have been adopted in existing computational models of trust. In some cases, trust is modeled as belonging to a finite set of qualitative labels, examples of which include the work by Azzedin and Maheswaran (2002c,b,a) and Abdul-Rahman and Hailes (1997). In the case of the former, the trust of one entity in another is a value belonging to the set $\{A, B, C, D, E, F\}$, and similarly in the latter, a member of the set $\{-1, 0, 1, 2, 3, 4\}$. Typically, these values are associated with linguistic labels that describe their intended meaning. For instance, Abdul-Rahman *et al.* attach labels to trust values as described in Table 2.1.

This relatively coarse set of values reflects a perceived difficulty in choosing continuous trust values with any meaningful degree of accuracy. In our view, however, this problem is specific to cases in which trust is elicited from a user[1]. As should become clear from what follows, there are meaningful methods of calculating continuous trust values when trust assessment becomes a fully automated task. We therefore argue that the difficulty in distinguishing between discrete trust levels compared to continuous levels limits the former's applicability to human elicited trust values.

Models that represent trust as a real-valued scalar include those developed by Marsh (1994) and Zacharia et al. (1999). Representative of these, and one that makes a good attempt to deal with many of the requirements outlined in Section 1.5, is the REGRET system (Sabater and Sierra, 2001, 2002), which includes three dimensions of trust: an individual dimension, a social dimension, and an ontological dimension. We shall examine each of these in turn below.

First, we consider the ontological dimension that is essentially concerned with the subjectivity of trust with respect to an individual truster. A trust value in REGRET is represented as a numeric value in the range $[-1, 1]$, with a value of 1 interpreted as *absolute* trust, and -1 interpreted as complete distrust. These values are attached to a particular context by a label; examples of which are `to_overcharge`, meaning that a

[1] To illustrate, consider trying to assess the probability of it raining tomorrow; is it possible to decide whether this probability is more likely to be 2.1 or 2.2?

trustee has a tendency to charge more for a service than the truster believes it is worth, and `quality_swindler`, meaning that, from the perspective of the truster, the trustee tends to supply services with unacceptable quality. The intended interpretation is that they relate to a particular trait of a trustee's behaviour.

An important element of the ontological dimension is that behavioural traits[2] of an agent can be defined in terms of other, lower level traits. For instance, a service provider could be assessed according to a trait labelled `swindler`, which is defined in terms of the traits, `to_overcharge` and `quality_swindler` mentioned earlier. From an implementation perspective, REGRET calculates trust values for such compositional traits as a weighted average of the trust values calculated for the base traits. The weights used in this calculation are considered to be dependent on an individual truster; they encode the agent's subjective definitions for these terms. Besides specifying how compositional traits can have trust values calculated, the ontological dimension can serve a communication role in that a reputation source can share its definition of compositional traits with other agents, so that they can decide how best to interpret reputation information from that provider.

The individual dimension of trust is based solely on the first-hand knowledge that a truster has about a trustee. In REGRET this is calculated based on past interactions that have occurred between the truster and trustee. For example, when a truster purchases a service from a trustee, the truster will have expectations about how the trustee will behave. Some of these expectations will be explicit, based on a contract between the truster and trustee for what the trustee should provide. Others will be implicit, based on the trustee's personal perspective on the world. A truster's individual trust level (with respect to a particular trait) is a function of the difference between the utility the truster would achieve if the trustee behaved according to these expectations and the actual utility gained from the interaction.

As well as providing a method for calculating these trust values, REGRET also provides two separate methods to measure the *reliability* of these values. Two different types of uncertainty determine the reliability of a trust value:

intrinsic uncertainty in trustee behaviour , which is estimated based on the variance of observed trustee behaviour.

uncertainty due to lack of evidence , for which REGRET uses a function that decreases logarithmically until a minimum value, in line with the number of observed interactions with a trustee and the time that has passed since those interactions.

In REGRET, both of these are measured using improvised functions. For example, REGRET calculates evidential uncertainty using Equation 2.1 (adapted from Sabater

[2] In REGRET, these are referred to as 'reputation types'. We use the term trait, so as not to confuse it with the concept of reputation as a collective opinion of a group with regards a particular entity.

and Sierra (2001)). Here, *noObs* is the number of observations a truster has made of a trustee's behaviour, and *itm* is a threshold number of observations, above which the truster considers its knowledge of a trustee to be completely reliable. In a similar way, intrinsic uncertainty is measured by another function, improvised from intuitive conclusions about what factors should affect its value. An overall reliability factor is then calculated as a weighted average of these two functions. One problem with this scheme is that it is unclear what value should be chosen for *itm*, and what weight should be used to generate the overall uncertainty value (counter to Requirement 1.5.3).

$$\text{evidential uncertainty} = \begin{cases} sin\left(\frac{\pi}{2 \cdot itm} \cdot noObs\right) & noObs \in [0, itm] \\ 1 & \text{otherwise} \end{cases} \tag{2.1}$$

Often, an agent will need to assess its trust in an entity with which it has little or no previous experience with. In this case, REGRET can draw upon the social dimension of trust, and there are three sources of information that fall under this heading: witness reputation, neighbourhood reputation, and system reputation.

Witness reputation is, as the name suggests, based on the opinions of third parties concerning a trustee. The influence of particular witness's opinion on the overall trust value is partly determined by the truster's trust in the reputation source to provide reliable information. This can be calculated by applying the formulae for individual trust — effectively, treating the ability to give reliable reputation information as just another trait.

Neighbourhood reputation assumes that the truster maintains a *sociogram*, which is a network structure describing the social relationships between agents in the environment. To calculate neighbourhood reputation, REGRET applies a set of static[3] fuzzy rules, where the antecedent of each rule is a condition on the relationships connecting the trustee to other agents. To illustrate, we might define a rule such as

$$\text{IF } coop(trustee, agent\_b) = high \text{ THEN } socialTrust = very\_bad,$$

where *high*, and *very_bad* are predefined fuzzy sets.

System reputation is also calculated according to a static set of fuzzy rules. In this case, the rules are defined according to the role that a trustee plays within an institutional structure — *seller* is an example of such a role. As with neighbourhood trust, system trust assumes that information about social roles is available to the truster.

[3] By static, we mean that REGRET must be preconfigured with a set of rules. REGRET cannot learn these rules for itself.

REGRET combines these different sources — individual and social dimensions — based on reliability functions defined over them. In addition, there is also an intrinsic preference ordering built in: direct interactions are intrinsically more reliable than witness reputation, witness reputation is more reliable than neighbourhood reputation, and neighbourhood reputation is more reliable than system reputation.

From our perspective, REGRET is significant because it satisfies a broad number of the requirements we described in Section 1.5. However, REGRET suffers limitations for at least two reasons. First, it assumes that certain information (for example, a sociogram) is available (Requirement 1.1). Second, there are several parameters in the model, optimal values for which are not known, and may be domain dependent (Requirement 1.5.3).

2.2.2.2 Probabilistic Models of Trust

Aside from fuzzy logic, the trust models we have looked at so far all make use of, essentially, hand-crafted representations of trust, and operations defined on these representations. This is by no means an invalid approach — in the end — the goal of assessing trust is to provide discriminatory information about trustees. If a solution works, then the approach is reasonable; how it is achieved is of lesser importance. However, there are existing formalisms for reasoning about uncertainty, which have well known beneficial properties, and are well grounded in mathematical theory. Of these, perhaps the most prominent is probability theory.

One noteworthy probabilistic trust model is detailed in Barber and Kim (2001). This provides a well grounded method for assessing the reliability of information sources, and shows how this can be used to combine conflicting information into a consist knowledge base. Unfortunately, the model is designed specifically to deal with such conflicts: it uses the statistical properties of the conflicts themselves to perform its task, and so cannot be applied to a more general setting.

More generally, probabilistic models that attempt to assess trustees on a broad range of services (which include those reviewed in the remainder of this section) have two things in common: First, they represent the outcome of an interaction with a trustee as a bistable event — either the trustee cooperates and fulfils its obligations to the truster, or it defects and does not. Second, they estimate the probability distribution for this binary variable based on direct or indirect (via reputation) observations of the trustee's past behaviour. Obviously, this simplification leaves clear room for improvement: if a truster's utility is dependent, not only on if a trustee performs a task, but also on how well the task is performed, then a bistable representation will fail to capture the relevant dynamics of the problem. Nevertheless, situations in which task performance does not carry much significance over and above task completion constitute an important subcase.

	Agent A	Agent B
successful	20	2
unsuccessful	20	2

Table 2.2: Frequencies of Successful and Unsuccessful Interactions with different Agents.

An example of such a system can be found in Wang and Vassileva (2003). Here, a trust mechanism is presented for use in a peer-to-peer file sharing environment. Trust in a particular provider is assessed according to several quality attributes, such as type of file requested, download speed and file quality. The system uses a naïve bayesian network, in which the probability of the provider being trustworthy (modeled as a binary variable) is dependent on each of the quality attributes considered. Here, *naïve* means that the effect of each attribute on the trustworthiness of a provider is assumed to be independent. Such assumptions are often made to simplify a problem domain, with solutions adopting them generally being robust when faced with minor violations. Whether the assumption is reasonable in the domain targeted by this model depends on the particular set of quality attributes used in a given instance of the model.

One factor which Wang and associates fail to account for is *evidential uncertainty*. Here, we differentiate evidential uncertainty from intrinsic uncertainty. We define intrinsic uncertainty to be uncertainty that is due to inherit unpredictability of a stochastic process. On the otherhand, we consider evidential uncertainty as uncertainty that is due to a lack of knowledge. To illustrate, consider observing successful and unsuccessful interactions with two agents, A and B, the frequencies for which are shown in Table 2.2. Using Wang's model, we would consider there to be no difference in the uncertainty surrounding agent A's behaviour and agent B's behaviour. However, intuition tells us that this is not the case, because we have interacted with A ten times more than B and can therefore be more certain about A's true behaviour. This highlights a failing common to all simple probabilities that is particularly important in domains where the frequency of observations is relatively low. We believe that trust assessment in large multi-agent systems is such a domain, because the likelihood of any two agents interacting a large number of times is fairly low. We therefore argue that accounting for both types of uncertainty is important and give further justification for this in Chapter 3.

Fortunately, to say that probability theory in general cannot account for evidential uncertainty would be incorrect. This is illustrated by the trust model presented by Jøsang and Ismail (2002), in which trust is modelled as a probability distribution for a binary event, a class of distributions commonly referred to as *Bernoulli* distributions. In addition however, they also model the *parameter* distribution of the Bernoulli distribution (DeGroot and Schervish, 2002a). Statistical models, such as Bernoulli distributions, are characterised by a set of parameters that determine their shape. In the case of a Bernoulli distribution, it is characterised by a single parameter — the probability of the variable being equal to one. The parameter distribution in this case, is the distribution over the possible values of that probability.

For simplicity, the authors choose to represent the parameter distribution as a Beta distribution. The advantage of this is that there is a special relationship between Bernoulli and Beta distributions. Specifically, consider a Bernoulli distribution for which the prior parameter distribution is a Beta Distribution. If we draw samples from this Bernoulli distribution under an i.i.d assumption[4], then the posterior parameter distribution, given the samples, will also be a Beta distribution. A family of distributions which exhibits this property for a statistical model is known as the model's *conjugate* family.

Effectively, the parameter distribution represents the evidential uncertainty surrounding the true intrinsic probability of a random variable; in this case, the intrinsic probability that a trustee will cooperate rather than defect. It can be used to reason about how much evidence is required to make a particular decision, or to compare the confidence levels different agents have in their knowledge about a trustee. Again, we discuss this further in Chapter 3. Moreover, by choosing a conjugate prior, the authors simplify the process of calculating, combining, and storing the parameter distribution associated with a trustee. For this reason Beta distributions are also applied to the field of trust by Mui et al. (2001) and Buchegger and Boudec (2003).

2.2.2.3 Dempster-Shafer Models of Trust

An alternative method for handling uncertainty can be found in Dempster-Shafer theory (Shafer, 1976). Dempster-Shafer provides a mechanism for forming degrees of belief about sets of hypotheses, based on available evidence. For example, imagine we have a set of two competing hypotheses $\{A, B\}$, of which only one can be true. Dempster-Shafer theory divides the total belief in the set between the elements of its superset[5], $\{\{A\}, \{B\}, \{A, B\}\}$. Essentially, belief in the set $\{A\}$ (and likewise for set $\{B\}$) represents the evidence supporting A as the true hypothesis. On the otherhand, belief in the set $\{A, B\}$ is belief that cannot be divided between A and B. This can be said to represent the evidential uncertainty surrounding A and B; because of this, Dempster-Shafer theory is often claimed as a solution to the inability of simple probabilities to capture this notion.

In particular, this is the rationale given for its use in Yu and Singh (2002). Here, the authors define a binary hypothesis set, in which the competing hypotheses are that an agent is trustworthy, and that it is not trustworthy. They consider scenarios in which trustees supply services, which are given a quality of service rating between 0

[4] This is a standard abbreviation for the assumption that samples are drawn independently from an identical distribution.

[5] By definition, the superset of a set S, is the set comprised of all subsets of S.

and 1. To gather evidence for the trustworthiness of an agent, they perform the following three steps. First, they break the range of quality values into three intervals, $[0 <= x < a], [a <= x < b], [b <= x <= 1]$, where a and b are arbitrary constants. Second, they count the proportion of recent[6] trustee interaction outcomes that fall in each of these three intervals. Finally, they take the proportion of outcomes in the lower interval as the belief that the trustee is untrustworthy, the proportion in the higher interval as the belief that the truster is trustworthy, and the proportion in the middle interval as the belief in the total set, {untrustworthy, trustworthy}. In line with Dempster-Shafer theory, belief in the total set is interpreted as the degree of uncertainty in whether the trustee is trustworthy or not.

The problem with this approach is twofold. First, there is no clear way to choose the values for the constants a and b, which violates Requirement 1.5.3. Second, the notion of trust that this representation captures is somewhat artificial. Consider as an example a trustee with whom a truster has (recently) interacted 1000 times. On each of these occasions, the trustee's quality of service was precisely 0.5. Here, the truster has chosen $a = 0.3$ and $b = 0.7$. According to Yu and Singh's model, this means that the truster is completely uncertain whether it trusts the trustee or not. Intuitively, this is not the concept of trust we want, since it violates Requirement 1.5.1. A more useful conclusion would be that the trustee provides an average quality of service of 0.5, with very low variance, so that there is a large degree of certainty regarding its behaviour. Furthermore, the rationale for using Dempster-Shafer, rather than probability, is somewhat unsound: We have already seen that probability theory can be made to account for evidential uncertainty in Section 2.2.2.2. In our view, Dempster-Shafer's main strength comes when the available evidence truly does support multiple hypotheses.

An alternative application of Dempster-Shafer, relevant to trust, is given by Jøsang (2002, 2001). Here, it is extended to form a logic for reasoning about uncertain probabilities, known as *subject logic.* This has grounding in both probability theory and Dempster-Shafer theory, and has propositional calculus as a special case. Significantly, from the perspective of trust, it defines two new operators for reasoning about third party opinions: the *consensus* operator and the *discounting* operator. In particular, they can be used to combine opinions from different sources, as is required when trust is based on the opinions of others (Section 2.2.3). The consensus operator is used to combine opinions from different sources when each source is equally and fully trusted to provide accurate information. The discounting operator plays a supportive role to the consensus operator: It is applied prior to the consensus operator, to any sources which are not fully trusted to provided accurate information. Its effect is to increase the evidential uncertainty surrounding the opinion. As a result it decreases the effect it would otherwise have when combined with other opinions.

[6] In their model, Yu and Singh only use the x most recent observations of a trustee's behaviour. This allows for the possibility that a trustee's behaviour changes over time, in which case old observations would be poor predictors of behaviour.

The justification for these two operates, is grounded in statistical theory. Specifically, a mapping is provided between the Dempster-Shafer notion of evidential uncertainty, and the beta distribution representation described in Section 2.2.2.2. The operators are thus shown to be equivalent to operations on the parameter distribution. In the case of the consensus operator, the grounding relation first assumes that each opinion concerns a Bernoulli distribution, and that they are each based on disjoint sets of samples from that distribution, under an i.i.d assumption. The result of the consensus operator is then shown to be equivalent to the probability that would result, if all the data are considered together. Although the assumptions behind this grounding are not expected to hold in general, it is expected to give reasonable results, even when they do not hold. The discounting operator is given a similar justification, which we do not describe in full here. Briefly, under certain conditions, it is shown to be equivalent to multiplying an opinion by the probability that it is true.

Overall, subjective logic provides a promising method for reasoning about uncertain probabilities. In particular, its grounding in probability theory gives a good justification for its use. There are, however, there are two points that must be brought to mind. First, the discounting operator does not say how the probability that a source is accurate should be calculated. This is an open question that may depend on the type of information available. Second, the definition is subjective, particularly in the case of its consensus and discounting operators, which make certain assumptions that may not be appropriate in every case. These should be questioned with respect to any application subjective logic is considered for.

2.2.3 Learning from Others

The basic problem of trust assessment is to estimate the behaviour of a trustee based on the available evidence. When this evidence is gathered indirectly via third party opinions (reputation), there are four additional factors that we must consider:

1. A third party may define observed properties in a different way from the truster. For instance, what one agent considers a good service, may not be what another considers good (Requirements 3.1 & 3.2).

2. Reports from several different reputation sources may be based on the same evidence, resulting in the correlated evidence problem (Requirement 3.3).

3. The behaviour of a trustee towards a third party may be different from its behaviour towards the truster (Requirement 3.1).

4. A reputation source may have no incentive to provide reputation or, if it does, it may have an incentive to misrepresent its knowledge about a trustee. We can subdivide the latter into positive discrimination (collusion) in which the reputation

source over estimates the beneficial qualities of a trustee, and negative discrimination in which the trustee's beneficial qualities are under estimated (Requirements 1.3 & 3.1).

Each of these factors manifests itself as a decrease in the predictive power of reputation — in other words noise — when compared to direct evidence. Thus, many trust models which employ reputation include noise reduction mechanisms to target one or more of these factors. Essentially, there are two basic method for detecting reputation noise: (1) endogenous methods, which attempt to identify noise from the statistical properties of the opinions expressed about a trustee; (2) exogenous methods, which use information other that the statistical properties of reputation[7]. The issue of incentives raised above has wider consequences for trust which we discuss further in Section 2.2.4. Here, we consider examples of exogenous and endogenous techniques that attempt to handle the above factors.

2.2.3.1 Exogenous Techniques

In Zacharia et al. (1999) two complementary reputation systems are introduced, called HISTOS and SPORAS. SPORAS is a simple trust model that is not context dependent, and that a truster can use when there is little information available about the other agents. To account for the unreliability of a reputation source, it simply weights its opinion by the truster's trust in the reputation source itself. The implicit assumption here is that if an agent can be trusted in general — for example, to provide a particular service — then it can be trusted to provide accurate information about other agents. Clearly, this assumption does not hold in general.

HISTOS on the otherhand, is a more sophisticated model suited to environments in which more information about a trustee's peers are available. It suffers from the same context independence as SPORAS, but takes on board the social relationships that exist between the truster, its reputations sources, and the trustee. Specifically, it (like SPORAS) treats trust as a transitive relationship, in which the trust of a truster in a trustee is a function of the trust of each reputation source in the trustee, and the trust of the truster in each reputation source. Unlike SPORAS, HISTOS builds a social network from the pairwise ratings that have previously been reported between agents. This is a directed graph, in which agents are represented by nodes, and edges between nodes represent the direct trust value of the parent node in the child node. The transitive trust relationship is then applied recursively along the directed paths between truster and trustee.

