Bootstrapping Trust Evaluations through Stereotypes

Chris Burnett
University of Aberdeen
King's College
Aberdeen, UK
cburnett@abdn.ac.uk

Timothy J. Norman
University of Aberdeen
King's College
Aberdeen, UK
t.j.norman@abdn.ac.uk

Katia Sycara
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA
katia@cs.cmu.edu

ABSTRACT

In open, dynamic multi-agent systems, agents may form short-term ad-hoc groups, such as coalitions, in order to meet their goals. Trust and reputation are crucial concepts in these environments, as agents must rely on their peers to perform as expected, and learn to avoid untrustworthy partners. However, ad-hoc groups introduce issues which impede the formation of trust relationships. For example, they may be short-lived, precluding agents from gaining the necessary experiences to make an accurate trust evaluation. This paper describes a new approach, inspired by theories of human organisational behaviour, whereby agents generalise their experiences with known partners as *stereotypes* and apply these when evaluating new and unknown partners. We show how this approach can complement existing state of the art trust models, and enhance the confidence in the evaluations that can be made about trustees when direct and reputational information is lacking or limited.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multi-Agent Systems

General Terms

Experimentation, Performance

Keywords

Trust, Stereotypes

1. INTRODUCTION

Trust is a vital concept in open and dynamic multi-agent systems, where diverse agents continually join, interact and leave. In such environments, some agents will inevitably be more trustworthy than others, displaying varying degrees of competence and self-interest in different interactions. When faced with the problem of choosing a partner with whom to interact, agents must evaluate the candidates and determine which one is the most trustworthy with respect to a given interaction and context. While the word 'trust' can denote a rich cognitive structure of beliefs [2], we define trust here pragmatically as the *degree of belief*, or subjective probability, with which a *trustor* believes a *trustee* will perform as expected when relied upon [4].

State-of-the-art trust approaches [6, 10, 16, 17] generally consider an agent's trust in a potential partner as a function of prior interactions with that partner, whether they are directly experienced or relayed by other agents in the society. The goal of these systems is primarily to maximise the accuracy of this function, and thereby minimise the number of unsatisfactory experiences an agent suffers when interacting with others. If an agent has insufficient direct evidence to form a confident evaluation of another, it can make use of *reputation* [15] in the society by obtaining the opinions of other agents who have previously interacted with it.

Initial cases exist, however, where previous direct and reputational evidence is unavailable. For example, at the beginning of a society's lifetime, no agents will have interacted before. This has been termed the "cold start problem" in recommender systems. In such cases, agents must either explore the population (e.g. by selecting random partners) or forego interaction altogether. Similarly, when a new and unknown trustee enters the system, it is generally not possible for any trustor to form an evaluation. Evidence can only be obtained when some agents opt to "take a chance" on the newcomer rather than interact with a better-known partner. We refer to these cases as the *cold-start* and *newcomer* cases respectively. In both cases, the problem is one of how to minimise the risk inherent in "bootstrapping" trust evaluations when interacting with new, unknown agents.

In this paper, we consider cases where agents interact in ad-hoc groups. These are organisational structures which form around a shared goal, and disband once that goal is satisfied. As such, they represent a recurring cold-start problem. Their life-spans are much shorter than that of the host system, and so agents may not have enough time to build a confident trust evaluation in their partners before the group dissolves. In addition, higher ratios of society size to team size will result in lower probabilities of agents encountering known partners in subsequent groups. Indeed, in dynamic systems, known agents might never be encountered again.

In order to address these issues, we propose a *stereotyping* approach, whereby agents can generalise their experiences with known partners in previous contexts in order to form tentative trust evaluations about unknown agents in new contexts. By ascribing trust evaluations to learned *classes* of individuals as well as individuals themselves, agents can make use of previous experiences and reputational opinions in contexts where this would not otherwise be possible.