[7] In Whitby et al. (2004) methods that use the trust of an agent in its reputation sources are considered exogenous. Here, we prefer to consider these as endogenous, if trust is based on the accuracy of past opinions expressed by reputation sources.

The REGRET system (Section 2.2.2.1) applies two different exogenous techniques to reputation noise reduction. The first of these applies the same transitive reasoning to trust as HISTOS and SPORAS, weighting a reputation source's opinion by the trust the truster has in that reputation source. However, REGRET's notion of trust is more expressive: it takes on board contextual issues such as the time a rating was given, and through its ontological dimension, can account for several different aspects of trust. For example, the trust of an agent as a reputation source may be built on its trust as a service provider, and the accuracy of any past opinions it has provided. Unfortunately, REGRET does not give specific guidance on the relative importance of such factors, nor how the accuracy of past opinions should be calculated, the latter of which, in itself, is not a trivial issue.

The second mechanism adopted by REGRET, specifically attempts to deal with the correlated evidence problem. The majority of trust models recognise witness propagation (Requirement 3.4) as a potentially severe source of correlation. The universal solution is to specify that a reputation source should only share its direct knowledge, and not pass on other agent's knowledge as its own. Further than that, the solutions offered by the literature vary. Most models assume independence, which can be justified if intersection between agent's world views are small. REGRET's solution is to carefully select reputation sources based on social network analysis. To do this, it uses an algorithm that divides a social network into groups of agents and then chooses reputation sources which are representative of those groups. The intuition is that a highly connected group of agents are likely to share the same knowledge, whereas loosely connected individuals are unlikely to share knowledge.

2.2.3.2 Endogenous Techniques

Of the endogenous techniques that exist, there are two basic approaches: first, we can estimate a reputation sources reliability by evaluating the accuracy of any past opinions it has expressed; or second, we can assume that the majority of opinions received about a trustee are representative of its behaviour. Examples of the latter include Whitby et al. (2004) and Dellarocas (2000).

Whitby and associates extend the Beta Reputation System (Section 2.2.2.2) by applying an iterative filtering algorithm. In each cycle, an interquantile range[8] is calculated for the set of opinions received about a trustee. Any opinions lying outside this range — that is, opinions that deviate significantly from the mean — are discarded. In the following cycle, the interquantile range is recalculated without the discarded opinions; the process continues until all remaining opinions are in range. Although this approach is reasonable, and has been shown to give encouraging results, there is no guarantee that

[8] The interquantile range of a dataset is a descriptive statistic that specifies a range of values in which a given percentage of the data lie.

any opinions will remain after the algorithm has been applied. This can occur when all opinions differ significantly from the mean. Therefore, the approach is only applicable when there is a clear consensus between a reasonable number of reputation sources.

Dellarocas adopts a slightly different approach. First, they attempt to prevent negative discrimination by controlled anonymity, by which reputation and services are distributed by a central institution that does not reveal the identity of producers or consumers to each other. The intuition here is that, because a reputation source does not know the true identity of the trustee, it cannot determine if it is friend or foe and so has no reason to discredit it. This approach does not account for positive discrimination because if a trustee and a reputation source collude, they could signal their identity to each other by other means, breaking anonymity.

To deal with this, the authors apply a clustering algorithm to separate a trustee's reputation into an upper and lower group of opinions. Since positive discrimination should appear more complementary of the trustee, such opinions are assumed to be in the upper cluster, which is discarded. In most cases discarding the upper cluster introduces a negative bias. Through empirical study, the authors argue that this bias is within acceptable bounds.

In our view, the main limitations of the Dellarocas approach lie in the applicability and effectiveness of controlled anonymity. Obviously, there are many cases in which a provider and consumer must be aware of each others identity for a transaction to take place, which limits the situations in which this can be applied. Where it can be applied, it cannot account for a reputation source who wishes to discredit all trustees other than itself, or assumes that any agent that does not signal its true identity is a foe.

More generally, however, the assumption that majority opinion is reliable does not hold, when there is a trustee with whom no agent has significant experience. In this case all benevolent reputation sources will report no information, while reputation sources with an incentive to lie, will report information. In light of this, most, if not all, of the reputation provided will be unreliable.

To alleviate these problems, we can consider the alternative endogenous approach, of assessing a reputation source based on the accuracy of its past opinions, as adopted by Yu and Singh (2003), who extend their previous work (Section 2.2.2.3) by applying a modified version of the Weighted Majority Algorithm (Littlestone and Warmuth, 1994). Essentially, their approach consists of three steps: first, the reputation of a trustee is calculated as a weighted average of reputation source opinions, with initially equal weights; second, after the result of an interaction with the trustee has been observed, the differences between each opinion and the observed result are calculated; third, the weights applied to each reputation source are adjusted relative to the difference between their stated opinion and the observed result.

There are two main advantages of this approach: First, under the reasonable assumption that a reputation sources past and future accuracy are correlated, the relative weight placed in inaccurate reputation sources will gradually decrease towards zero. Second, unlike Dellarocas and Whitby, this approach does not require the majority of reputation opinions to be accurate, and so does not suffer the consequences associated with that assumption.

2.2.4 Mechanism Design

The techniques described so far have all addressed trust by attempting to assess trust based on available knowledge. An alternative approach, *mechanism design* (Dash et al., 2003), is to design a system in such a way that it is in the best interest of the agents to behave favourably towards each other. An established research area in its own right, this is not always explicitly tied to issues of trust, but from a trust perspective, it reduces the uncertainty surrounding a trustee's willingness to behave well. However, uncertainty in trustee behaviour cannot be removed completely. Generally, to manipulate a trustee's interests, we must assume that it is rational, which may not be the case for a variety of reasons, not least that an agent may have contracted a virus. Also, affecting an agent's willingness does not affect the uncertainty surrounding its capabilities. In light of this, we view mechanism design as complementary to trust assessment, rather than a replacement for it. Here, we illustrate how it can be used to simplify trust assessment problems, by reviewing some of the methods that lie in the intersection between trust and mechanism design.

Many of the trust models considered above include recommendations that can be considered as mechanism design. For instance, in HISTOS and SPORAS, it is not be possible for an agent to have a reputation value lower than that of a new unknown agent entering the system for the first time. If this were possible, and agents were able to change identity at no cost, then agents with low reputation would simply create a new identity to improve their standing. Unfortunately, this approach may lead agents never to trust new agents if measures are not taken to ensure otherwise[9].

An important problem not addressed above, is the incentive that an agent has to act as a reputation source. Clearly, if agents share information about trustee behaviour, they can increase their combined expected utility. However, this is not sufficient to persuade individual agents not to freeload, taking advantage of any available information, while not sharing any of their own. Jurca and Faltings attempt to alleviate this problem by introducing side payments for reputation (Jurca and Faltings, 2003). This obviously provides an incentive for an agent to supply reputation information. However, it does

[9] Perhaps an alternative approach could be to charge agents a penalty if they choose to leave the system with a reputation value less than that of a new agent.

not distinguish between accurate and inaccurate reputation. To rectify this, they suggest following conditions should be guaranteed.

- Agents that report truthfully the result of every interaction with another agent, should not lose utility.
- Agents that report reputation incorrectly should gradually lose utility.

To ensure these conditions, Jurca and Faltings suggest that agents should only be paid for their opinion if it matches the next opinion received about the same trustee, from a different source. Unfortunately, this approach fails if most agents provide false information, if agents collude to provide matching false reports, or if agents hold multiple identities to outwit the truster.

A more robust solution is provided by Dash et al. (2004), who introduce the concept of *trust-based mechanism design*, which attempts to explicitly handle issues of trust through mechanism design. In their approach, suppliers are allocated to consumers by a central institution (henceforth referred to as the centre). To aid the centre in making a good allocation, the consumer informs the centre of its preferences with regards the allocation and all the information it currently knows about potential suppliers. Furthermore, the consumer either receives or makes a payment to the institution based on the effect its information has on the overall utility of the system. Based on these two components, it can be shown that it is in the best interest of a consumer to provide its reputation information fully and accurately.

One notable exception, however, is the possibility of agents colluding under certain conditions. A key premise is that agents will truthfully reveal their utility functions for an allocation, because to do otherwise risks decreasing the agent's utility in the allocation. This does not preclude the agent from omitting preferences it may hold that do not effect its allocation. For instance, suppose an agent wishes to decrease or increase the chances of another agent receiving a good allocation, and that it may further this goal by reporting inaccurate reputation. Provided the effect of this inaccuracy does not affect its own allocation, then it may do so without penalty, and for this reason, the approach only satisfies Requirement 1.3 when all an agent's preferences concern its own allocation.

2.3 Discussion

In this chapter, we have addressed four key points. First, we set the scene by giving an overview of multi-agent system research. Second, we considered methods for representing trust, and assessing trust based on evidence directly available to the truster. Third, we reviewed mechanisms for taking account of third party opinions, bearing in mind the extra challenges this source of trust imposes. Finally, we described the complementary role of mechanism design with regards trust assessment; that is, how it can simplify the trust assessment problem, by reducing the uncertainty in a trustee's behaviour *a priori.* In this section, we summarise the main points made throughout the chapter, and identify key challenges for future research into trust assessment.

In Section 2.2.1 we considered work that concentrates on the cognitive aspects of trust — the core beliefs that a truster must hold to rationally be in a state of trust with a trustee. The main contribution of this work, is that is helps to better understand the nature of trust, and the factors that contribute to it. However, it is not always clear how these core beliefs can be elicited from a truster's environment.

In contrast to this, the REGRET system demonstrates how a wide range of evidence can be brought together and used to assess trust in a given context. These sources include previous interactions with a trustee, third party opinions, information about other agents in the same group as the trustee, the relationship between the truster and the trustee, and general assumptions about trustee behaviour. REGRET thus gives a reasonable assessment of a trustee, both when there is a significant amount of information available, and when information is scarce. The main disadvantage of REGRET is that it is based on ad-hoc formulae, which require many parameter settings with no obvious optimal values.

The two main alternatives to ad-hoc formulae, as found in REGRET, include Dempster Shafer theory, and probability theory. An example application of Dempster-Shafer theory to trust is given by Yu and Singh who show how Dempster Shafer theory can be used to assess trust, based on previously observed interactions with a trustee. Although their method is sound in general, the way in which they ground trust in observed interactions is somewhat arbitrary.

Of the probabilistic trust models, the majority represent trust as the probability of a binary event; that is, the probability that a trustee will cooperate or defect. These models generally provide a sound statistical basis for calculating trust biased on available evidence, and offer an attractive alternative to ad-hoc formulae for trust assessment. However, by modelling a trustee's possible actions simply as co-operation or defection, they ignore the effect that quality of service provided by a trustee may have on a trustee. In addition, they remain dependent on (direct or indirect) observations of the trustee's own behaviour, and do not consider other sources of information, such as those explored by REGRET.

The vast majority of these trust models rely on third party opinions. Using such opinions, however, imposes several addition concerns that do not arise from knowledge directly available to the truster, because a third party's own preferences and world view inflict a bias on its opinions. To deal with this, there are two types of approaches: exogenous approaches, which is based on the statistical properties of the opinions received by a truster; and endogenous techniques, which make use of other information available to the truster.

Examples of endogenous techniques can be found in REGRET, which uses information about social relationships between the truster, its reputation sources, and the trustee, to assess the reliability of a given source. Exogenous techniques can be divided into two sets, according to the assumptions they make. First, we can assume that the majority of opinions received about an agent will be representative of its behaviour. Models that adopt this approach can run into problems in several scenarios; for example, if no agents have knowledge about a trustee, the only agents that will report information about the trustee will be those that have an incentive to lie. Second, we can assume that the accuracy of a given reputation source will be correlated with its previous accuracy (as done by Yu and Singh). This is an altogether more reasonable assumption, particularly if we account for the length of time that has past between opinions from a given source[10]. Significantly, if a reputation source attempts to mislead a truster, it will decrease its opportunity to do so in the future.

Against this background, we identify the following key challenges that we believe should be addressed by future research. We consider each of these in turn, in the proceeding Chapters.

Assessing reputation source accuracy — The approach taken by Yu and Singh of judging reputation sources by the perceived accuracy, we believe, is promising. However, some other elements of their model, particularly how they ground trust in observed opinions, leave room for improvement. Therefore, in our view, the general principle of accuracy assessment warrants further investigation, and so we adopt it as part of our own mechanism for inaccuracy filtering (see Sections 3.2, 3.3 and 4.1.3).

Combining different types of evidence — The multitude of information sources employed by REGRET give this model great flexibility. However, its ad-hoc formulae leave a lot to be desired. We believe that further work is needed to investigate the combination of heterogeneous types of evidence in a more principled manner. We discuss one possible method for this in Section 5.2, in which information based on observations of a trustee's behaviour is combined with information about other agents similar to the trustee.

Exploring trustee behaviour — None of the models that we have reviewed have considered the plight of new service providers entering an already well established system. Such service providers may never be given the opportunity to have their trustworthiness assessed, since existing agents may choose to stick with

[10] By accounting for the length of time between opinions received from a reputation source, we allow for the possibility that its mean accuracy may have changed during the intermitting period.

the providers they already know (Requirement 1.6). This type of problem is addressed by reinforcement-learning, but we believe there are interesting problems when trust and exploration are considered together. We discuss this further in Section 5.2.

Chapter 3

A Probabilistic Framework for Modelling Trust & Reputation

One of the key questions we identify in the previous chapter, is how to assess the accuracy of a reputation source, so that we can determine the influence it should have on a truster's judgement of a trustee. How we address this problem is intimately tied to how the precise semantics of a reputation source's opinion; that is, to assess the accuracy of a statement, we must be clear about its exact meaning. Meaning and accuracy can therefore not be considered in isolation, and any means of assessing the accuracy of an opinion must do so with reference to its meaning. With this in mind, we present in the current chapter a set of guidelines for assessing reputation source accuracy with reference to one particular representation of trust. Specifically, we have chosen to represent trust using a set of probability distributions relating to a trustee's behaviour, similar to the beta Reputation System (Section 2.2.2.2). We have chosen to frame trust in this way, because we believe it provides unambiguous semantics, and that it captures the important aspects of trust (Requirement 1.5).

Furthermore, we believe that the framework discussed in this chapter provides an improvement of the Beta Reputation System, because it suggests how to assess trust in an agent, when its actions are not treated as binary events. This does not mean that binary action spaces do not present an important problem. For instance, in many cases, a truster may not care how a trustee fulfils its obligations, but only that it does so. Thus, it may be appropriate to record only that a trustee fulfils its obligations, or that it does not. In Chapter 4 we show how the framework presented in this chapter can be applied to such cases; however, in Chapter 5, we discuss the application of the framework beyond this case.

The current chapter consists of three parts. First, Section 3.1 outlines how a truster can assess and represent its trust in an agent, based on direct observations of its past behaviour. Second, Section 3.2 shows how, based on this representation, the knowledge that different agents hold about a trustee can be communicated as reputation. Finally, Section 3.3, describes the process of assessing reputation source accuracy, and thus how to adjust the effect that an third party opinion has on a truster's judgement of a trustee. Underpinning the discussion in each of these sections is some basic basic statistical terminology, which is detailed in Appendix A.

3.1 Trust Assessment based on Direct Observations

Before we can discuss our basic approach, we must give a formal definition of our problem. In a MAS consisting of n agents, we denote the set of all agents as $\{a_1, a_2, ..., a_n\} = \mathcal{A}$. Over time, distinct pairs of agents may interact with one another. Each such interaction consists of a truster, $a_{tr} \in \mathcal{A}$; a trustee, $a_{te} \in \mathcal{A}$; and a *context*, $\mathcal{C}$. The context $\mathcal{C}$ specifies state information that is relevant to the outcome of the interaction; for instance, $\mathcal{C}$ could include the type of service that a_{te} is requested to provide, and specify terms of agreement that define acceptable quality of service (QoS). In this context, we refer to quality of service as the measurable aspects of a service which affect its desirability; for example, frame rate could be considered a QoS measure for a video stream.

During an interaction, a_{te} is obliged to provide a service to a_{tr}. The actions that a_{te} takes during an interaction, determines the QoS characteristics of the service provided by a_{te}, and ultimately the reward that a_{tr} receives for the interaction. For the purposes of trust, a_{tr} therefore monitors and records QoS information during service provision. Specifically, let $\mathcal{S}_\mathcal{C}$ be the set of possible actions that a_{te} can take during an interaction, in context $\mathcal{C}$; for instance, if the requested service is to provide movie content, then $\mathcal{S}_\mathcal{C}$ would be the set of all possible data streams. Agent a_{tr} monitors service provision by using a function $Q(a_{tr}, \mathcal{C}, \mathcal{S}_\mathcal{C})$ that maps $\mathcal{S}_\mathcal{C}$ onto a set of QoS measurements, denoted $\mathcal{O}^\mathcal{C}$. The QoS measurement taken by a_{tr} for an interaction with a_{te} at time t in context $\mathcal{C}$ is denoted as $O^t_{a_{tr},a_{te}} \in \mathcal{O}^\mathcal{C}$; in some cases, we may omit the time superscript if the interaction time is irrelevant to the discussion. The set of all observations made by a_{tr} of a_{te}, between times t and $t+n$, is denoted $O^{t:t+n}_{a_{tr},a_{te}}$. Furthermore, we define time to be positive, and denote the current time as t'; in this way, all outcomes observed between a truster and trustee up until the current time are denoted $O^{0:t'}_{a_{tr},a_{te}}$.

Against this background, we model the behaviour of a_{te} towards a_{tr} in context $\mathcal{C}$, as a probability distribution, $b(x|\theta_{a_{tr},a_{te}})$ where $x \in \mathcal{O}^\mathcal{C}$ and $\theta_{a_{tr},a_{te}} \in \Theta^\mathcal{C}$. Here, $\theta_{a_{tr},a_{te}}$ is a parameter vector which specifies the shape of the distribution and $\Theta^\mathcal{C}$ is the parameter space, for context $\mathcal{C}$. Essentially, $b(x|\theta_{a_{tr},a_{te}})$ characterises the intrinsic probability with which a_{te} chooses its actions. In general, a_{tr} cannot determine the true value of the parameter $\theta_{a_{tr},a_{te}}$, because it does not necessarily have complete information about the trustee's intentions and capabilities; instead, we estimate $\theta_{a_{tr},a_{te}}$ by Bayesian Analysis.

In Bayesian analysis, we estimate the true value of a parameter, such as $\theta_{a_{tr},a_{te}}$ by first estimating its distribution, which we call the *parameter distribution.* Essentially, the parameter distribution tells us about the evidential uncertainty surrounding a_{te}'s behaviour: if we know little about a_{te}, then the parameter distribution will be close to uniform; on the other hand, if we know much about a_{te}, then the parameter distribution may be highly peaked around one possible value for $\theta_{a_{tr},a_{te}}$.

We assess the parameter distribution in two cases: first, we define $\theta_{a_{tr},a_{te}}$'s *prior* distribution, $d(\theta_{a_{tr},a_{te}})$, which summarises the assumptions that a_{tr} has about a_{te} before observing its actual behaviour; second, once a_{tr} has interacted with the a_{te} it updates the prior-parameter distribution to form the posterior distribution, $d(\theta_{a_{tr},a_{te}}|O^{0:t'}_{a_{tr},a_{te}})$; here, if $O^{0:t'}_{a_{tr},a_{te}} = \emptyset$, then $d(\theta_{a_{tr},a_{te}}) = d(\theta_{a_{tr},a_{te}}|O^{0:t'}_{a_{tr},a_{te}})$. Like, $b(x|\theta_{a_{tr},a_{te}})$, the shape of the parameter distribution is also determined by its own parameter vector. To distinguish between these two vectors, we refer to the latter as the hyperparameter, and denote it as $\phi_{a_{tr},a_{te}} \in \Phi^{\mathcal{C}}$, where $\Phi^{\mathcal{C}}$ is the hyperparameter space, for context $\mathcal{C}$.

Now, to find an appropriate estimate of $\theta_{a_{tr},a_{te}}$, we choose a value $\vartheta \in \Theta^{\mathcal{C}}$, which minimises an appropriate loss function, $L(\theta_{a_{tr},a_{te}}, \vartheta)$, according to the posterior-parameter distribution $d(\theta_{a_{tr},a_{te}}|O^{0:t'}_{a_{tr},a_{te}})$. Thus, ϑ is a *bayes estimate* of $\theta_{a_{tr},a_{te}}$ (Appendix A). Essentially, this gives us a basic level of trust based on a truster's direct interaction history with the trustee, and any prior assumptions that the truster may hold.

There are two important points about this that we make use of in subsequent sections. First, the posterior distribution is a function of the prior distribution and the dataset (Equation 3.1). Second, since the estimate ϑ is based on the parameter distribution, ϑ is also a function of the hyperparameter, $\phi_{a_{tr},a_{te}}$. Moreover, in this case of the posterior distribution, ϑ is a function of the prior distribution's hyperparameter, denoted $\phi^{prior}_{a_{tr},a_{te}}$, and the observed data $O^{0:t'}_{a_{tr},a_{te}}$ (Equation 3.2).

$$\phi^{post}_{a_{tr},a_{te}} = f(O^{1:t}_{a_{tr},a_{te}}, \phi^{prior}_{a_{tr},a_{te}}) \tag{3.1}$$

$$\theta_{a_{tr},a_{te}} \approx \vartheta = g(\phi^{post}_{a_{tr},a_{te}}) \tag{3.2}$$

3.2 Reputation Communication Framework

By concentrating on trustee observations, we can frame the challenge of assessing trust using reputation, as a generalisation of the basic technique we describe in the preceding section. In this case however, we no longer have one set of observations, but multiple, possibly overlapping datasets. This situation is illustrated in Figure 3.1. Here, a_{tr} has three sets of data on which to base its assessment of a_{te}: its own dataset $O^{1:t}_{a_{tr},a_{te}}$, and the dataset of two other agents, a_1 and a_2. The ideal solution to this situation would be to apply Equation 3.1 to the union of these datasets; in this case, the problem becomes equivalent to direct trust assessment. However, there are two basic obstacles

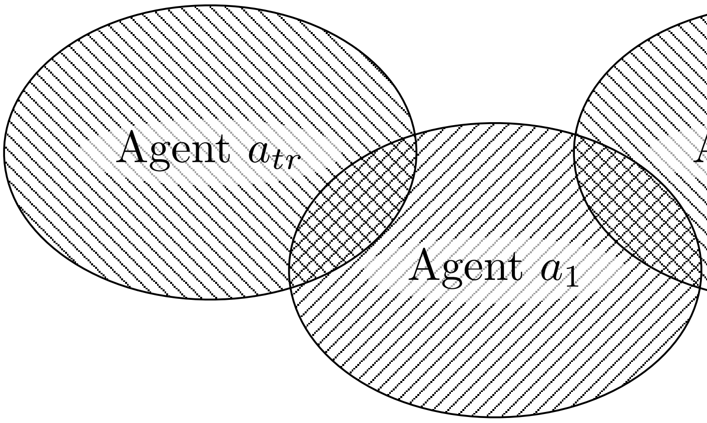

FIGURE 3.1: Venn Diagram of Overlapping Reputation Datasets

to this. First, in most scenarios of interest, we expect there to be a cost associated with communication between agents. Therefore, it would generally not be practical for agents to transmit their entire datasets to each other. Second, for reasons we have already covered, we can expect datasets for reputation sources to contain noise not associated with direct sources (Section 2.2.3). In this section, we concentrate on how and what information should be communicated between trusters and their reputation sources to alleviate these issues.

When a reputation source provides an opinion about a trustee, the important point is not that it should transmit the data on which it bases its opinion; rather, it should express only information relevant to assessing the trustee, and the opinion itself. To enable this, each reputation source, $a_{rep} \in \mathcal{A}$, should have a function r, such that $\mathcal{R}_{a_{rep},a_{te}} = r(O^{0:t}_{a_{rep},a_{te}})$. Here, $\mathcal{R}_{a_{rep},a_{te}}$ is the *opinion* of a_{rep} about a_{te}, and r is the *opinion function.* In the interest of simplicity, we assume there is one shared definition of r for all agents, and that the datasets on which each agent bases its reported opinion do not intersect. Note that these are not general requirements: different definitions of r may be acceptable, provided agents communicate the differences and interpret them appropriately; if agents' data do intersect then intersections could either be estimated or specified as part of $\mathcal{R}_{a_{rep},a_{te}}$. Despite this, we consider it outside the scope of our work to offer solutions in cases were these assumptions do not hold. With this in mind, we now specify the basic conditions that the opinion function r should satisfy.

Condition 1 *(Objectivity).* In general, trust is a subjective quality because it is assessed from the unique perspective and experience of the truster; for instance, a trustee may behave differently towards one agent than it does towards another. Subjectivity causes problems when assessing reputation because a truster must account for the differences between its own perspective, and the perspective of its reputation sources. That said, the opinion function is one source of subjectivity that we can control, by insisting that it is based only on objective measures[1]. Many existing trust models allow for trust to be expressed using subjective terms such as *good* or *bad*, and mention that the meaning of such terms may be different for different agents. Usually, such terms are subjective because they depend on bias or preferences that are individual to the truster. If a truster is human, then such subjectivity may be hard to tease apart from the underlying facts. However, in our case, we are only interested in software agents; so, we believe that subjectivity can be avoided, by not introducing it into the opinion function in the first place.

Condition 2 *(Composition).* When we assess trust based on direct observations, Equation 3.1 states that we require two things: the prior hyperparameter, $\phi^{prior}_{a_{tr},a_{te}}$; and the observations, $O^{0:t'}_{a_{tr},a_{te}}$. When we wish to take on board reputation, the set of observations is now $O_{complete} = O^{0:t'}_{a_{tr},a_{te}} \bigcup_{i=1}^{n} O^{1:t}_{a_i,a_{te}}$, where agents $a_1, ..., a_n$ are a_{tr}'s reputation sources. However, since we can only obtain information about $O^{1:t}_{a_i,a_{te}}$ through $\mathcal{R}_{a_i,a_{te}}$, we must find a way of calculating $\phi^{post}_{a_{tr},a_{te}}$ based on the opinions, rather than the complete dataset itself. To do this, we choose a function h such that:

$$\begin{aligned} &\exists f \quad \phi^{post}_{a_{tr},a_{te}} = f(s(O_{complete}), \phi^{prior}_{a_{tr},a_{te}}) \quad \text{where,} \\ &s(O_{complete}) = h(\mathcal{R}_{a_{tr},a_{te}}, \mathcal{R}_{a_1,a_{te}}, ..., \mathcal{R}_{a_n,a_{te}}) \end{aligned} \tag{3.3}$$

Here, we use the function h to combine all information obtained from a truster's reputation sources. For convenience, we make h a function of truster's own opinion, so that it can be treated in the same way; then, we ensure that s is a *decomposable statistic* of $O_{complete}$ (Definition 3.1), for which h is a *comprisal* function, and the opinion function r is a corresponding *constitute* function. In this way, we ensure that, no matter which way the observations in $O_{complete}$ divided between the reputation sources, the result will always be as if the truster made all the observations directly. This however, makes certain assumptions which we address in the proceeding section.

Definition 3.1 (Decomposable Statistic). Assume that X is a set of random variables corresponding to a set of observations. Then, a statistic is *any* function $s(X)$ of X (derived from Upton and Cook (2002)). Now assume that $X_1, ..., X_n$ are disjoint subsets of X, such that $\bigcup_{i=1}^{n} X_i = X$. A function $s(X)$ is a decomposable statistic, if it is a statistic and Equation 3.4 holds.

$$\exists h \quad \exists r \quad s(X) = h(r(X_1), ..., r(X_n)) \tag{3.4}$$

In this case, a function r is called a *constitute* function of $s(X)$ and the corresponding function h is called the *comprisal* function.

Condition 3 *(Minimal Communication).* In itself, Condition 2 does not guarantee the cost of communicating an opinion is within any bound of optimal. Rather, it formally

[1] This actually follows from the assumption that all agents share a common definition of the opinion function r; however, we believe that objectivity is an important condition, which warrants an explicit mention.

defines a set of functions which would produce the same result, had all the observations been made directly by the truster. To further ensure that only the required information is sent, we should attempt to choose a reputation function r, such that it is a *minimal constitute function* of statistic $s(O_{complete})$ (Definition 3.2).

Definition 3.2 (Minimal Constitute Function). Assume that s is a decomposable statistic and r is a constitute function of s. Then, r is a minimal constitute function of s if and only if r is a function of every other constitute function of s. Put formally,

$$\forall v \in C \quad \exists f \quad r = f(v), \quad \text{where } C \text{ is the set of constitute functions for } s.$$

3.3 Coping with Inaccurate Reputation

From the previous section, we have a fully specified framework for assessing trust based on reputation. However, for this framework to provide reasonable results for a given truster-trustee pair, then the following conditions must hold.

Condition 4. If $a_{tr} \in \mathcal{A}$ is a truster and $R \subseteq \mathcal{A}$ is the set of all reputation sources which a_{tr} consults about a trustee $a_{te} \in \mathcal{A}$, then the behaviour of a_{te} towards all members of $\{a_{tr}\} \bigcup R$ must be equal.

Condition 5. If $a_{tr} \in \mathcal{A}$ is a truster and $R \subseteq \mathcal{A}$ is the set of all reputation sources which a_{tr} consults about a trustee $a_{te} \in \mathcal{A}$, then all members of R must report their information about a_{te} truthfully and accurately.

Essentially, Conditions 4 to 5 ensure that observations made by a truster's reputation sources are representative of the actual behaviour a trustee is likely to have towards the truster. Unfortunately, we cannot expect these conditions to hold in general, so we must develop methods for coping with cases in which they are violated. Many of the trust models we review in Chapter 2 include methods for coping with some of these conditions. However, as we state in Section 2.3, each has its own set of downfalls. To address some of these downfalls, we define a two-step filtering mechanism: First, we calculate the probability that an agent will provide an accurate opinion given its past opinions, and later observed interactions with the trustees, for which those opinions were given. Second, based on this value, we reduce the distance between a rater's opinion and a prior belief that all possible values for an agent's behaviour are equally probable. Once all the opinions collected about a trustee have been adjusted in this way, the opinions are aggregated using the techniques described in Section 3.2.

To describe our reputation filtering mechanism in more detail, we must introduce some additional notation. The way in which we assess the accuracy of a reputation source's opinion is dependent on its actual value. For simplicity, we refer to the current opinion (of reputation source a_{rep}) under consideration as $\mathcal{R}^r$. With this in mind, we base decisions on five parameter distributions, all of which are defined for the trust parameter vector $\theta_{a_{tr},a_{te}}$ (Section 3.1). First, $d(\theta_{a_{tr},a_{te}}|\phi_c)$ is the parameter distribution that results when a truster assesses a trustee, based on its direct observations and all the opinions it receives about that trustee (Equation 3.3). Second, $d(\theta_{a_{tr},a_{te}}|\phi_r)$ is the parameter distribution that results when trust is based *only* on the considered opinion, $\mathcal{R}^r$, assuming a uniform prior. Third, $d(\theta_{a_{tr},a_{te}}|\phi_{c-r})$ is equivalent to $d(\theta_{a_{tr},a_{te}}|\phi_c)$, except that the opinion $\mathcal{R}^r$ is ignored. Fourth, $d(\theta_{a_{tr},a_{te}}|\phi_o)$ is a distribution based on directly observed interactions of trustee behaviour (Section 3.3.1). Finally, $d(\theta_{a_{tr},a_{te}}|\phi^{uni})$ is the uniform distribution, which represents an opinion of no information. In this discussion, we need to refer to the following properties for each of these distributions: the hyperparameter vector, denoted ϕ; the estimate based on the parameter distribution, ϑ; the standard deviation, denoted σ; and the expected value, denoted E. In each case, we link each property to the appropriate distribution by giving the corresponding subscript; for instance, $d(\theta_{a_{tr},a_{te}}|\phi_c)$ has expected value $E[\theta_{a_{tr},a_{te}}|\phi_c]$, standard deviation σ_c, hyperparameter vector ϕ_c, and estimate ϑ_c. In the following subsections we describe this technique in more detail: Section 3.3.1 details how the probability of accuracy is calculated and then Section 3.3.2 shows how opinions are adjusted and the combined reputation obtained.

3.3.1 Estimating the Probability of Accuracy

The first stage in our solution is to estimate the probability that a rater's stated opinion of a trustee is accurate. However, to do this we need to be more precise about what it means for an opinion to be accurate. Recall that an opinion $\mathcal{R}_{a_{rep},a_{te}}$, is essentially a summary of the observations an opinion source has had of a trustee's behaviour. If an opinion is fully trusted, the effect that the opinion has on the overall distribution $d(\theta_{a_{tr},a_{te}}|\phi_{a_{tr},a_{te}})$ is as if the underlying observations had been made directly by the truster. For this to produce reasonable results, those observations should be representative of the trustee's true behaviour towards the truster; we therefore require some measure of how representative an opinion is of a trustee's behaviour.

One way of doing this is to consider what an opinion actually tells us about the likely value of $\theta_{a_{tr},a_{te}}$. Specifically, it tells us that, according to the current knowledge of a_{rep}, the distribution of the parameter $\theta_{a_{tr},a_{te}}$ is $d(\theta_{a_{tr},a_{te}}|\phi_r)$. What we would like to do, is verify how correct this distribution is. Unfortunately, this is, in general, impossible because the distribution $d(\theta_{a_{tr},a_{te}}|\phi_r)$, is not only a function of the trustee's behaviour, but also a function of the number of interactions a_{rep} has had with a_{te}. If we did somehow know how many interactions have occurred between a_{rep} and a_{te}, we could verify this element of the opinion immediately. However, we assume that this information is not generally available to the truster.

What we can validate to an extent is $E[\theta_{a_{tr},a_{te}}|\phi_r]$. Imagine for the moment that $d(\theta_{a_{tr},a_{te}}|\phi_r)$ is the true distribution of $\theta_{a_{tr},a_{te}}$. Then, if we estimate our own distribution for $\theta_{a_{tr},a_{te}}$ (through repeated interactions with a_{te}) we should find that the expected value of our own distribution, converges on $E[\theta_{a_{tr},a_{te}}|\phi_r]$ as the number of observations increases. In our scenario, we have a very limited number of trustee observations with which to validate a given opinion in this way. This is because an agent has the right to change its opinion about a trustee: we could compare an opinion to observations of a trustee's behaviour indefinitely, but if the reputation source has since changed it mind, we would be judging it on a belief it no longer holds.

Fortunately, if we are interested in a reputation source's general accuracy, we don't have to consider its beliefs about any particular trustee in isolation: if we estimated a parameter distribution based on all observations (regardless of the trustee's identity), for which a_{rep} gave an opinion with a corresponding expected value E', then the expected value of our distribution would still converge on E', if a_{rep} generally provides accurate opinions. This result clearly follows because if we have n sets of samples, each with mean E', then the union of those sets will also have mean E'.

This still leaves us with one outstanding problem: since $\Phi^{\mathcal{C}}$ can in general be infinite, the probability of a_{rep} giving the same value for $E[\theta_{a_{tr},a_{te}}|\phi_r]$ twice is essentially zero. This again thwarts our attempts to gather enough observations to form a reliable parameter distribution with which to validate $E[\theta_{a_{tr},a_{te}}|\phi_r]$. Instead, the best we can do is consider all opinions for which $E[\theta_{a_{tr},a_{te}}|\phi_r]$ lies in a certain interval, and thus estimate the probability that the true mean lies within that interval; we call this probability the probability of accuracy, and denote it as $\rho_{a_{tr},a_{rep}}$. Bearing in mind that this only validates $E[\theta_{a_{tr},a_{te}}|\phi_r]$, we must assume that if $E[\theta_{a_{tr},a_{te}}|\phi_r]$ is unreliable, then so is $d(\theta_{a_{tr},a_{te}}|\phi_r)$, and if $E[\theta_{a_{tr},a_{te}}|\phi_r]$ is reliable, then $d(\theta_{a_{tr},a_{te}}|\phi_r)$ is likely to be reliable also. However, in Section 3.3.2, we take steps to prevent reputation sources manipulating trust assessment, by exaggerating the elements of $d(\theta_{a_{tr},a_{te}}|\phi_r)$ we cannot directly validate.

We now describe this process in more detail. First, let $\mathcal{H}_{a_{tr},a_{rep}}$ be the complete set of pairs of form $(\mathcal{R}_{a_{rep},a_x}, O_{a_{tr},a_x})$. Here, a_x is any member of $\mathcal{A}$, and O_{a_{tr},a_x} is the outcome of an interaction for which, prior to a_{tr} observing this outcome, a_{rep} gave the opinion $\mathcal{R}_{a_{rep},a_x}$. Second, divide the parameter space $\Theta^{\mathcal{C}}$ into disjoint intervals $\Theta^{\mathcal{C}}_1, ..., \Theta^{\mathcal{C}}_n$, such that $\bigcup_{i=1}^{n} \Theta^{\mathcal{C}}_i = \Theta^{\mathcal{C}}$. Third, calculate $E[\theta_{a_{tr},a_{te}}|\phi_r]$, and find the interval $\Theta^{\mathcal{C}}_r$ which contains its value. Fourth, find the subset $\mathcal{H}^r_{a_{tr},a_{rep}} \subseteq \mathcal{H}_{a_{tr},a_{rep}}$, which comprises all pairs for which the opinion falls in $\Theta^{\mathcal{C}}_r$. Now, we use the observations contained in $\mathcal{H}^r_{a_{tr},a_{rep}}$ to calculate the hyperparameter vector of $d(\theta_{a_{tr},a_{te}}|\phi_o)$; using this parameter distribution we now calculate $\rho_{a_{tr},a_{rep}}$ as the portion of the total mass of $d(\theta_{a_{tr},a_{te}}|\phi_o)$ that lies in the interval $\Theta^{\mathcal{C}}_r$.