Cite as: Bootstrapping Trust Evaluations through Stereotypes, C. Burnett, T. J. Norman and K. Sycara, *Proc. of 9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lespérance, Luck and Sen (eds.), May, 10–14, 2010, Toronto, Canada, pp. 241-248
Copyright © 2010, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

2. STEREOTYPES

Modern dynamic human organisations, such as temporary work groups and global virtual organisations, face similar problems with regard to trust as do their analogues in multi-agent systems. However, researchers in these fields have found that participants in temporary systems display trusting behaviours even when they have no access to the direct or reputational sources of evidence traditionally thought to be necessary for trust to form. In order to address this phenomenon, the theory of *swift* trust was developed [7, 13]. Swift trust is a tentative, probationary form of trust based on, among other things, stereotypical impressions formed about similar agents in previous contexts. Agents identify classes of partners based on their visible *features*, and use these classes to form behavioural expectations about unknown agents. Such stereotypical reasoning is held to be critical in reducing the complexity of human social decision making [12].

Agents interacting in MASs can adopt a similar stereotyping approach in order to make trust evaluations when evidence is unavailable. However, unlike human stereotypes, which are the product of affective and cultural forces [11], and often negative in nature, agents can build stereotypes which attempt to model their observations as accurately as possible. Where relationships exist between agent features and performance, a stereotyping approach can help agents to avoid the need to make a random partner selection. The assumption that agent features may provide useful predictors of future behaviour for a trust evaluation is a reasonable one. For example, an agent may learn that partners from a particular organisation tend to perform poorly in certain contexts, or that agents who have performed competently in one role can therefore be trusted in another.

To motivate our approach, consider the example of professional qualifications in human societies. Professional qualifications allow people who meet the requirements set by an awarding body to present a feature to the society signalling competence in a particular domain. When such individuals interact in the society, these signals allow partners to establish a positive, albeit tentative form of trust despite the absence of direct or reputational experiences. In this example, the degree of trust conferred on an accredited individual is a function of the degree to which the awarding body is trusted to ensure only competent individuals display the diagnostic features. In many situations, however, such prior knowledge of the significance of features may not be available. In the remainder of this paper, we examine the general case where trustors must discover for themselves the relationship between features and trustworthiness.

However, an important requirement of stereotypes is that they yield to concrete evidence about an individual when it is available. While stereotypes may provide useful estimates in initial conditions, they are based on generalisations, and should therefore carry less weight than direct evidence with an individual.

We do not attempt to capture here the rich cognitive notions of stereotype formation as it applies to humans, but rather view a stereotype as a function $S : \vec{F} \rightarrow T$ mapping feature vectors of agents, $\vec{F}$, to initial stereotypical trust estimates about those agents, T. This allows us to evaluate a model of stereotypes in the context of a general trust evaluation mechanism.

3. APPROACH

The model we propose here can be applied to any trust mechanism that uses numerical ratings to compare and exchange opinions. We demonstrate its use with a simple probabilistic model based on Subjective Logic which considers direct, reputational and stereotypical sources of information.

We assume here a society of agents, A, which comprises a set of trustors $X \subseteq A$, and trustees $Y \subseteq A$. Each trustor $x \in X$ maintains a set of *opinions* O_x, the structure of which will be defined in the following sections. In addition, we define $R_x \subseteq X$ to be the set of recommender agents visible to x, and $Y_x \subset Y$ to be the set of candidate trustees visible to x. Each agent $x \in X$ possesses a trust evaluation function $E_x(y, t, O_x, R_x)$ which returns a degree of trust for a trustee $y \in Y_x$ with respect to a task t, given a set of existing opinions held by x, O_x, and those of the visible recommenders, R_x. The behaviour of the function E_x is described in the course of this section.

Since the primary aim of our trust model is to support partner selection, we assume that, when presented with a number of potential candidates Y_x, a trustor x will always select the highest rated candidate in Y_x for a task t, denoted $C_{x:t}$, according to that trustor's evaluation function:

$$C_{x:t} = \arg\max_{y \in Y_x} E_x(y, t, O_x, R_x) \tag{1}$$

3.1 Subjective Logic

Subjective Logic (SL) [8] is a belief calculus which allows agents to express opinions as degrees of belief, disbelief and uncertainty about propositions. For binary propositions, such as "agent y is trustworthy with respect to issue t", opinions are equivalent beta probability density functions, and so they are compatible with other probabilistic trust approaches [9]. We adopt SL as a trust representation because it provides an intuitive way of capturing an agent's degree of belief in another, the degree of uncertainty about that belief, and the *a priori* belief in the absence of evidence. We will briefly outline the fundamental concepts of SL that are important to our approach, namely belief representation, evidence, probability expectation and the base rate.