An example of this process is illustrated in Figure 3.2. Here, the parameter space $\Theta^{\mathcal{C}}$ is instantiated as the set of real numbers in the range $[0, 1]$; this is then divided into

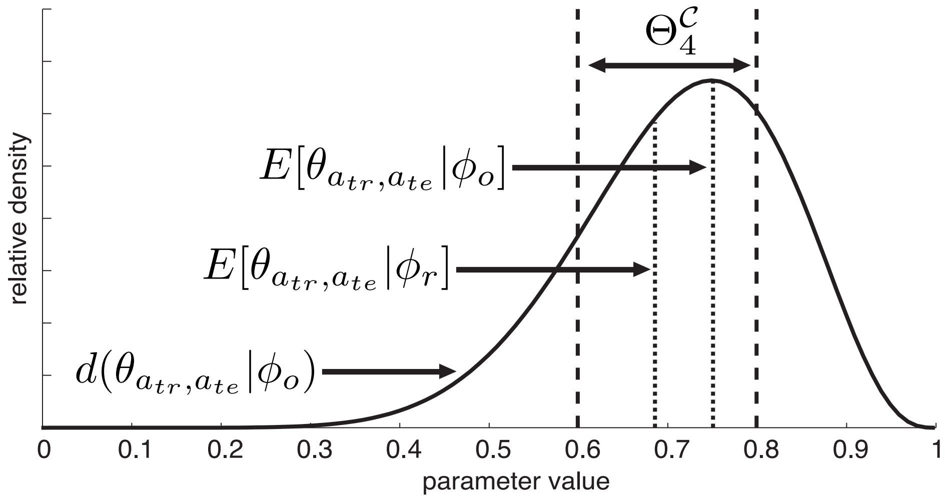

FIGURE 3.2: Illustration of $\rho_{a_{tr},a_{rep}}$ Estimation Process, for *sparamvec* $\in [0, 1]$

five intervals, $\Theta_1^{\mathcal{C}} = [0, 0.2], ..., \Theta_5^{\mathcal{C}} = [0.8, 1]$. The reputation source, a_{rep}, has provided a_{tr} with an opinion for which $E[\theta_{a_{tr},a_{te}}|\phi_r]$ is in $\Theta_4^{\mathcal{C}}$; thus, we calculate ϕ_o based on all previous interaction outcomes, for which a_{rep} provided an expected value in $\Theta_4^{\mathcal{C}}$ to a_{rep}. As can been seen in the figure, the parameter distribution based on these outcomes, $d(\theta_{a_{tr},a_{te}}|\phi_o)$, is peaked inside $\Theta_4^{\mathcal{C}}$, with a significant proportion of its mass also in $\Theta_4^{\mathcal{C}}$. Integrating $d(\theta_{a_{tr},a_{te}}|\phi_o)$ over $\Theta_4^{\mathcal{C}}$ will thus give us a significantly high value for $\rho_{a_{tr},a_{rep}}$. If subsequent outcome-opinion pairs were also to follow this trend, then $d(\theta_{a_{tr},a_{te}}|\phi_o)$ would become increasingly peaked inside this interval; therefore $\rho_{a_{tr},a_{rep}}$ would converge to one. On the other hand, if subsequent outcomes disagreed with their corresponding opinions, then $\rho_{a_{tr},a_{rep}}$ would approach 0. One implication of this technique is that the number of bins effectively determines an acceptable margin of error in opinion provider accuracy: a larger set of opinion providers will have their estimated accuracy converge to 1 if bin sizes are large, compared to if bin sizes are small.

3.3.2 Adjusting Reputation Source Opinions

Once we have calculated $\rho_{a_{tr},a_{rep}}$, we then adjust $\mathcal{R}^r$ such that its effect on trust is decreased in line with its probability of accuracy; to do this, we define a mapping function $\mathcal{R}^a = m(\mathcal{R}^r, \rho_{a_{tr},a_{rep}})$, where $\mathcal{R}^a$ is the adjusted opinion, with corresponding hyperparameter ϕ_a. Intuitively, if $\rho_{a_{tr},a_{rep}} = 1$, then the effect that $\mathcal{R}^r$ has on $d(\theta_{a_{tr},a_{te}}|\phi_c)$ should remain unchanged; if $\rho_{a_{tr},a_{rep}} = 0$, $\mathcal{R}^r$ should have no effect; and if $0 < \rho_{a_{tr},a_{rep}} < 1$, the effect should be reduced by some proportion. Apart from this, the precise definition of the mapping function depends on the way in which the parameter distribution is modelled; therefore, we cannot define it completely here. Instead, in this section, we give a set of guidelines which we believe any instantiation of the function must satisfy. Specifically, we identify three aspects of an opinion that the mapping function should account for: (1) the size of dataset the opinion is (reportedly) based on, (2) the conclusion the opinion supports, (3) the ability of a reputation source to manipulate the trust assessment process. In the remainder of this section, we shall consider each of these points in turn.

Dataset size — If the size of a reputation source's dataset is large compared to all other data a truster bases its decision on, then the truster's opinion will move significantly towards that reputation sources opinion. On the other hand, if the reputation sources dataset is empty, the truster's opinion will not change at all[2]. This shows that we can reduce the effect an opinion has, by adjusting it so that the size of the underlying data we pay attention to, is effectively reduced.

Opinion Conclusion — The conclusion that a reputation source's opinion supports is also significant. Mainly, this is because the opinion's effect is relative to the amount of data the truster has from other sources; if the truster has little alternative evidence, then it would move its opinion significantly in the direction of the reputation source opinion — even if the reputation source opinion was itself based on little evidence. This means only adjusting an opinion such that the size of the underlying dataset is reduced, will not prevent a truster placing undue emphasis on an untrustworthy opinion. This is particularly important when little is known by anyone about a certain trustee, in which case the only opinions based on reportedly high datasets, will be from agents with an incentive to lie.

Opinion Manipulation — Imagine halving the size of the dataset to reduce its impact on trust; if a reputation source knew that this was our intention, it could simply double the reported size of its dataset, to counteract our adjustment. This is one example of how a reputation source could potentially manipulate its opinion to counteract the adjustment procedure; any instantiation of the mapping function should therefore include measures to prevent such interference. One way to do this is to use the standard deviation, or the variance, of the opinion parameter distribution, $d(\theta_{a_{tr},a_{te}}|\phi_r)$. The change in the parameter distribution decreases as the size of the dataset increases: as the parameter distributions variance approaches zero (due to increase in dataset size) the effect of any new data becomes minimal; moreover, when the variance is zero, no new data will change the distribution. This means that by adjusting according to the distribution variance, we make it increasingly hard for a reputation source to bias a truster, by exaggerating the weight of its own evidence.

From these three points, we conclude that the mapping function should decrease the size of the adjusted dataset according to the distribution variance, and move the conclusion

[2] This result is due to the way we combine reputation: recall for Section 3.2 that trust based on opinions is equivalent to trust based on the union of the truster's own observations, and those of its reputation sources.

it supports in the direction of the truster's own prior opinion. This latter condition will introduce a bias towards the trusters own opinion, but we believe this bias is justified to guard against the effect of unreliable opinions.

Chapter 4

TRAVOS: A Trust Model for Boolean Action Spaces

In this chapter, we introduce a trust model called TRAVOS (Trust and Reputation system for Agent Based Virtual OrganisationS), which instantiates the framework described in the previous chapter for boolean action spaces. By this we mean we concentrate on scenarios in which a trustee can only behave in one of two ways during an interaction: either it can cooperate, and fulfil its obligations to the truster; or it can defect, breaking its obligations to the truster. In turn, we assume that the truster's utility is only dependent on which of these actions a trustee takes, and we define its quality of service function $Q(a_{tr}, \mathcal{C}, \mathcal{S}_{\mathcal{C}})$ as a binary function that simply records which action a trustee took, during a given interaction.

We divide the chapter into four sections. First, Section 4.1 describes the process of instantiating the framework in general, and then describes the boolean instantiation used in TRAVOS. Second, Section 4.2 presents a method for collecting reputation information that can be used to calculate trust using the framework instantiation. Third, Section 4.3 describes how TRAVOS is applied as part of a system for management agent-based VOs, including a walk-through scenario outlining its use. Finally, Section 4.4 gives an empirical evaluation of TRAVOS, through computer simulation.

4.1 Instantiating the Framework for Boolean Action Spaces

Essentially, the process of instantiating the framework starts by considering the definition of the quality of service function $Q(a_{tr}, \mathcal{C}, \mathcal{S}_\mathcal{C})$, followed by finding appropriate definitions for the various elements described in Chapter 3. Overall, the process involves the following six steps:

1. define $Q(a_{tr}, \mathcal{C}, \mathcal{S}_\mathcal{C})$ and thus define the quality of service space $\mathcal{O}^\mathcal{C}$
2. find a suitable parameter model to represent the behaviour distribution, $b(x \in \mathcal{O}^\mathcal{C} | \theta_{a_{tr},a_{te}})$
3. choose an appropriate parameter model for the parameter distribution, $d(\theta_{a_{tr},a_{te}} | \phi_{a_{tr},a_{te}})$
4. choose an appropriate loss function $L(\theta_{a_{tr},a_{te}}, \vartheta)$, and thus derive the optimal definition for the bayes estimator function, ϑ
5. define the reputation function, r, according to the conditions set out in Section 3.2
6. instantiate the reputation mapping function, $\mathcal{R}^a = m(\mathcal{R}^r, \rho_{a_{tr},a_{rep}})$

To describe the instantiation used in TRAVOS, we now discuss each of these steps in turn: Section 4.1.1 describes the parameter models used in TRAVOS (Steps 1 to 3); Section 4.1.2 defines the optimal estimator for $\theta_{a_{tr},a_{te}}$ (Step 4); and Section 4.1.3 instantiates the reputation mechanism (Steps 5 to 6).

4.1.1 Parameter Models for Binary Action Space

As we have already stated, in TRAVOS, both the trustee action space and the quality of service function are binary: either a trustee cooperates, constituting a successful interaction for the truster, or the trustee defects, constituting an unsuccessful interaction for the truster. The definition of the quality of service space is thus given by Equation 4.1.

$$\mathcal{O}^\mathcal{C} = \begin{cases} 1 & \text{if contract is fulfilled by } a_{te} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

This binary definition means that a series of observations of trustee behaviour, such as $O^{0:t'}_{a_{tr},a_{te}}$ can be considered as a set of Bernoulli trials, and thus is drawn from a Bernoulli distribution; that is, the (intrinsic) probability distribution is completely described by the probability that $O_{a_{tr},a_{te}} = 1$ — we simply subtract this from 1, to obtain the probability that $O_{a_{tr},a_{te}} = 0$. The parameter $\theta_{a_{tr},a_{te}}$ is therefore a real number in the range $[0, 1]$, which represents $p(O_{a_{tr},a_{te}} = 1)$ (Equation 4.2).

$$\theta_{a_{tr},a_{te}} = p(O_{a_{tr},a_{te}} = 1), \quad \text{where } \theta_{a_{tr},a_{te}} \in [0, 1] \tag{4.2}$$

In the interest of simplicity, we adopt the standard practice of choosing a conjugate prior for the parameter distribution (DeGroot and Schervish, 2002a). In case of Bernoulli distributions, the conjugate family is the set of beta distributions. In this respect TRAVOS is therefore similar to the Beta Reputation System (Section 2.2.2.2). The hyperparameter space, $\Phi^{\mathcal{C}}$, now takes on the form of the standard parameters of the beta distribution (Equation 4.3). Specifically, the beta distribution has two parameters, typically denoted α and β, both of which are positive real numbers. These parameters determine the shape of the distribution through the probability density function (Equation 4.4), the expected value of the distribution (Equation 4.5) and the variance (Equation 4.6).

$$\Phi^{\mathcal{C}} = \{(\alpha,\beta)|\alpha > 0 \wedge \beta > 0\} \tag{4.3}$$

$$d(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int U^{\alpha-1}(1-U)^{\beta-1}dU} \tag{4.4}$$

$$E[\theta|\alpha,\beta] = \frac{\alpha}{\alpha+\beta} \tag{4.5}$$

$$\sigma^2 = \frac{\alpha\cdot\beta}{(\alpha+\beta)(\alpha+\beta+1)} \tag{4.6}$$

With this in mind, we can now show how the various aspects of the beta distribution can be applied to the framework. In particular, for a given prior, $\phi^{prior}_{a_{tr},a_{te}} = (\alpha^{prior},\beta^{prior})$, the posterior hyperparameter, $\phi^{post}_{a_{tr},a_{te}} = (\alpha^{post},\beta^{post})$, is calculated by counting the number of successful interactions (Equation 4.7) and the number of unsuccessful interactions (Equation 4.8) in the interaction history, $O^{0:t'}_{a_{tr},a_{te}}$; and then adding these values to the α and β parameters as shown in Equations 4.9 & 4.10. This is a well known result, a derivation of which is given in DeGroot and Schervish (2002a).

$$m_{a_{tr},a_{te}} = |\{o \in O^{0:t'}_{a_{tr},a_{te}}|o = 1\}| \tag{4.7}$$

$$n_{a_{tr},a_{te}} = |\{o \in O^{0:t'}_{a_{tr},a_{te}}|o = 0\}| \tag{4.8}$$

$$\alpha^{post} = \alpha^{prior} + m_{a_{tr},a_{te}} \tag{4.9}$$

$$\beta^{post} = \beta^{prior} + n_{a_{tr},a_{te}} \tag{4.10}$$

The effect of updating the parameter distribution in light of observations is illustrated in Figure 4.1. Here, adding observations, and thus increasing α and β, decreases the distribution variance, making the distribution more peaked. The proportion of successful and unsuccessful interactions, along with the prior, determine where on the interval $[0,1]$ the distribution peaks. A high α value compared to β (usually resulting from a high proportion of successful outcomes) causes the distribution mode to occur close to 1. Intuitively, this is correct, because it supports the conclusion that the intrinsic probability of $O_{a_{tr},a_{te}} = 1$ is also close to 1.

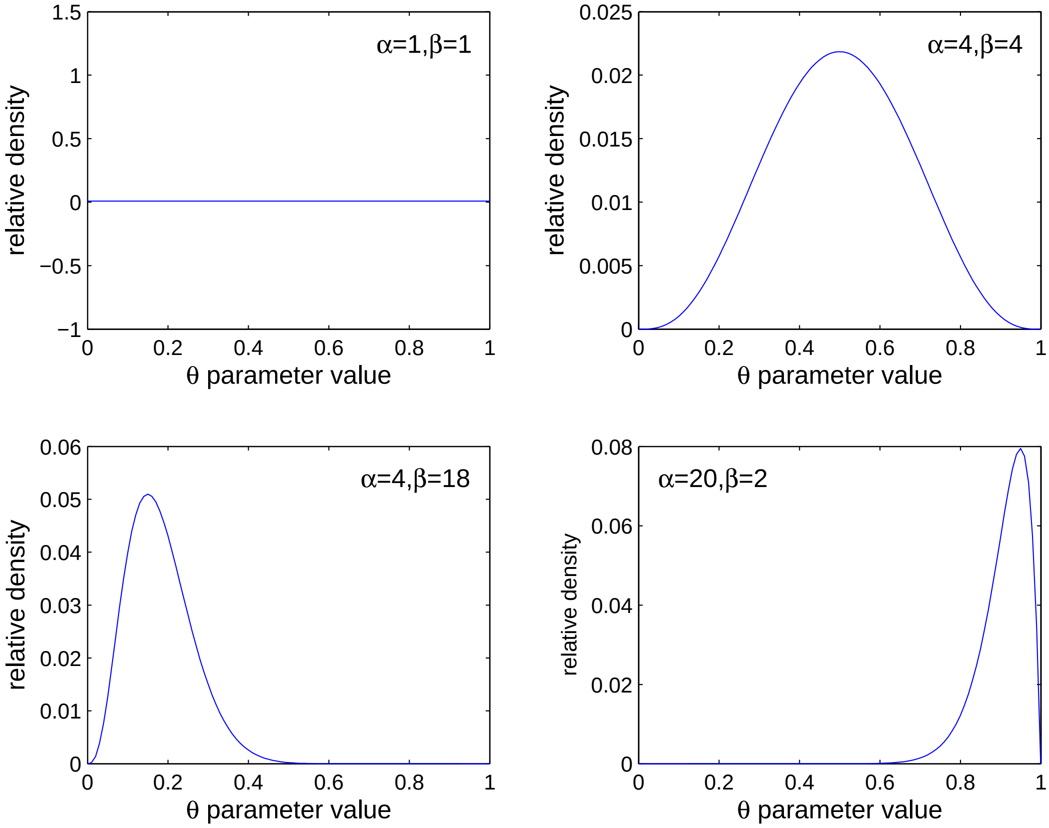

FIGURE 4.1: Example beta pdf plots; note that when $\alpha = 1, \beta = 1$ (top-left) the distribution is uniform on the interval $[0, 1]$.

4.1.2 The Parameter Estimator

Now that we have defined the distribution parameter distribution, we can derive the estimator function for the parameter vector. To do this, we choose the *mean squared error* as the loss function because this encourages our estimate to be, on average, as close to the true parameter value as possible. To find the estimator, we find the value which minimises the expected value of this loss function. Through Theorem 4.1, we see that, in general, the estimator which minimises mean squared error, is the expected value, $E[\theta_{a_{tr},a_{te}}]$; therefore, in this case, the optimal estimate ϑ is given by Equation 4.5.

Theorem 4.1 (Bayes estimate using mean squared error). *Assume that θ is a parameter with probability distribution $p(\theta)$, where $a \leq \theta \leq b$. Then, the bayes estimate (which we denote ϑ) that minimises the expected mean squared error (Equation 4.11) is the expected value $E[\theta]$ according to the distribution $p(\theta)$.*

$$\textit{mean squared error} = L(\theta, \vartheta) = (\vartheta - \theta)^2 \tag{4.11}$$

proof: *First, we differentiate the expected mean squared error with respect to ϑ. Then, we set the derivative equal to 0 to find the minimum point.*

$$E[L(\theta,\vartheta)] = \int_a^b (\vartheta-\theta)^2 \cdot p(\theta) \quad d\theta \quad \textit{(by definition)} \tag{4.12}$$

$$\frac{d}{d\vartheta} \quad E[L(\theta,\vartheta)] = \frac{d}{d\vartheta}\int_a^b (\vartheta-\theta)^2 \cdot p(\theta) \quad d\theta \quad = \quad 0 \tag{4.13}$$

$$\frac{d}{d\vartheta} \quad E[L(\theta,\vartheta)] = \int_a^b \left[\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta)\right] \quad d\theta \quad = \quad 0 \tag{4.14}$$

evaluating the derivative we get

$$\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta) = p(\theta)\cdot\left[\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2\right] + (\vartheta-\theta)^2 \cdot \left[\frac{d}{d\vartheta} \quad p(\theta)\right] \tag{4.15}$$

$$\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta) = p(\theta)\cdot\left[\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2\right] + (\vartheta-\theta)^2 \cdot 0 \tag{4.16}$$

$$\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta) = p(\theta)\cdot\left[\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2\right] \tag{4.17}$$

$$\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta) = p(\theta)\cdot\left[\frac{d}{d(\vartheta-\theta)} \quad (\vartheta-\theta)^2\right]\cdot\left[\frac{d}{d\vartheta} \quad \vartheta-\theta\right] \tag{4.18}$$

$$\frac{d}{d\vartheta} \quad (\vartheta-\theta)^2 \cdot p(\theta) = p(\theta)\cdot[2\cdot(\vartheta-\theta)]\cdot 1 \tag{4.19}$$

$$\tag{4.20}$$

substituting into the integral we get

$$0 = \int_a^b 2(\vartheta-\theta)\cdot p(\theta) \quad d\theta \tag{4.21}$$

$$0 = \int_a^b 2\vartheta\cdot p(\theta) - 2\theta\cdot p(\theta) \quad d\theta \tag{4.22}$$

$$0 = 2\cdot\left[\int_a^b \vartheta\cdot p(\theta) \quad d\theta\right] - 2\cdot\left[\int_a^b \theta\cdot p(\theta) \quad d\theta\right] \tag{4.23}$$

$$\int_a^b \vartheta\cdot p(\theta) \quad d\theta = \int_a^b \theta\cdot p(\theta) \quad d\theta \tag{4.24}$$

$$\int_a^b \vartheta\cdot p(\theta) \quad d\theta = E[\theta] \tag{4.25}$$

$$\vartheta\cdot\int_a^b p(\theta) \quad d\theta = E[\theta] \tag{4.26}$$

$$\vartheta\cdot 1 = E[\theta] \quad \textit{(probability distribution integrates to 1)} \tag{4.27}$$

$$\vartheta = E[\theta] \tag{4.28}$$

Hence, the bayes estimator which minimises expected mean squared error is $E[\theta]$.

4.1.3 Instantiating the Reputation Mechanism

To allow reputation to be used during trust assessment, we must define the reputation function (Section 3.2), and instantiate the opinion mapping function. First, let us consider the reputation function. In Section 3.2, we state that there three three conditions that the reputation function should satisfy: objectivity, composition and (optionally) minimalism. One function which satisfies all of these conditions is the pair of frequencies for successful and unsuccessful interactions observed by the reputation source with the trustee (Equation 4.29).

$$r(O^{0:t'}_{a_{rep},a_{te}}) = (m_{a_{rep},a_{te}}, n_{a_{rep},a_{te}}), \quad \text{where } m_{a_{rep},a_{te}} = \text{successful frequency} \\ \text{and } n_{a_{rep},a_{te}} = \text{unsuccessful frequency} \tag{4.29}$$

This function is objective because it is dependent only on the behaviour of the trustee, and not directly[1] on the identity of the observing reputation source. The function satisfies the composition condition, because the posterior parameter distribution can be written as follows.

$$\phi^{post}_{a_{tr},a_{te}} = f(s(O_{complete}), \phi^{prior}_{a_{tr},a_{te}}) \quad \text{(from Equation 3.3)} \tag{4.30}$$

$$\phi^{post}_{a_{tr},a_{te}} = \left(\alpha^{prior} + M_{a_{tr},a_{te}}, \quad \beta^{prior} + N_{a_{tr},a_{te}}\right), \quad \text{where} \tag{4.31}$$

$$s(O_{complete}) = (M_{a_{tr},a_{te}}, N_{a_{tr},a_{te}}) \tag{4.32}$$

$$M_{a_{tr},a_{te}} = \sum_{a_i \in \mathcal{S} \bigcup \{a_{tr}\}} m_{a_i,a_{te}} \tag{4.33}$$

$$N_{a_{tr},a_{te}} = \sum_{a_i \in \mathcal{S} \bigcup \{a_{tr}\}} n_{a_i,a_{te}} \tag{4.34}$$

Here, $\mathcal{S}$, is the set of reputation sources consulted by a_{tr}; $\phi^{post}_{a_{tr},a_{te}}$ is a function of the prior $\phi^{prior}_{a_{tr},a_{te}} = (\alpha^{prior}, \beta^{prior})$ and is a statistic of the complete set of observations, $O_{complete}$. The statistic $s(O_{complete})$ is a decomposable function for which r is a constitute function, because s can be obtained by summing the elements of r from each reputation source (Equation 4.31). Finally, r is a minimal function of s because r and s have identical definitions:

$$s : O = \{o | o = 1 \vee o = 0\} \longrightarrow \{(m, n) | m > 0 \wedge n > 0\},$$
$$r : O = \{o | o = 1 \vee o = 0\} \longrightarrow \{(m, n) | m > 0 \wedge n > 0\},$$
$$\text{where } m = \sum_{o \in O} o \text{ and } n = \sum_{o \in O} (1 - o)$$

This means that r is a decomposable statistic, for which all constitute functions of s are also constitute functions of r; hence, r is a minimal constitute function of s.

[1] Of course, the trustee may change its behaviour depending on the identity of the observer, but the function itself is still objective.

Our final task is to instantiate the reputation mapping function. To do this, we first map the parameter space on to the expected value and variance of the parameter distribution (Equations 4.5 and 4.6). Now, to define the opinion mapping function, we reduce the euclidean distance between the vector $(E[\theta_{a_{tr},a_{te}}|\phi_r], \sigma_r^2)$ for the opinion distribution, and the equivalent vector for the uniform distribution, $(E[\theta_{a_{tr},a_{te}}|\phi^{uni}], \sigma_{uni}^2)$ (Equations 4.35 and 4.36). From this, the adjusted opinion, $\mathcal{R}^a$, can be determined using Equations 4.37 to 4.40; in all of these equations, we use the over bar to denote properties belonging to the adjusted opinion (see Appendix B for a derivation of Equations 4.37 and 4.38).

$$\bar{E} = E[\theta_{a_{tr},a_{te}}|\phi^{uni}] + \rho_{a_{tr},a_{rep}} \cdot (E[\theta_{a_{tr},a_{te}}|\phi_r] - E[\theta_{a_{tr},a_{te}}|\phi^{uni}]) \quad (4.35)$$

$$\bar{\sigma}^2 = \sigma_{uni}^2 + \rho_{a_{tr},a_{rep}} \cdot (\sigma_r^2 - \sigma_{uni}^2) \quad (4.36)$$

$$\bar{\alpha} = \frac{\bar{E}^2 - \bar{E}^3}{\bar{\sigma}^2} - \bar{E} \quad (4.37)$$

$$\bar{\beta} = \frac{(1-\bar{E})^2 - (1-\bar{E})^3}{\bar{\sigma}^2} - (1-\bar{E}) \quad (4.38)$$

$$\bar{m}_{a_{rep},a_{te}} = \bar{\alpha} - 1 \quad , \quad \bar{n}_{a_{rep},a_{te}} = \bar{\beta} - 1 \quad (4.39)$$

$$\mathcal{R}^r = (\bar{m}_{a_{rep},a_{te}}, \bar{n}_{a_{rep},a_{te}}) \quad (4.40)$$

Defining the mapping function in this way, satisfies the guidelines described in Section 3.3.2 for the following reasons. First, adjusting the variance effectively reduces the size of the underlying dataset which we pay attention to. Second, since we adjust the variance linearly, we are increasingly skeptical of ever larger datasets; thus, we reduce the ability of a reputation source to counteract the adjustment process by manipulating its opinion. Finally, by moving the expected value of the distribution towards uniform (i.e. a value of 0.5) we make the adjusted opinion more conservative, thus reducing the effect of untrusted opinions in cases where no other information is available.

4.2 Reputation Gathering in TRAVOS

In the preceding sections, we show how, by using the framework, reputation information can be used along with a truster's direct experience to assess the trustworthiness of an agent. However, apart from assessment, there are two other issues that a practical trust and reputation system should include: (1) agents require some mechanism to obtain opinions from reputation sources, and (2) agents must decide when it is necessary to obtain reputation information. The latter is important, because if a truster has sufficient direct evidence with which to judge a trustee, the cost of obtaining reputation information may outweigh its benefits. We now consider each of these issues in turn.

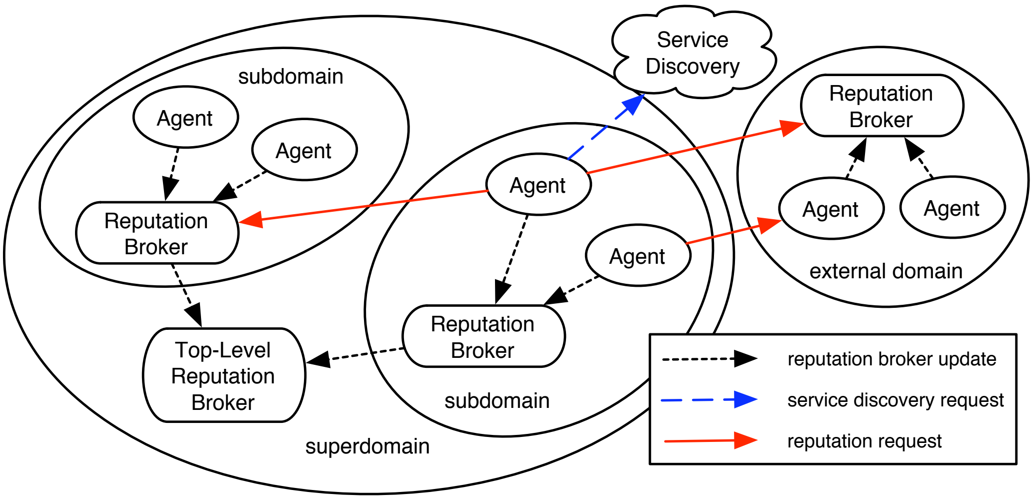

FIGURE 4.2: Reputation Brokering System

4.2.1 Reputation Brokering

The problem with obtaining opinions in large systems, is that directly querying many agents may entail a significant communication overhead. Therefore, agents must do one or more of the following: (1) choose a subset of agents to query, (2) employ some method of streamlining reputation. Our solution to this problem is illustrated in Figure 4.2.

We assume that each agent in a system belongs to exactly one primary *domain*. Here, a domain may correspond to an organisation or department in the real world, to which the agent is responsible. This view is in line with the vision of systems, such as the Grid, in which computing resources belonging to different organisations may be used together. Within each domain, there is a *reputation broker* agent, which is responsible for aggregating the opinions of all other agents within its domain; that is, the opinion of a reputation broker about a trustee is an aggregation of the opinions of all other agents within its domain. In addition, domains can be arranged in a hierarchy such that brokers in subdomains report to a broker in an overall domain. It this way, a top level broker aggregates all the opinions of agents in each of its subdomains.

Reputation Brokers provide a point of contact for external agents looking to receive reputation information. When a truster requires reputation, it first uses a service discovery system (such as described in Section 4.3) to identify domains that advertise having information about trustee's in some general context[2]. For example, companies which make use of grid-based storage space, may advertise having knowledge about vendors of such storage space.

[2] Here, we do not address the issue of at what level domains should advertise information. For example, if a department within a company is mainly responsible for certain information, should the department be the advertised point of contact, or the organisation it belongs to?

```
========================================================================
  Each time an interaction outcome is observed do the following
========================================================================
IF interaction successful
  SET m[trustee_id] = m + 1
ELSE
  SET n[trustee_id] = n + 1
END IF

========================================================================
  Periodically do the following
========================================================================
FOR ALL i = trustee_id
  IF m[i] $\neq$ 0 OR n[i] $\neq$ 0
    add m[i] and n[i] to update message
  END IF
  SET m[i] = 0
  SET n[i] = 0
END LOOP
SEND update message to reputation broker
```

Figure 4.3: Reputation broker update algorithm, performed by reputation sources

Once a truster has received a list of appropriate domains, it can choose to request an opinion from either the main reputation broker for that domain, or other brokers or individual agents within that domain. Although we do not specify how a truster should make this choice, there is an obvious trade-off in granularity. By requesting information from a top-level broker, the truster can receive all the information known by the domain in a single message. However, in this case, a truster can only judge the accuracy of the broker's domain as a whole (using the techninques described in Section 4.1.3). On the other hand, if a truster contacted several agents within a domain, it could judge their accuracy individually, thus identifying the most reliable contacts within an organisation. Here, it is important that a truster should avoid using a reputation source at the same time as any reputation broker, which that source reports to. The reason for this is correlated evidence (Requirement 3.3): since the broker's opinion is based on those agents which report to it, using a reputation source along with its broker would amount to counting the reputation source's opinion twice!

We now describe how a reputation broker's opinion is formed. Each broker periodically receives updates regarding any newly observed interaction outcomes, from the agents within its own domain of responsibility. These updates take the same form as normal reputation opinions in TRAVOS (Equation 4.29) except that they are only based on observations that have occurred since the last update the observer sent to its broker. This process is summarised algorithmically in Figure 4.3. When a reputation broker receives an opinion from within its domain about a trustee a_{te}, it updates its own opinion about a_{te} using Equations 4.41 and 4.42. In this way, the broker's opinion can be compared to that of a single agent, which has observed all the interaction outcomes recorded by the agents within the broker's domain.

$$m_{a_{broker},a_{te}} = m_{a_{broker},a_{te}} + \sum_{a_i \in \mathcal{D}} m^*_{a_i,a_{te}} \tag{4.41}$$

$$n_{a_{broker},a_{te}} = n_{a_{broker},a_{te}} + \sum_{a_i \in \mathcal{D}} n^*_{a_i,a_{te}} \tag{4.42}$$

where $(m^*_{a_i,a_{te}}, n^*_{a_i,a_{te}})$ is the update message from a_i about a_{te}, and $\mathcal{D}$ is the set of agents in a_{broker}'s domain.

4.2.2 When to Seek Reputation

In some cases, an agent may decide that it is sufficiently confident in its own knowledge about a trustee, to avoid acquiring reputation information to improve its estimate. Two reasons for this are the communication cost of reputation acquisition, and the inherit unreliability of reputation compared to direct observations. One simple method of doing this is to calculate the posterior probability that the true value for $\theta_{a_{tr},a_{te}}$ lies within an acceptable margin of error around the estimate. We can calculate this using the parameter distribution as follows. First, we decide on an acceptable error margin, $\vartheta a_{tr} a_{te} \pm \epsilon$, where ϵ is the acceptable distance from $\vartheta a_{tr} a_{te}$. Second, we integrate the parameter distribution over the area define by the error margin. To do this, we use the beta probability density function as shown in Equation 4.43. We refer to the resulting value as the *confidence* value of $\vartheta a_{tr} a_{te}$, which we denote as $\gamma_{a_{tr},a_{te}}$. Finally, we choose a threshold for this probability, above which we consider the accuracy of the estimate as acceptable; we denote this threshold as τ.

$$\gamma_{a_{tr},a_{te}} = \frac{\int_{\vartheta_{a_{tr},a_{te}}+\epsilon}^{\vartheta_{a_{tr},a_{te}}-\epsilon} B^{\alpha-1}(1-B)^{\beta-1}dB}{\int_0^1 U^{\alpha-1}(1-U)^{\beta-1}dU}, \quad \text{where } (\alpha,\beta) = \phi_{a_{tr},a_{te}} \tag{4.43}$$

4.3 An Application to Agent-Based VOs

In this section, we describe the role of TRAVOS in the CONOISE-G system (Patel et al., 2005; Shao et al., 2004). The CONOISE-G system (Constraint Oriented Negotiation in Open Information Seeking Environments for the Grid) seeks to, *"support robust and resilient virtual organisation formation and operation. It aims to provide mechanisms to assure effective operation of VOs in the face of disruptive and potentially malicious entities in dynamic, open and competitive environments."*[3] More specifically, CONOISE-G provides methods by which agents operating in a grid environment can form dynamic resource coalitions (VOs), in order to fulfil their goals. Here, by dynamic we mean that the membership of a VO may change over its lifetime. This can happen for various reasons; for instance, a particular member's resources may fail, requiring a

[3] This quote is taken from http://www.conoise.org/

new member to make up the shortfall. In the following subsections we give an overview of the CONOISE-G system (Section 4.3.1), following by a trust-orientated scenario of how TRAVOS is used in CONOISE-G (Section 4.3.2).

4.3.1 System Overview

In essence, the CONOISE-G architecture comprises several different agents, including *system agents* and *service providers* (SPs), as shown in Figure 4.4. The former are those needed to achieve core system functionality for VO formation and operation, while the latter are those involved in the VO itself. Moreover, SPs are responsible for overseeing the life cycle of a VO, which consists for three stages: (1) formation, (2) operation and (3) dissolution. The formation of a VO consists of three steps:

1. **Resource Discovery** — A particular SP, acting either on its own behalf, or on behalf of a user, identifies a need for a number of resources, which it cannot supply (efficiently) by itself. To fulfil this need, the SP instigates VO formation, by requesting a list of other SPs, which can supply the required resources; it obtains this list from the Yellow Pages Agent (YP), which performs a service discovery role (Deora et al., 2004). At this point, the SP which places the request for service, takes on the VO Manager (VOM) role for the potential VO, as illustrated in Figure 4.4.

2. **Resource Assessment** – After receiving a response from the YP, the VOM invites the identified providers to bid for the requested services. Once, all such bids are received, the VOM generates an expected utility function for each bid based on the price offered per resource unit, trust and the advice given by the Quality of Service Assessor (QoSA). The QoSA, based on Deora et al. (2003), is an external service which rates how well a given SP is likely to perform. Its role can be viewed as similar to that of a reputation provider in TRAVOS, in that it provides extra information about a trustee's likely behaviour. However, the nature of its assessment and its underlying assumptions are different from that of reputation sharing in TRAVOS, and therefore it must be treated differently.

   In our approach, we first estimate the SP's behaviour distribution (as described in Section 4.1.2) thereby estimating the probability that the SP will fulfil its obligations to the VOM. Then, we use the QoSA's assessment of an SP to provide an alternative estimate of this probability, and combine these two estimates using a suprabayesian Approach (Keeney and Raiffa, 1976). In general, the combined probability should be more accurate than either of the individual estimates, since it incorporates the knowledge of both the QoSA, the VOM (in its role as a truster) and the VOM's reputation sources. The combined probability is then used to calculate the expected utility for the VOM, for each possible number of resource units it can purchase from the bidding SP.

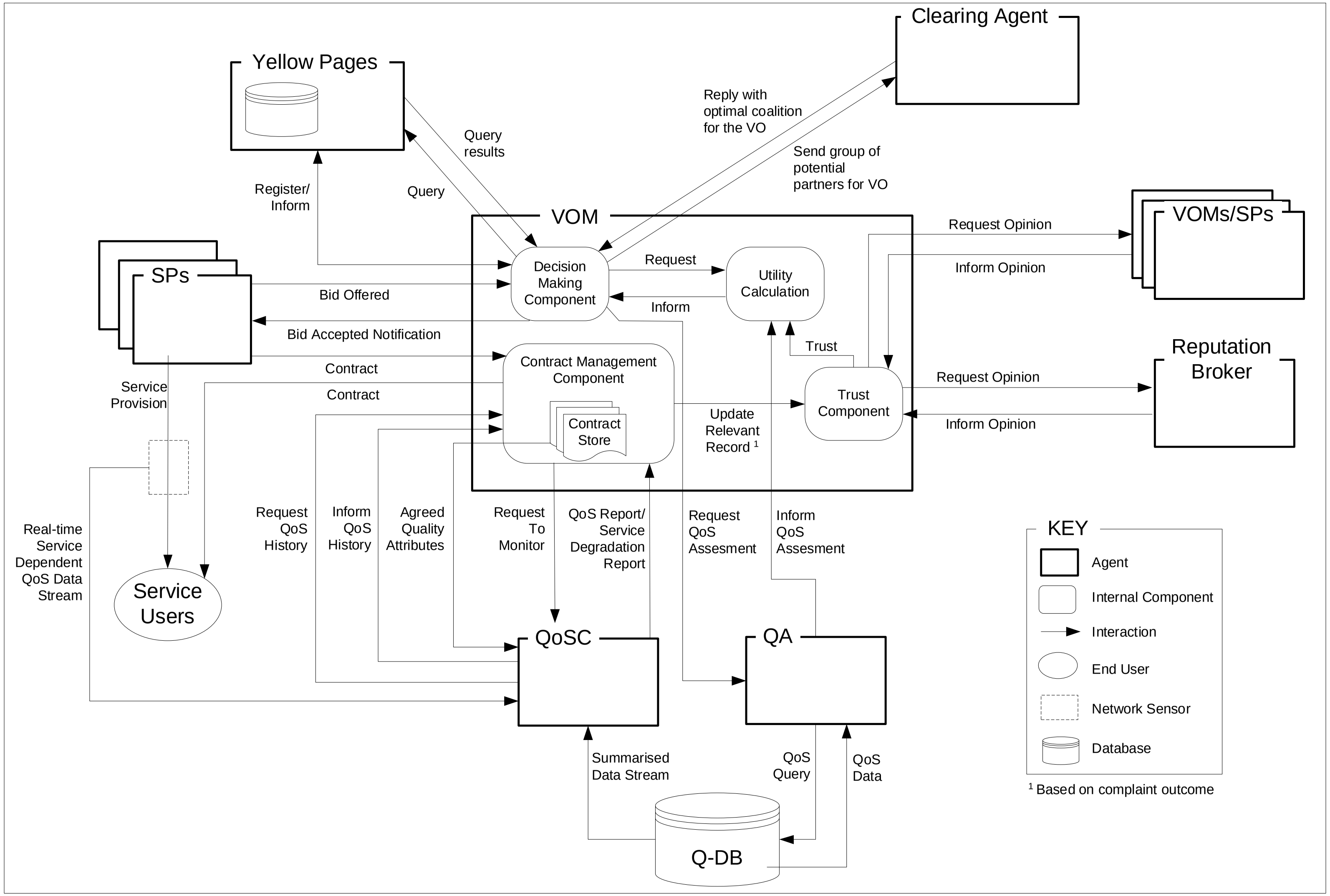

FIGURE 4.4: CONOISE-G Architecture

3. **Resource Allocation** – Once we have an expected utility function for each bidder, we employ the Clearing Agent (CA), which finds the optimal resource allocation[4] for the set of bidding SPs (Dang and Jennings, 2002). The resulting allocations are reported back to the VOM, which then sends 'hired' messages to each of the successful bidders, informing them of the quantity of each resource they are asked to provide.