3.1.1 Belief representation

An opinion held by an agent x about agent y performing a task t in Subjective Logic is represented as a tuple:

$$\omega^x_{y:t} = \langle b^x_{y:t}, d^x_{y:t}, u^x_{y:t}, a^x_{y:t} \rangle \quad \text{where } b^x_{y:t} + d^x_{y:t} + u^x_{y:t} = 1, \quad \text{and } a^x_{y:t} \in [0, 1]. \tag{2}$$

In the above opinion representation, $b^x_{y:t}, d^x_{y:t}, u^x_{y:t}, a^x_{y:t}$ represent the degrees of *belief*, *disbelief*, *uncertainty* and the *base rate* (or *a priori* degree of belief) respectively. In each case, the superscript identifies the belief owner, and the subscript represents the belief *target*, i.e. the agent and task that the opinion pertains to. Each trustor x maintains a set of such tuples, O_x, which stores its opinions about known trustees in the system.

3.1.2 Evidence aggregation

Opinions are formed on the basis of evidence aggregated from different sources which, in turn, are represented as observed frequencies of positive and negative experiences. A

rating held by an agent x about y in *evidence representation* is a pair $\langle r^x_{y:t}, s^x_{y:t}\rangle$, where $r^x_{y:t}$ is the number of positive experiences observed by x about y, and $s^x_{y:t}$ is the number of observed negative experiences. Equation 3 shows how the $r^x_{y:t}$ and $s^x_{y:t}$ parameters are used to produce an opinion:

$$b^x_{y:t} = \frac{r^x_{y:t}}{(r^x_{y:t} + s^x_{y:t} + 2)}$$
$$d^x_{y:t} = \frac{s^x_{y:t}}{(r^x_{y:t} + s^x_{y:t} + 2)}$$
$$u^x_{y:t} = \frac{2}{(r^x_{y:t} + s^x_{y:t} + 2)} \quad (3)$$

3.1.3 Probability Expectation Value

An opinion's probability expectation value can be used as a single valued trust metric, suitable for ranking potential partners. Equation 4 shows how a probability expectation value $P(\omega^x_{y:t})$ is calculated from an opinion $\omega^x_{y:t}$. We use the term *rating* to mean $P(\omega^x_{y:t})$ for a particular opinion $\omega^x_{y:t}$.

$$P(\omega^x_{y:t}) = b^x_{y:t} + a^x_{y:t} \cdot u^x_{y:t} \quad (4)$$

By using Equations 4 and 3 together, we can obtain for a given evidence pair $\langle r^x_{y:t}, s^x_{y:t}\rangle$ a probability expectation value $P(\omega^x_{y:t})$.

3.1.4 Base rate

The base rate parameter $a^x_{y:t}$ represents the *a priori* degree of trust x has about y performing task t, before any evidence has been received. It determines the effect that the uncertainty parameter $u^x_{y:t}$ will have on the resultant probability expectation value. The default value of $a^x_{y:t}$ is 0.5, which means that before any positive or negative evidence has been received, both outcomes are equally likely. This means that before any evidence has been received, $P(\omega^x_{y:t})$ = 0.5, which is the least informative value it can take. Values of $a^x_{y:t} > 0.5$ will result in more uncertainty being converted to belief, and conversely disbelief for $a^x_{y:t} < 0.5$.