Once the VO is formed, the operational phase begins. During this stage, the VOM may request the QoS Consultant (QoSC) to monitor any services provided by any members of the VO. The QoSC informs the VOM if and when an SP diverges from its agreed service level. When the QoS provision of a service in the VO falls below an acceptable level of service, or some breach of contract is observed, the QoSC alerts the VOM, which initiates a VO re-formation process. During this stage, the Contract Management component of the VOM, decides whether a breach of contract has actually occurred; and if so, which SP is to blame. Based on this result, the VOM updates its trust component, recording either a successful of unsuccessful outcome for any terminated contracts.

Meanwhile, the VOM issues another message to the YP requesting a list of SPs that can replace the resources of the failed SP. As before, the YP identifies possible SPs, bids are received and evaluated, resulting in the CA determining the best SP to replace the failed provider. At this point, the VOM re-forms the VO with the new SP replacing the old one, and instructs the QoSC to stop monitoring the old SP and to monitor the new one instead. A similar process may also take place if another SP, not currently in the VO, sends the VOM a competitive offer on resources it receives from current VO members. This process is facilitated by a publish and subscribe service offered by the YP: the VOM may register interest in SPs that provide particular resources, in response to which the YP will inform the VOM any time a new SP offering such services appears in the system.

4.3.2 Walk-through Scenario

This section provides an agent-based VO scenario in which we demonstrate the use of TRAVOS. We begin by stating that there is a need to create a VO to meet a specific requirement to provide a composite multimedia communication service to an end user. This consists of the following basic services: text messaging, HTML content provision and phone calls (this example is taken from Norman et al. (2003)). Now, assume agent a_1 has identified this need and wishes to capitalise on the market niche. However, a_1 only has the capability to provide a text messaging service. It can only achieve its goal by forming a VO with an agent that can supply a service for phone calls and one

[4] Alternatively, if there are significant time constraints, the CA can find an allocation which is within some bound of optimal.

for HTML content. For simplicity, we assume that each agent in the system has the ability to provide only one service. Agent a_1 is aware of three agents that can provide a phone call service, and its interaction history with these is shown in Table 4.1. Similarly, it is aware of three agents that are capable of providing HTML content, and its past interactions with these entities are given in Table 4.2. We also assume that a trusters prior parameter distribution for all agents is uniform:

$$\alpha^{prior} = 1, \quad \beta^{prior} = 1$$

Agent	**Past interactions**	
	Successful	**Unsuccessful**
a_2	17	5
a_3	2	15
a_4	18	5

Table 4.1: Agent a_1's interaction history with phone call service provider agents.

Agent	**Past interactions**	
	Successful	**Unsuccessful**
a_5	9	14
a_6	3	0
a_7	18	11

Table 4.2: Agent a_1's interaction history with HTML content service provider agents.

Agent a_1 would like to choose the most trustworthy phone call and HTML content service provider from the selection. The following describes how this is achieved using TRAVOS.

4.3.2.1 Calculating Trust

Using the information from Tables 4.1 and 4.2, a_1 can determine the number of successful interactions n, and the number of unsuccessful interactions m, for each agent it has interacted with. Feeding these into Equations 4.9 and 4.10, a_1 can obtain a parameter distribution which summarises each agent's likely behaviour in future interactions; for example, the shape parameters α and β, for a_2, are calculated as follows:

Using Table 4.1: $n_{a_1,a_2} = 17$, $m_{a_1,a_2} = 5$.

Using Equations 4.9 & 4.10: $\alpha = 17 + 1 = 18$ *and* $\beta = 5 + 1 = 6$.

The hyperparameter for each agent is then used estimate the probability that each agent would cooperate on any future interaction. In line with Section 4.1.2, we calculate this estimate as the expected value of the parameter distribution (Equation 4.5); for example, the estimate, ϑ_{a_1,a_2}, for a_2 is calculated as follows:

Using Equation 4.5: $\vartheta_{a_1,a_2} = \frac{\alpha}{\alpha+\beta} = \frac{18}{18+6} = 0.75$.

The above estimate gives a_1 an assessment of a_2's likely behaviour based on direct interactions. However, as discussed in Section 4.2, a_1 may wish to determine if the accuracy of this estimate is sufficient to avoid the need to gather reputation. To do this, we calculate the posterior probability that the true value for θ_{a_1,a_2} lies within an acceptable margin of error around the estimate. We can calculate this using the parameter distribution as follows. First, we decide on an acceptable error margin, $\vartheta a_1 a_2 \pm \epsilon$, where ϵ is a suitable value, such as 0.2. Second, integrate the parameter distribution over the area define by the error margin. Finally, we decide upon some threshold for this probability, above which we decide the estimate as an acceptable level of accuracy; for example, we could define a threshold τ as 0.95. The proceeding example illustrates this calculation for a_1's estimate for a_2, using $\epsilon = 0.2$; we denote the resulting confidence value as γ_{a_1,a_2}.

$$\gamma_{a_1,a_2} = \frac{\int_{\vartheta_{a_1,a_2}+\epsilon}^{\vartheta_{a_1,a_2}-\epsilon} B^{\alpha-1}(1-B)^{\beta-1}dB}{\int_0^1 U^{\alpha-1}(1-U)^{\beta-1}dU} = \frac{\int_{0.95}^{0.55} B^{\alpha-1}(1-B)^{\beta-1}dB}{\int_0^1 U^{\alpha-1}(1-U)^{\beta-1}dU} = 0.98$$

Agent	α	β	ϑ_{a_1,a_x}	γ_{a_1,a_x}
a_2	18	6	0.75	0.98
a_3	3	16	0.16	0.98
a_4	19	6	0.76	0.98
a_5	10	15	0.40	0.97
a_6	4	1	0.8	0.87
a_7	19	12	0.61	0.98

TABLE 4.3: Agent a_1's calculated trust and associated confidence level for HTML content and phone call service provider agents.

The hyperparameters, estimate and associated confidence for each agent, a_2 to a_7, which a_1 computes using TRAVOS, are shown in Table 4.3. From this, it is clear that the trust values for agents a_2, a_3 and a_4, all have a confidence above τ (=0.95). This means that a_1 does not need to consider the opinions of others for these three agents. Agent a_1 is able to decide that a_4 is the most trustworthy out of the three phone call service provider agents and chooses it to provide the phone call service for the VO.

4.3.2.2 Calculating Reputation

The process of selecting the most trustworthy HTML content service provider is not as straightforward. Agent a_1 has calculated that out of the possible HTML service providers, a_6 has the highest trust value. However, it has determined that the confidence it is willing to place in this value is 0.87, which is below that of τ and means that a_1 has not yet interacted with a_6 enough times to calculate a sufficiently confident trust value. In this case, a_1 has to use the opinions from other agents that have interacted with a_6, and form a reputation value for a_6 that it can compare to the trust values it has calculated for other HTML providers (a_5 and a_7).

Lets assume that a_1 is aware of three agents that have interacted with a_6, denoted by a_8, a_9 and a_{10}, whose opinions about a_6 are $(15, 46)$, $(4, 1)$ and $(3, 0)$ respectively. Agent a_1 can obtain hyperparameters based solely on the opinions provided as follows.

Opinions from providers: $a_8 = (15, 46), a_9 = (4, 1)$ and $a_3 = (3, 0)$

Using Equations 4.33 & 4.34: $N = 15 + 4 + 3 = 22, \quad M = 46 + 1 + 0 = 47$

Using Equation 4.31: $\alpha = 22 + 1 = 23, \quad \beta = 47 + 1 = 48$

Having obtained the shape parameters, a_1 can obtain an estimate for a_6 using Equation 4.5, as follows:

Using Equation 4.5: $\vartheta_{a_1,a_6} = \frac{\alpha}{\alpha+\beta} = \frac{23}{23+48} = 0.32$

Now a_1 is able to compare the trust in agents a_5, a_6 and a_7. Before calculating the trustworthiness of a_6, agent a_1 considered a_6 to be the most trustworthy (see Table 4.3). Having calculated a new trust value for agent a_6 (which is lower than the first assessment), agent a_1 now regards a_7 as the most trustworthy. Therefore a_1 chooses a_7 as the service provider for the HTML content service.

4.3.2.3 Handling Inaccurate Opinions

The method a_1 uses to assess the trustworthiness of a_6, as described in Section 4.3.2.2, is susceptible to errors caused by reputation providers giving inaccurate information. In our scenario, suppose a_8 provides the HTML content service too, and is in direct competition with a_6. Agent a_1 is not aware of this fact, which makes a_1 unaware that a_8 may provide inaccurate information about a_6 to influence its decision on which HTML content provider agent to incorporate into the VO. If we look at the opinions provided by agents a_8, a_9 and a_{10}, which are $(20, 46)$, $(4, 1)$ and $(3, 0)$ respectively, we can see that the opinion provided by a_8 does not correlate with the other two. Agents a_9 and a_{10} provide a positive opinion of a_6, whereas agent a_8 provides a very negative opinion. Suppose that a_8 is providing an inaccurate account of its experiences with a_6. We can use the mechanism discussed in Section 3.3 to allow a_1 to cope with this inaccurate information, and arrive at a better decision that is not influenced by self-interested reputation providing agents (such as a_8).

Before we show how TRAVOS can be used to handle inaccurate information, we must assume the following. Agent a_1 obtained reputation information from a_8, a_9 and a_{10} on several occasions, and each time a_1 recorded the opinion provided by a reputation provider and the actual observed outcome (from the interaction with an agent to which the opinion is applied). Each time an opinion is provided, the outcome observed is

Agent	**Weighting**	**Adjusted Values**			
		μ	σ	α	β
a_8	0.0039	0.5	0.29	1.0091	1.0054
a_9	0.78	0.65	0.15	5.8166	3.1839
a_{10}	0.74	0.62	0.17	4.3348	2.6194

Table 4.4: Agent a_1's adjusted values for opinions provided by a_8, a_9 and a_{10}.

	$[0, 0.2]$		$[0.2, 0.4]$		$[0.4, 0.6]$		$[0.6, 0.8]$		$[0.8, 1]$		**Total**
	n	**m**	**n**	**m**	**n**	**m**	**n**	**m**	**n**	**m**	
a_8	2	0	11	4	0	0	0	0	2	3	25
a_9	0	2	1	3	0	0	22	10	6	4	30
a_{10}	1	3	0	2	0	0	18	8	5	3	25

Table 4.5: Observations made by a_1 given opinion from a reputation source. n represents that the interaction (to which the opinion applied) was successful, and likewise m means unsuccessful.

recorded by updating a frequency bin corresponding to the interval $\Theta_r^{\mathcal{C}}$, which the received opinion belongs to. Agent a_1 keeps information of like opinions in bins as shown in Table 4.5. For example, if a_8 provides an opinion that is used to obtain a trust value of 0.3, then the actual observed outcome (successful or unsuccessful) is stored in the $0.2 < E[\theta_{a_{tr},a_{te}}|\phi_r] \leq 0.4$ bin.

Using the information shown in Table 4.5, agent a_1 can calculate the weighting to be applied to the opinions from the three reputation sources by applying the technique described in Section 3.3.1. In so doing, agent a_1 uses the information from the bin, which contains the opinion provided, and integrates the beta distribution between the limits defined by the bin's boundary. For example, a_8's opinion falls under the $0.2 < E[\theta_{a_{tr},a_{te}}|\phi_r] \leq 0.4$ bin. In this bin, agent a_1 has recorded that $n = 15$ and $m = 3$. These n and m values are used to obtain a beta distribution, $d(\theta_{a_{tr},a_{te}}|\phi_o)$, which is then integrated between 0.2 and 0.4 to give a probability of accuracy $\rho_{a_1,a_8} = 0.0039$ for a_6's opinion. Then, by using Equations 4.35 and 4.36, agent a_1 can calculate the adjusted mean and standard deviation of the opinion, which in turn gives the adjusted α and β parameters for that opinion. The results from these calculations are shown in Table 4.4.

Summing the adjusted values for α and β from Table 4.4, a_1 can obtain a more reliable value for the trustworthiness of a_6. Using Equation 4.5, a_1 calculates an estimate $\vartheta_{a_1,a_6} = 0.62$ for a_6. This means that from the possible HTML content providers, a_1 now sees a_6 as the most trustworthy and selects it to be a partner in the VO. Unlike a_1's decision in Section 4.3.2.2 (when a_7 was chosen as the VO partner), here we have shown how a reputation provider cannot influence the decision made by a_1 by providing inaccurate information.

4.4 Empirical Study

In this section, we demonstrate the advantages that TRAVOS offers to the state of the art, through empirical evaluation. We divide our discussion into three parts. First, Section 4.4.1 describes the simulation environment and overall methodology used to perform our experiments. Second, Section 4.4.2 compares the reputation component of TRAVOS to the Beta Reputation System (BRS) (see Sections 2.2.2.2 & 2.2.3.1 for more detail). We have chosen this model as a benchmark, because it shares the same basic representation of trust as TRAVOS. Any difference in performance can therefore be attributed to the novel properties of TRAVOS, rather than those it shares with the earlier system. Finally, Section 4.4.3 investigates the component performance of TRAVOS; that is, how TRAVOS performs when a truster uses both its direct experience of a trustee and reputation, and when it uses either source of evidence in isolation. This allows us to show how TRAVOS behaves when different types of information are available, and that using both types of information is in general better than using one or the other independently.

4.4.1 Experiment Methodology

Evaluation of TRAVOS took place using a simulated marketplace environment, consisting of three distinct sets of agents: provider agents $\mathcal{P} \subset \mathcal{A}$, consumer agents $\mathcal{C} \subset \mathcal{A}$, and reputation source agents $\mathcal{S} \subset \mathcal{A}$. For our purposes, the role of any $c \in \mathcal{C}$ is to evaluate $\vartheta_{c,p}$ for all $p \in \mathcal{P}$. The behaviour of each provider and reputation source agent is set before each experiment. Specifically, the behaviour of a provider $p_1 \in \mathcal{P}$ is determined by the parameter θ_{c,p_1} as described in Section 3.1. Here, reputation sources are divided into three types that define their behaviour: *accurate* sources report the number of successful and unsuccessful interactions they have had with a given consumer without modification; *noisy* sources add gaussian noise to the beta distribution determined from their interaction history, rounding the resulting expected value if necessary to ensure that it remains in the interval $[0, 1]$; and *lying* sources attempt to maximally mislead the consumer by setting the expected value $E[\theta_{c,p}|\phi_r]$ to $1 - E[\theta_{c,p}|\phi_r]$.

Against this background, all experiments consisted of a series of episodes in which a consumer was asked to assess its trust in all providers $\mathcal{P}$. Based on these assessments, we calculate the consumer's mean estimation error for the episode (Equation 4.44). This gives us a measure of the consumer's performance on assessing the provider population as a whole. The value of this metric will vary depending on the distribution of values of $\theta_{c,p}$ over the provider population. For simplicity, all the results described in the next sections have been acquired for a population of 101 providers with values of $\theta_{c,p}$ chosen

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

Table 4.6: Reputation Source Populations

uniformly between 0 and 1 at intervals of 0.01.

$$avg\_estimate\_err = \frac{1}{N}\sum_{i=1}^{n} abs(\vartheta_{c,p_i} - \theta_{c,p_i}) \tag{4.44}$$

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

4.4.2 TRAVOS Against the Beta Reputation System

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

To show variation in performance depending on reputation source behaviour, we ran experiments with populations containing accurate and lying reputation sources, and populations containing accurate and noisy sources. In each case, we kept the total number of sources equal to 20, but ran separate experiments in which the percentage of accurate sources was set to 0%, 50% and 100% (see Table 4.6). Now figure 4.5 shows the mean estimation error of TRAVOS and BRS with these different reputation source populations averaged over 50 independent episodes in each experiment. To provide a benchmark, the figure also shows the mean estimation error of a consumer $c_{0.5}$, which keeps $\vartheta_{c_{0.5},p} = 0.5$ for all $p \in \mathcal{P}$. Results are plotted against the number of previous interactions that have occurred between the consumer and each reputation source.

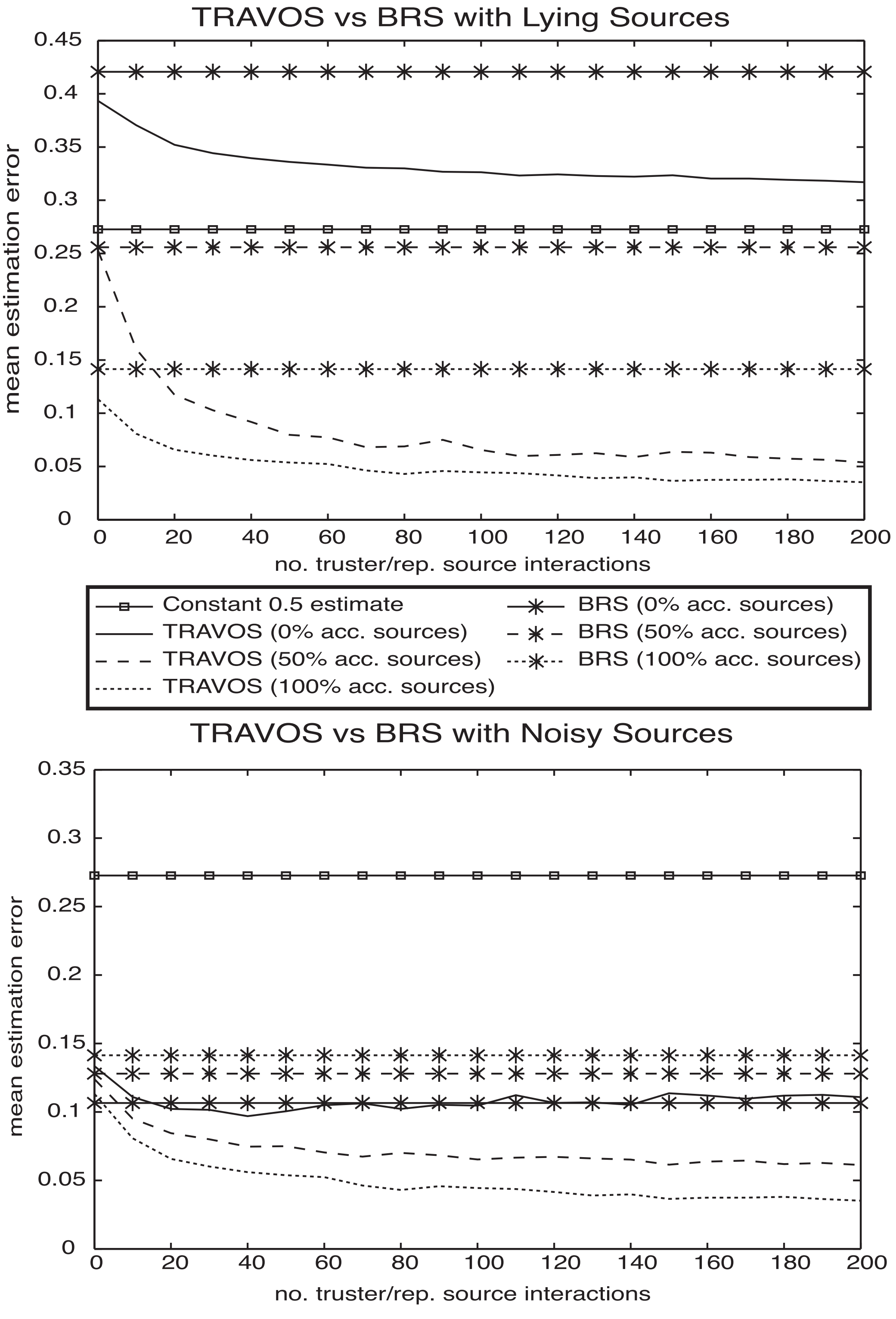

Figure 4.5: TRAVOS Reputation System vs BRS

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

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

4.4.3 TRAVOS Component Performance

To evaluate the overall performance of TRAVOS, we compared three versions of the system that used the following information respectively: direct interactions between the consumer and providers; direct provider experience and reputation; and reputation information only. In these experiments, we varied the number of interactions between the consumers and providers, and kept the number of consumer to reputation source interactions constant at 10. We used the same reputation source populations as described in Section 4.4.2. The mean estimation errors for a subset of these experiments are shown in Figure 4.6. Using only direct consumer to provider experience, the mean estimation error decreases as the number of consumer to provider interactions increases. As would be expected, using both information sources when the number of consumer to provider interactions is low, results in similar performance to using reputation information only. However, in some cases, the combined model may provide marginally worse performance than using reputation only.[5] This can be attributed to the fact that TRAVOS will always put more faith in direct experience than reputation.

With a population of 50% lying reputation sources, the combined model is misled enough to temporarily increase its error rate above that of the direct only model. This is a symptom of the relatively small number of consumer to reputation source interactions

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

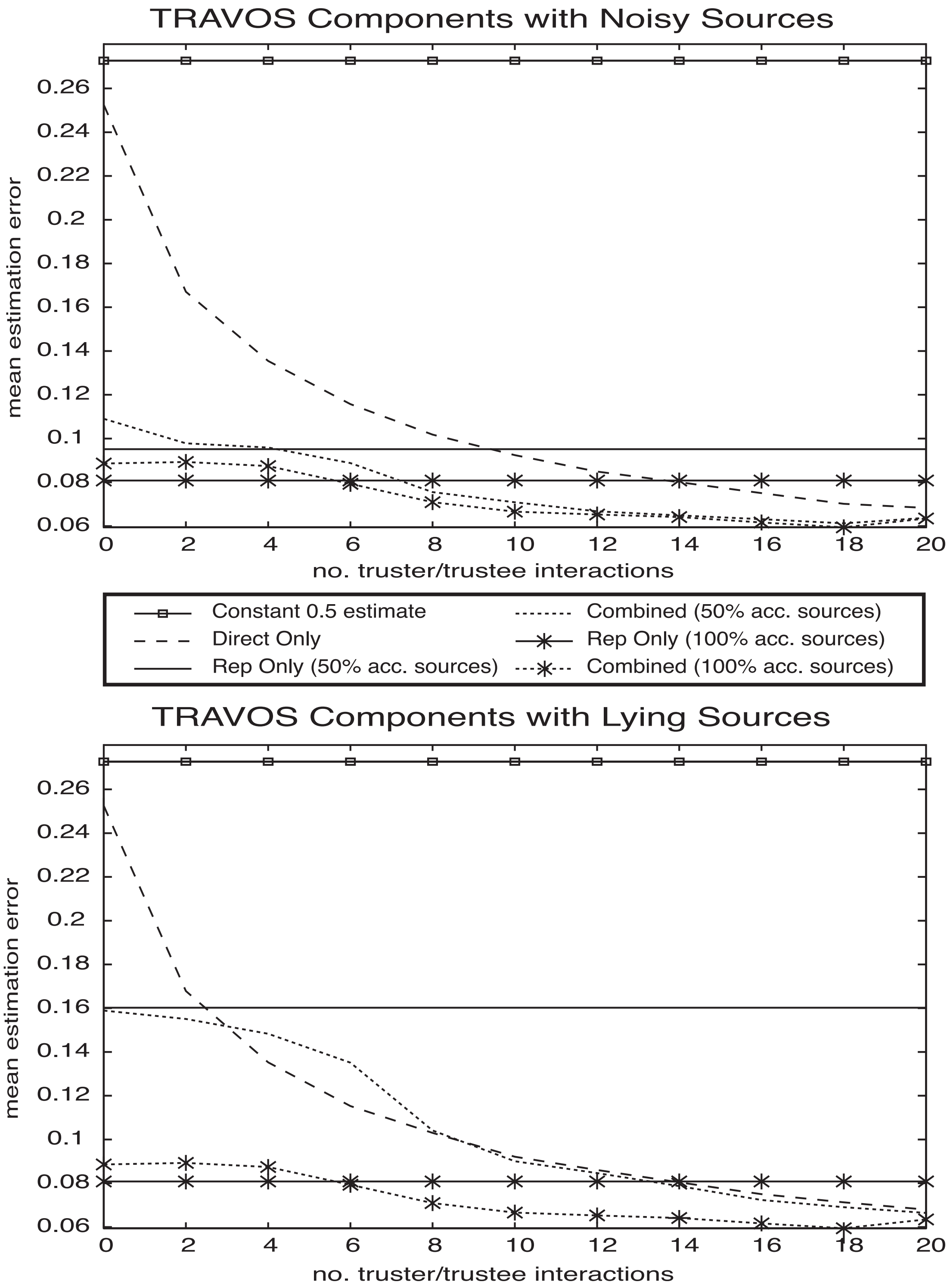

Figure 4.6: TRAVOS Component Performance

(10), which is insufficient for the consumer to completely discount all the reputation information as unreliable. The effect disappears when the number of such interactions is increased to 20. However, these results are not illustrated graphically here.

Chapter 5

Future Work and Conclusions

5.1 Current Research Contribution

In this section we consider the main contributions to the state or the art, which are presented in this document. With respect to the requirements set out in Section 1.5, our trust model TRAVOS (Chapter 4), together with the framework (Chapter 3), provide at least a partial solution to many of the outlined conditions. However, many existing trust models also satisfy a significant number of these requirements, at least to some extent. Therefore, we concentrate our discussion here to those requirements to which we provide a significant advancement over competing models. In particular, we look at trust representation (Requirement 1.5), and reputation accuracy assessment (Requirement 3.1). To contrast our work against existing systems, we also consider three other models which, in our view, are representative of the status of the art: REGRET (Section 2.2.2.1), the Beta Reputation System (BRS) (Sections 2.2.2.2 and 2.2.3.1) and Yu and Singh's model (Section 2.2.2.3 and 2.2.3.1). We divide the section into two subsections: (1) Section 5.1.1 considers trust representation issues and (2) Section 5.1.2 considers reputation issues.

5.1.1 Trust Representation

To show that our work satisfies the representation requirement (Requirement 1.5), we show that we provide a solution to each of its three subrequirements. First, in Chapter 3, we show how the uncertainty surrounding a trustee's behaviour can be represented using two separate probability distributions: (1) the probability distribution of a trustee's behaviour during an interaction with the truster, denoted as $b(x \in \mathcal{O}^{\mathcal{C}}|\theta_{a_{tr},a_{te}})$; and (2) the distribution of the parameter $\theta_{a_{tr},a_{te}}$, which models the uncertainty surrounding the true distribution of the trustee's behaviour. In effect, the behaviour distribution $b(x \in \mathcal{O}^{\mathcal{C}}|\theta_{a_{tr},a_{te}})$ models the intrinsic uncertainty surrounding the trustee's behaviour, while the parameter distribution represents the evidential uncertainty; thus we satisfy Requirement 1.5.1.

Second, we satisfy the grounding requirement (Requirement 1.5.2) by applying Bayesian Analysis. Specifically, in Section 3 we show how standard statistical theory can be used to estimate of a trustee's behaviour distribution, given a set of observations from past interactions with the trustee. In Section 4.1, we show how this general approach can be applied to cases where the trustee's behaviour is treated as a binary event: either the trustee cooperates, fulfilling its obligations to the truster, or it defects.

Third, our model does not include any parameters for which there is no obvious way to choose a reasonable value (Requirement 1.5.3). The choice of statistical models used to represent trustee behaviour (an example of which is found in Section 4.1), the definition of the reputation adjustment function (Section 3), and the choice of loss function (Section 3.1), can be made intuitively by considering the properties of the target domain. Furthermore, the size of interval used to split up the parameter distribution (Section 3.3.1) is a trade off between margin of error deemed acceptable for a reputation source's opinion, and how fast to model learns to disregard in accurate reputation sources.

The Beta Reputation System satisfies Requirement 1.5, for similar reasons, since it shares the same basic representation of trust as TRAVOS. However, through our framework, we provide a clear path to developing probabilistic models of trust in non-binary cases; the Beta Reputation System on the other hand, is limited to binary cases. In comparison to this, Yu and Singh's model does not satisfy Requirement 1.5 for the reasons stated in Section 2.2.2.3. Similarly, the REGRET system suffers from its requirement for a number of parameter settings, with no obvious way to choose reasonable values (see Section 2.2.2.1). For instance, the way in which REGRET measures evidential uncertainty (Equation 2.1) requires a threshold to be specified for the number of trustee observations, above which it is assumed their is no evidential uncertainty. In is unclear how this threshold should be determined.

This problem is compounded when complete uncertainty is calculated, and when reputation is taken into account. Complete uncertainty is calculated as a weighted average of intrinsic uncertainty and evidential uncertainty. Similarly, when reputation is accounted for, complete uncertainty is calculated as a weighted average of the complete uncertainty measurements reported by each reputation source and the truster itself. Again, it is unclear how these weights should be determined. Moreover, the validity is this approach is open to question. Intuitively, evidential uncertainty should decrease monotonically as evidence increases, a condition which is not upheld by the weighted mean approach. In contrast, TRAVOS has clear justifications for its uncertainty representations, grounded in statistical theory.

One additional advantage also results from adopting a probabilistic approach is that we provide a clear path toward guiding decision making in two ways. First, probability theory underpins decision theory (Russell and Norvig, 2003b), in which it is stated that a rationale agent should always make a decision such that it maximises its expected utility. Expected utility calculations rely upon the existence of a probability distribution over the possible states of the world that effect the actual utility a decision maker receives. In the context of trust, a trustee's behaviour constitutes at least part of the state information relevant to decision making. In this respect, both the behaviour distribution, and the parameter distribution can potentially take part in expected utility calculations (an example of how the behaviour distribution can be used in this way is given in Section 4.3.1). Finally, the parameter distribution also summarises how much the truster's current knowledge decreases the uncertainty in the trustee's behaviour. In this way, it can be used to decide when extra evidence is required to make a reasonable decision; a simple method for doing this is presented in Section 4.2.

5.1.2 Reputation

Then main problem with using reputation to assess trust is that reputation sources may be unreliable. In Sections 3.3 and 4.1.3 we present a solution to Requirement 3.1 which has the following key properties.

Statistical grounding — When all a truster's reputation sources are considered accurate, a truster's assessment is the same as if it directly observed all the interactions reported by its reputation sources. This gives a sound justification for this particular combination scheme, which is shared by the Beta Reputation System (although as stated in the preceding section, we also offer a path toward non-binary action spaces). Additionally, when some reputation sources are not completely accurate, the result is as if all such sources made a smaller number of more conservative observations. Thus, we reduce the effect of unreliable sources on the final result.

Accuracy assessed on individual basis — Many existing mechanisms for handling inaccurate reputation, including the BRS, assume that the majority of reputation sources are accurate. One important case where this assumption does not hold is when no agents are familiar with a trustee. In this case, the only opinions that will be reported, will be from agents with an incentive to mislead the truster. Yu and Singh's solution is similar to ours, in that it compares past opinions given by a particular source to subsequent observations of trustee behaviour. In general, this approach does not rely on the majority accurate assumption, and so provides a significant advance over methods which do. However, as stated in the previous section, Yu and Singh's model does not additionally satisfy Requirement 1.5.

5.2 Future Work

In our future work, we wish to build upon the assessment capability of our current system in the following four ways. First, we plan to develop methods for assessing a trustee based on the behaviour of similar agents, especially when there is little direct information about the trustee available. Second, we plan to enhance the mechanism by which a truster decides if it needs to seek reputation about a trustee (see Section 3.1). Third, we plan to extend TRAVOS so that it can handle non-binary representations of a trustee's behaviour. We discuss each of these in more detail in the subsections below.

5.2.1 Group Behaviour Priors

In Section 3.1, we mention that a truster's assessment of trustee is, not only based on observations of the trustee's behaviour, but also on the preconceptions a truster has about a trustee. In line with Bayesian Analysis, these preconceptions are summarised by a prior distribution over the trustee's possible behaviour. However, we did not specify how a truster should come about such a prior. One way to do this, would be to consider observations of other agents that are somehow similar to the trustee under consideration. Obviously, such observations will not be as informative as observations of the trustee itself, so we can not simply treat them as if they were. Instead, we plan to find a suitable method for forming the prior, which accounts for the amount of evidence we have about agents similar to a trustee, and the variance in the behaviour that occurs between such agents. Clearly, both these factors should influence the effect of such evidence on our assessment of a trustee: the amount of evidence we have about a group tells us how much we actually know about the group's behaviour, while the variance tells us how similar we should expect the behaviour of the trustee to be to the average behaviour of the group. Although other models such as REGRET already provide mechanisms for considering group behaviour along with individual behaviour, these mechanisms generally require the specification of large numbers of arbitrary parameters thus breaking Requirement 1.5.3. We believe we can over come this downfall by stronger use of statistical techniques, just as we have done in our current work.

5.2.2 Assessment of Truster Knowledge

In Section 4.2.2, we describe how a truster can decide if it has sufficient knowledge to judge a trustee, by calculating the probability that the true behaviour of the trustee lies within some bound of the truster's estimate. However, we do not specify how high this bound should be, or how high the probability value should be to indicate a sufficient amount of information. To rectify this limitation, we need to quantify the accuracy required by the truster, during the decision making process of which trustee assessment is part. To do this, we assume that when an agent must make a decision, it is presented with a finite number of alternative actions, of which it must choose one. We further assume that an agent makes this choice using decision theory, by choosing the option which maximises its expected utility. For example, when a truster must choose between several competing service providers it performs the following three steps: (1) using TRAVOS, estimate the probability distribution for each service provider's behaviour; (2) based on this estimate, calculate the expected utility for each possible choice of service provider; (3) choose the service provider which maximises expected utility.

Unfortunately, since we can only estimate the true behaviour distribution of a service provider, we can only estimate expected utility. Thus, there remains a risk that a truster will not choose the best service provider. Our intention is to quantify this risk by considering the evidence we have about all competing service provider's as a whole. In particular, we wish to identify, which, if any, of the competing service providers the truster does not know enough about to make a sound decision. We can then use this as a basis for deciding when to seek reputation, and which agents to seek reputation for.

5.2.3 A Trust Model for Continuous Action Spaces

A binary representation of a trustee's behaviour is appropriate when a truster only cares about whether or not a trustee fulfils its obligations. However, if there is some notion of *how well* a trustee fulfils its obligations that affects a trusters utility, then non-binary cases should be considered. Therefore, it is our intention to extend the TRAVOS model to include a non-binary instantiation of our framework.

5.2.4 Implications of Reputation in Group Learning

An important implication of trust assessment is that a truster will generally choose to interact with agents which, according to the knowledge of the truster, provide better than average performance. Although this seems reasonable, it raises the possibility that a small number of service providers could quickly gain a monopoly position for certain types of service: new agents entering a system may never get a foothold in the market, because no clients will be willing to take a chance on unknown entities (Requirement 1.6). In human society, this problem is solved by exploration. Although people may generally stick with suppliers that they know, they may occasionally take a risk with a new supplier to judge its performance. In machine learning, such exploration usually falls under the domain of Reinforcement Learning (Sutton and Barto, 1998), which traditionally considers the problem of individual learners exploring their environment. Recently however, research in Reinforcement Learning has progressed to consider groups of learning entities. Generally, this type of work considers one of two types of problem: (1) agents are self-interested entities, which attempt to learn about each other's behaviour in a competitive environment (Tran and Cohen, 2004); (2) agents are co-operative members of a team, which attempt to increase group knowledge efficiently, by coordinating their actions (Dutta et al., 2004). In the former case, agents do not generally share the knowledge that they learn. In the latter, agents do share knowledge, however they assume that all such knowledge is expressed truthfully.

In our view, agents which share reputation information, effectively bridge the gap between these two types of problem. To some extent, trusters are self-interested agents which attempt to learn about the behaviour of other agents, to choose the best interaction partners. To achieve this however, trusters may share information they have about their piers in the form of reputation. This is therefore a cooperative learning problem, with the complication that reputation cannot be assumed to be accurate. With this in mind, we plan to investigate the use of our current work in combination with reinforcement techniques.

5.3 Summary

In this document, we review the current state of the art in automated trust assessment techniques, and identify several key requirements that such techniques should satisfy. In Section 2.3 we highlight three of these that warrant further investigation: (1) assessing reputation source accuracy, which is required when a truster judges a trustee based on third party opinions; (2) combining different types of evidence, which is required when there is not enough of any one type of evidence for a truster to assess a trustee; (3) exploration of trustee behaviour, which involves taking a calculated risk and interacting with certain agents, to better assess their true behaviour.

Our current research has focused on the first of these problems. Specifically, we have developed a probabilistic framework for assessing trust based on direct observations of a trustee's behaviour and indirect observations, made by a third party (reputation). A significant contribution of this framework is that it provides a well founded mechanism for estimating reputation source accuracy. We have instantiated this framework for cases in which a trustee's behaviour is representation as a binary event (for example, cooperate or defect) and, through empirical evaluation, we show that this is robust against inaccurate reputation sources. In future work, we intend to enhance the assessment capability of our current system, in particular addressing the final two issues identified above.

Appendix A

Statistical Concepts

Underpinning much of the work in this document are some standard techniques from statistical estimation, and bayesian analysis in particular. Thus, knowledge of these techniques is beneficial to understanding our work. For this reason, we now give a brief overview of some of the basic statistical techniques we use in the previous chapters. A more detailed treatment of this topic can be found in Lee (2004); DeGroot and Schervish (2002b).

A common problem in statistics is to estimate the (population) probability distribution of a random variable; that is, for some random variable X, we wish to know the probability that X will take on a given value from its domain of possible values. Although, generally speaking, we do not know the true distribution, we may be able to observe a *sample* of values drawn from that distribution. Usually, we assume that each sample value has been independently drawn from the same distribution, which is commonly referred to as the i.i.d. assumption (independently drawn from the same identical distribution). This is useful because, providing it holds, the distribution of the sample will converge on the population distribution as the size of the sample increases, which simplifies the estimation process.

Given a sample drawn from a distribution under an i.i.d. assumption, there are many competing ways for estimating the population distribution; for example *maximum likelihood estimation* (DeGroot and Schervish, 2002c) and *bayes estimation* (DeGroot and Schervish, 2002d). What these techniques have in common is that they start with the assumption that the population distribution can be completely characterised by a parameter vector θ and that estimating the population distribution amounts to finding the best value for θ. The main difference between them can be seen by considering *Bayes rule* (Equation A.1).