The base rate parameter provides a means to incorporate the predictions of our stereotyping model back into the trust evaluation process. We model the effects of stereotypes in SL by using the model's predictions as the base rate. That is, for a given agent y, the base rate $a^x_{y:t} = S(\vec{F}_y)$. In this way, stereotypes represent our *a priori* degrees of belief. For example, when no evidence has been received for a particular trustee, we have maximum uncertainty, i.e. $\omega^x_{y:t} = (0, 0, 1, 0.5)$. In this case, $a^x_{y:t}$ alone determines the value of $P(\omega^x_{y:t})$. However, as more evidence is received, the value of $u^x_{y:t}$ decreases, and so the effect of $a^x_{y:t}$ also decreases. This satisfies our fundamental condition that a stereotype must yield to concrete evidence as it is obtained.

Due to the additivity requirement of the $b^x_{y:t}$, $d^x_{y:t}$ and $u^x_{y:t}$ parameters, the opinion spaces of agents can be visualised as a triangular (ternary) plot (Figure 1), with the top vertex representing maximum uncertainty, the bottom left representing maximum disbelief, and the bottom right representing maximum belief. The distance from the midpoint of the leftmost edge represents the degree of belief, the distance from the midpoint of the rightmost edge represents disbelief, and the distance from the bottom edge represents the degree of uncertainty.

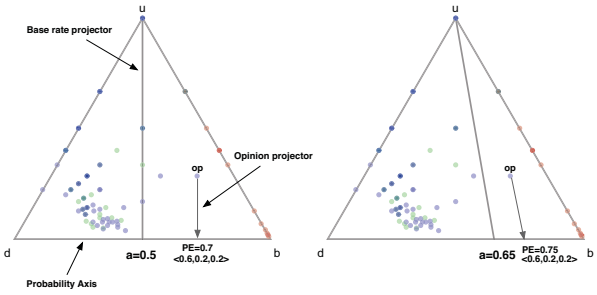

Figure 1: Sample opinion space for a trustor, with base rate projector shown.

The bottom edge of the triangle represents the classical probability axis. The base rate value $a^x_{y:t}$ is plotted here. In calculating $P(\omega^x_{y:t})$, opinions are *projected* onto this axis following a line parallel to the base rate projector line (originating at the uncertainty vertex and ending at a on the probability axis). Figure 1 shows an example opinion space with two different base rates. The leftmost opinion has $a^x_{y:t} = 0.5$ (an unbiased opinion) while the rightmost opinion has $a^x_{y:t} = 0.65$. This causes the probability expectation value for the example opinion to become shifted from 0.7 to 0.75.

3.1.5 Reputation

Reputation in probabilistic trust systems is often calculated by aggregating the $r^x_{y:t}$ and $s^x_{y:t}$ parameters from reputation providers [9, 17]. Since we make the simplifying assumption that all agents report their experiences truthfully and accurately[1], the result of the aggregation of evidence provided by a set of recommender agents R is a combined evidence pair $\langle r'^x_{y:t}, s'^x_{y:t}\rangle$ equivalent to one where the evaluating agent had observed all the aggregated experiences itself:

$$r'^x_{y:t} = r^x_{y:t} + \sum_{\rho \in R_x} r^\rho_{y:t} \qquad s'^x_{y:t} = s^x_{y:t} + \sum_{\rho \in R_x} s^\rho_{y:t} \quad (5)$$

Once the evidence parameters have been aggregated, an opinion and rating for the combined evidence can be calculated using Equations 3 and 4.

3.2 Learning Stereotypes

As we have mentioned, the goal of our stereotyping mechanism is to identify a function $S : \vec{F} \rightarrow T$, where $\vec{F}$ is an agent's feature vector (a vector of discrete values) and T is the expected probability of a good outcome (a continuous real value). When an agent has a collection of experiences with other agents described by feature vectors, we can make use of existing machine learning techniques for learning associations between sets of discrete attributes and continuous classes. Specifically, we employ the M5 model tree learning [1, 14] algorithm[2]. This algorithm shares some similarities with decision tree classifiers, in that it recursively

[1]While the issue of deception remains an open problem, some techniques for addressing this assumption have been investigated [3, 16]

[2]We use the M5 implementation of Weka [18], a popular open-source machine learning toolkit written in Java.