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{A.1}$$

An ideal way to estimate θ would be to choose a value for θ that maximises $P(\theta|X)$; that is, the posterior probability of the parameter vector, given the sample data. Bayes rule tells us that this is proportional to $P(X|\theta)P(\theta)$, because $P(X)$ must be chosen such that the distribution integrates to 1. To do this directly requires that we know the prior distribution $P(\theta)$, which summarises all the information we have about θ, excluding the information provided by the sample data. In maximum likelihood estimation, we choose to ignore any prior information and instead choose an estimate that maximises the *likelihood function*, $P(X|\theta)$. This avoids the sometimes difficult issue of specifying $P(\theta)$ and works particularly well when the sample size is large, which thus reveals significant information about the population distribution.

Bayes estimation on the other hand, chooses to tackle the problem head on. Here, we do specify the prior distribution for the parameter θ, allowing us to calculate the full posterior distribution $P(\theta|X)$. Based on this parameter distribution, we then choose an estimate for θ that minimises some measure of the cost, or loss, to the statistician.

More specifically, we specify a loss function $L(\vartheta, \theta)$, and choose an estimate ϑ of θ that minimise the posterior mean of the loss function. An estimate chosen in this way is known as a *bayes estimate*. A typical choice loss function is the mean squared error (Equation A.2), which we choose to minimise the distance between the estimated and 'true' value of θ.

$$\text{mean squared error} = L(\theta, \vartheta) = (\vartheta - \theta)^2 \tag{A.2}$$

Although bayes estimation requires us to specify the prior parameter distribution, we use it to estimate the behaviour of a trustee for two reasons. First, in large multi-agent systems, the probability that any two agents have interacted a significant number of times may be quite low. This means that the number of observations, and thus the sample size may also be low. However, observations of a trustee's behaviour are not the only potential source evidence about a trustee; for example, we may look at the behaviour of other agents, similar to a trustee. We therefore believe that it is possible to specify prior distributions for the parameter, which will allow reliable predictions of trustee behaviour even if no previous observations of a trustee's behaviour are available (for an example, see Section 5.2.1). Second, the parameter distribution provides a useful summary of the amount of information we have about a trustee's behaviour, which we use in Sections 3.3 and 4.2.

Appendix B

Parameter Mapping for the Beta Distribution

In this appendix, we provide two theorems, which show how the α and β parameters of a beta distribution can be calculated, if we know the variance and mean of the distribution. Specifically, Theorem B.1 shows how α can be derived in terms of the distribution mean (denoted μ) and the variance (denoted σ^2), and then how, given this, β can be determined from α and μ. Following this, Theorem B.2, gives an alternative expression for β, in terms of σ^2 and μ only.

Theorem B.1. *Given Equations B.1 & B.2, the parameters of the beta distribution, α & β can be derived from the distribution variance (denoted σ^2) and the mean (denoted μ).*

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{B.1}$$

$$\sigma^2 = \frac{\alpha\cdot\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{B.2}$$

Proof: *First of all, we express β in terms of μ and α:*

$$\begin{aligned}
\mu &= \frac{\alpha}{\alpha+\beta} \quad \text{(from definition)} && \text{(B.3)}\\
(\alpha+\beta)\cdot\mu &= \alpha && \text{(B.4)}\\
\alpha+\beta &= \alpha/\mu && \text{(B.5)}\\
\beta &= \alpha/\mu - \alpha && \text{(B.6)}
\end{aligned}$$

Now substitute for β in equation B.2 and simplify:

$$\begin{aligned}
\sigma^2 &= \frac{\alpha(\alpha/\mu-\alpha)}{(\alpha+(\alpha/\mu-\alpha))^2(\alpha+(\alpha/\mu-\alpha)+1)} && \text{(B.7)}\\
\sigma^2 &= \frac{\alpha^2/\mu-\alpha^2}{(\alpha/\mu)^2(\alpha/\mu+1)} && \text{(B.8)}\\
\sigma^2 &= \frac{\alpha^2/\mu-\alpha^2}{(\alpha/\mu)^3+(\alpha/\mu)^2} && \text{(B.9)}\\
\sigma^2 &= \frac{\alpha^2/\mu-\alpha^2}{\alpha^3/\mu^3+\alpha^2/\mu^2} && \text{(B.10)}\\
\sigma^2 &= \frac{1/\mu-1}{\alpha/\mu^3+1/\mu^2} && \text{(B.11)}\\
\sigma^2 &= \frac{\mu^2-\mu^3}{\alpha+\mu} && \text{(B.12)}
\end{aligned}$$

Now arrange to find α:

$$\begin{aligned}
\sigma^2(\alpha+\mu) &= \mu^2-\mu^3 && \text{(B.13)}\\
\sigma^2\cdot\alpha+\sigma^2\cdot\mu &= \mu^2-\mu^3 && \text{(B.14)}\\
\sigma^2\cdot\alpha &= \mu^2-\mu^3-\sigma^2\cdot\mu && \text{(B.15)}\\
\alpha &= (\mu^2-\mu^3-\sigma^2\cdot\mu)/\sigma^2 && \text{(B.16)}\\
\alpha &= \frac{\mu^2-\mu^3}{\sigma^2}-\mu && \text{(B.17)}
\end{aligned}$$

From Equations B.6 and B.17, α and β can be expressed as follows, thus proving the theorem.

$$\alpha=\frac{\mu^2-\mu^3}{\sigma^2}-\mu,\quad \beta=\frac{\alpha}{\mu}-\alpha$$

Theorem B.2. *The β parameter of the beta distribution can be expressed only in terms of μ and σ as shown in Equation B.18. We prove this in two ways: first, by considering the properties of the beta distribution; and second, by substitution.*

$$\beta=\frac{(1-\mu)^2-(1-\mu)^3}{\sigma}-(1-\mu) \tag{B.18}$$

Proof through the properties of the Beta Distribution: *Imagine that we have two beta distributions: distribution d with parameters α and β, and distribution $\hat{d}$ with parameters $\hat{\alpha}$ and $\hat{\beta}$. Similarly, we denote the mean of $\hat{d}$ as $\hat{\mu}$ and the variance of $\hat{d}$ as $\hat{\sigma}$.*

Now assume that $\hat{\alpha}=\beta$ and $\hat{\beta}=\alpha$. From this we know that $\hat{\sigma}=\sigma$ since:

$$\frac{\alpha\cdot\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=\frac{\beta\cdot\alpha}{(\beta+\alpha)^2(\beta+\alpha+1)}=\frac{\hat{\alpha}\cdot\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)} \tag{B.19}$$

and $\hat{\mu}=(1-\mu)$ since:

$$\begin{aligned}
\hat{\mu}+\mu &= \frac{\alpha}{\alpha+\beta}+\frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} && \text{(B.20)}\\
\hat{\mu}+\mu &= \frac{\alpha}{\alpha+\beta}+\frac{\beta}{\beta+\alpha} && \text{(B.21)}\\
\hat{\mu}+\mu &= 1 && \text{(B.22)}\\
\hat{\mu} &= 1-\mu && \text{(B.23)}
\end{aligned}$$

We can now prove Equation B.18 as follows:

$$\beta = \hat{\alpha} = \frac{\hat{\mu}^2 - \hat{\mu}^3}{\hat{\sigma}} - \hat{\mu}, \quad \textit{(from Equation B.17)} \tag{B.24}$$

$$\beta = \frac{(1-\mu)^2 - (1-\mu)^3}{\sigma} - (1-\mu), \quad \textit{(by substitution)} \tag{B.25}$$

Proof by Substitution: *We now show that Equation B.18 is true by substituting Equation B.17 into Equation B.6 as follows:*

$$\beta = \frac{\alpha}{\mu} - \alpha \tag{B.26}$$

$$\beta = \left[\frac{\mu^2 - \mu^3}{\sigma} - \mu\right] / \mu - \left[\frac{\mu^2 - \mu^3}{\sigma} - \mu\right] \tag{B.27}$$

$$\beta = \left[\frac{\mu - \mu^2}{\sigma} - 1\right] - \left[\frac{\mu^2 - \mu^3}{\sigma} - \mu\right] \tag{B.28}$$

$$\beta = \frac{(\mu - \mu^2) - (\mu^2 - \mu^3)}{\sigma} - (1-\mu) \tag{B.29}$$

$$\beta = \frac{\mu - 2\mu^2 + \mu^3}{\sigma} - (1-\mu) \tag{B.30}$$

To show that Equations B.18 and B.30 are equivalent, we expand $(1-\mu)^2 - (1-\mu)^3$.

$$(1-\mu)^2 = 1 - 2\mu + \mu^2 \tag{B.31}$$

$$(1-\mu)^3 = (1 - 2\mu + \mu^2)(1-\mu) \tag{B.32}$$

$$(1-\mu)^3 = (1 - 2\mu + \mu^2) - (\mu - 2\mu^2 + \mu^3) \tag{B.33}$$

$$(1-\mu)^3 = 1 - 2\mu + \mu^2 - \mu + 2\mu^2 - \mu^3 \tag{B.34}$$

$$(1-\mu)^3 = 1 - 3\mu + 3\mu^2 - \mu^3 \tag{B.35}$$

$$(1-\mu)^2 - (1-\mu)^3 = (1 - 2\mu + \mu^2) - (1 - 3\mu + 3\mu^2 - \mu^3) \tag{B.36}$$

$$(1-\mu)^2 - (1-\mu)^3 = 1 - 2\mu + \mu^2 - 1 + 3\mu - 3\mu^2 + \mu^3 \tag{B.37}$$

$$(1-\mu)^2 - (1-\mu)^3 = \mu - 2\mu^2 + \mu^3 \tag{B.38}$$

$$\beta = \frac{(1-\mu)^2 - (1-\mu)^3}{\sigma} - (1-\mu) = \frac{\mu - 2\mu^2 + \mu^3}{\sigma} - (1-\mu) \tag{B.39}$$

Hence Equations B.18 and B.30 are equivalent and therefore Equation B.18 is true.

Bibliography

A. Abdul-Rahman and S. Hailes. A distributed trust model. In *Proceedings of the 1997 workshop on New security paradigms*, pages 48–60, Langdale, Cumbria, UK, 1997. ACM Press.

F. Adelstein, S. Gupta, R. Golden, and L. Schweibert. *Fundamentals of Mobile and Pervasive Computing*. McGraw-Hill, first edition, November 2004. ISBN 0071412379.

F. Azzedin and M. Maheswaran. Evolving and managing trust in grid computing systems. In *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 2002)*, volume 3, pages 1424–1429, May 2002a.

F. Azzedin and M. Maheswaran. Integrating trust into grid resource management systems. In *IEEE Proceedings. International Conference on Parallel Processing*, pages 47–54, August 2002b.

F. Azzedin and M. Maheswaran. Towards trust-aware management in grid computing systems. In *Cluster Computing and the Grid 2nd IEEE/ACM International Symposium (CCGRID2002)*, pages 419–424, May 2002c.

K. S. Barber and J. Kim. Belief revision process based on trust: Agents evaluating reputation of information sources. In Rino Falcone, Munindar P. Singh, and Yao-Hua Tan, editors, *Trust in Cyber-societies*, volume 2246 of *Lecture Notes in Computer Science*, pages 73–82. Springer-Verlag Heidelberg, 2001.

T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.

M. Blaze, J. Feigenbaum, and J. Lacy. Decentralized trust management. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, pages 164–173, 1996.

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

C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer Academic Publishers, 2001.

V. D. Dang and N. R. Jennings. Polynomial algorithms for clearing multi-unit single item and multi-unit combinatorial reverse auctions. In *Proceedings of the Fifteenth European Conference on Artificial Intelligence*, pages 23–27, 2002.

P. Dasgupta. Trust as a commodity. In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, chapter 4, pages 49–72. Basil Blackwell, 1988. Reprinted in electronic edition from Department of Sociology, University of Oxford, 2000.

R. K. Dash, D. C. Parkes, and N. R. Jennings. Computational mechanism design: A call to arms. *IEEE Intelligent Systems*, 18(6):40–47, 2003.

R. K. Dash, S. D. Ramchurn, and N. R. Jennings. Trust-based mechanism design. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, New York, USA, 2004.

M. DeGroot and M. Schervish. *Probability & Statistics*, chapter 6, pages 335–337. Addison-Wesley, third edition, 2002a.

M. DeGroot and M. Schervish. *Probability & Statistics*. Addison-Wesley, 3rd edition, 2002b.

M. DeGroot and M. Schervish. *Probability & Statistics*, chapter 6, pages 355–362. Addison-Wesley, third edition, 2002c.

M. DeGroot and M. Schervish. *Probability & Statistics*, chapter 6, pages 346–353. Addison-Wesley, third edition, 2002d.

Chrysanthos Dellarocas. Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems. In *Proceedings of the 21st International Conference on Information Systems*, pages 520–525, Brisbane, Australia, December 2000.

V. Deora, J. Shao, W. A. Gray, and N. J. Fiddian. A quality of service management framework based on user expectations. In *Proceedings of the First International Conference on Service Oriented Computing*, pages 104–114, 2003.

V. Deora, J. Shao, G. Shercliff, P.J. Stockreisser, W.A. Gray, and N.J. Fiddian. Incorporating QoS specifications in service discovery. In *Proceedings of Second International Web Services Quality Workshop (WQW 2004)*, 2004.

P. S. Dutta, S. Dasmahapatra, S. R. Gunn, N. R. Jennings, and L. Moreau. Cooperative information sharing to improve distributed learning. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 828–835, New York, USA, 2004.

R. Falcone, G. Pezzulo, and C. Castelfranchi. Fuzzy approach to a belief-based trust computation. In R. Falcone, S. Barber, L. Korba, and M. Singh, editors, *Trust,* *Reputation and Security: Theories and Practice*, volume 2631 of *Lecture Notes in Artificial Intelligence*, pages 73–86. Springer, 2003.

I. Foster, N. R. Jennings, and C. Kesselman. Brain meets brawn: Why grid and agents need each other. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 8–15, New York, USA, July 2004.

I. Foster and C. Kesselman, editors. *The Grid: Blueprint for a New Computing Infrastructure.* Morgan Kaufmann, 2nd edition, 2004.

I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 15(3): 200–222, 2001.

E. Friedman and P. Resnick. The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.

D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, chapter 13, pages 213–237. Basil Blackwell, 1988. Reprinted in electronic edition from Department of Sociology, University of Oxford.

D. Gollmann. *Computer Security.* John Wiley and Sons Ltd, 1998. ISBN 0-471-97844-2.

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

A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, June 2001.

A. Jøsang. Subjective evidential reasoning. In *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, Annecy, France, July 2002.

A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Conference on Electronic Commerce*, Bled, Slovenia, June 2002.

R. Jurca and B. Faltings. Towards incentive-compatiable reputation management. In R. Falcone, S. Barber, L. Korba, and M. Singh, editors, *Trust, Reputation and Security: Theories and Practice*, volume 2631 of *Lecture Notes in Artificial Intelligence*, pages 138–147. Springer, 2003.

R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley, New York, 1976.

P. M. Lee. *Bayesian Statistics: An Introduction.* Hodder Arnold, third edition, 2004. ISBN 0-340-81405-5.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

S. Marsh. *Formalising Trust as a Computational Concept.* PhD thesis, University of Stirling, 1994.

B. Misztal. *Trust in Modern Societies: The Search for the Bases of Social Order.* Polity Press, 1996.

L. Mui, M. Mohtashemi, C. Ang, P. Szolovits, and A. Halberstadt. Ratings in distributed systems: A bayesian approach. In *Proceedings of the 11th Workshop on Information Technologies and Systems*, New Orleans, USA, December 2001.

T. J. Norman, A. Preece, S. Chalmers, N. R. Jennings, M. Luck, V.D. Dang, T. D. Nguyen, V. Deora, , J. Shao, A. Gray, and N. J. Fiddian. Conoise: Agent-based formation of virtual organisations. In *Proceedings of 23rd SGAI International Conference on Innovative Techniques and Applications of AI*, pages 353–366, Cambridge, UK, 2003.

J. Patel, W. T. L. Teacy, N. R. Jennings, M. Luck, S. Chalmers, N. Oren, T. J. Norman, A. Preece, P. M. D. Gray, G. Shercliff, P. J. Stockreisser, J. Shao, W. A. Gray, N. J. Fiddian, and S. Thompson. Agent-based virtual organisations for the grid. In *Proceedings 1st International Workshop on Smart Grid Technologies*, Utrecht, Netherlands, 2005.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Cambridge University Press, 1988.

C. P. Pfleeger. *Security in Computing.* Prentice Hall, 2002. ISBN 0-130-35548-8.

S. D. Ramchurn, B. Deitch, M. K. Thompson, D. C. de Roure, N. R. Jennings, and M. Luck. Minimising intrusiveness in pervasive computing environments using multi-agent negotiation. In *Proceednings of the 1st International Conference on Mobile and Ubiquitous Systems*, pages 364–372, Boston, USA, 2004.

Stuart Russell and Peter Norvig. *Artificial Intelligence A Modern Approach.* Prentice Hall, 2nd edition, 2003a.

Stuart Russell and Peter Norvig. *Artificial Intelligence A Modern Approach*, chapter 15, pages 537–550. Prentice Hall, 2003b.

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

J. Sabater and C. Sierra. Social regret, a reputation model based on social relations. *SIGecom Exchanges, ACM*, pages 44–56, 2002. 3.1.

G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ, 1976.

J. Shao, W. A. Gray, N. J. Fiddian, V. Deora, G. Shercliff, P. J. Stockreisser, T. J. Norman, A. Preece, P. M. D. Gray, S. Chalmers, N. Oren, N. R. Jennings, M. Luck, V. D. Dang, T. D. Nguyen, J. Patel, and W. T. L. Teacy. Supporting formation and operation of virtual organisations in a grid environment. In *Proceedings of the UK OST e-Science 2nd All Hands Meeting (AHM'04)*, Nottingham, UK., September 2004.

D. Canfield Smith, A. Cypher, and J. Spohrer. Kidsim: Programming agents without a programming language. *Communications of the ACM*, 37(7):54–67, July 1994.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, 1998.

T. Tran and R. Cohen. Improving user satisfaction in agent-based electronic marketplaces by reputation modelling and adjustable product quality. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, pages 828–835, New York, USA, 2004.

S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham, C. Kesselman, T. Maguire, T. Sandholm, P. Vanderbilt, and D. Snelling. Open grid services infrastructure (ogsi). Technical report, The Globus Alliance, 2003.

G. Upton and I. Cook. "statistic". In *Dictionary of Statistics.* Oxford University Press, 2002.

Y. Wang and J. Vassileva. Bayesian network-based trust model. In *Proceedings of IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 372–378, Halifax, Canada, October 2003.

A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the 7th International Workshop on Trust in Agent Societies*, New York, USA, 2004.

M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, June 1995.

B. Yu and M. P. Singh. An evidential model of distributed reputation management. In *Proceedings of First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, volume 1, pages 294–301. ACM Press, 2002.

Bin Yu and Munindar P. Singh. Detecting deception in reputation management. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems*, pages 73–80, Melbourne, Australia, July 2003. ACM Press.

G. Zacharia, A. Moukas, and P. Maes. Collaborative reputation mechanisms in online marketplaces. In *Proceedings of 32nd Hawaii International Conference on System Sciences*, volume 8. IEEE Computer Society Press, 1999.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

L. A. Zadeh. Fuzzy logic and approximate reasoning. *Synthese*, 30:407–428, 1975.