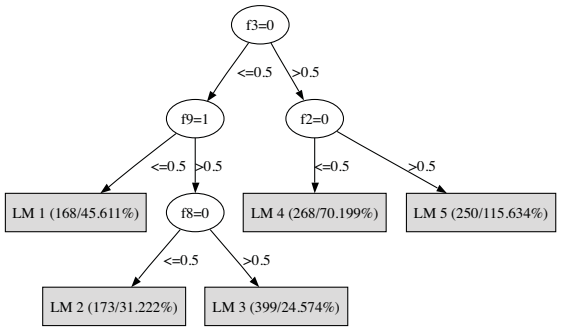

Figure 2: Example model tree for a stereotype.

constructs a tree for classification. However, while leaves of decision trees are class labels, the leaves of a model tree are linear regression models, which are used to predict the target value (in our case, a probability expectation value).

Figure 2 shows an example model tree representing a learned stereotype, with agent features as nodes, feature values as paths, and linear models as leaves.

We define a *learning interval* L which determines the number of experiences an agent must accumulate before building (or re-building) its stereotype model. Once an agent has obtained L experiences, the stereotyping process proceeds as follows. For each opinion $\omega^x_{y:t} \in O_x$, unless $u^x_{y:t} = 1$ (totally uncertain opinions add no knowledge to the model) we add the example $(\vec{F}_y, P(\omega^x_{y:t}))$ to the training set. The tree is then constructed. In this way, an agent may build a model of the relationships between the observable features of trustees and the degrees of trust placed in them. Subsequently, when evaluating an agent y' for which we have no evidence, the model tree can be used to obtain a predicted trust value for $P(\omega^x_{y':t})$. With this value, we can create a new opinion about y', setting the predicted value as the base rate. This satisfies our requirements for the S function.

3.3 Stereotypical Reputation

While stereotypes can help in initial cases, they still require a number of examples of interactions to be built up in order to construct a predictive model. New trustors entering the system are initially at a disadvantage while they gather the experiences from which to form a stereotype, while existing trustors may already possess useful stereotypes. We propose here a way to extend the use of stereotypical evaluations to the reputational case, by allowing new trustors to make use of *stereotypical reputation* gathered from experienced trustors who have already constructed stereotypes.

When evaluating a given agent y, a trustor x will perform a stereotype query when the following conditions hold: (1) x cannot produce a direct evaluation for y, (2) no reputational evidence about y can be found, and (3) x cannot yet form its own stereotype about y.

In this case, x can ask reputation providers if they are able to provide stereotypical reputation of y, in lieu of experiential evidence. Even if no agent in the system has interacted with y before, it is possible that some agents may have already constructed a stereotype from interactions with similar agents to y. On receiving a stereotype query about y from x, a reputation provider r checks whether it already has a stereotypical evaluation for y. If so, the value is returned. Otherwise, r attempts to form a stereotypical evaluation of y and returns the result to x.

Once all stereotypical ratings have been received, x uses the following equation to derive a mean bias for y:

$$SR^x_{y:t} = \frac{\sum_{\rho \in R_x} c_\rho s^\rho_{y:t}}{\sum_{\rho \in R_x} c_\rho} \tag{6}$$

Equation 6 shows how stereotypical reputation $SR^x_{y:t}$ is calculated as the weighted mean of all returned stereotypical ratings $s^\rho_{y:t}$, weighted by the provider's confidence in its stereotype model c_ρ. In our model, this is given by the root mean squared error (RMSE) of the stereotype model of an agent ρ, S_ρ. This provides a measure of the model's accuracy as a function of the differences between the actual opinions and those predicted by the model. Therefore the closer a stereotype fits the observed experiences, the more weight its output is given when calculating the mean:

$$c_\rho = 1 - \sqrt{\frac{\sum_{\omega^\rho_{y:t} \in O_\rho} \left(P(\omega^\rho_{y:t}) - S_x(\vec{F}_y)\right)^2}{|O_\rho|}} \tag{7}$$

In this way, new trustors with few experiences can leverage the stereotypes of more experienced agents in order to make evaluations about new trustees.

Table 1: Trustee Profiles

Profile Id	Mean	StDev	f_1	f_2	f_3	f_4	f_5	f_6
p_1	0.9	0.05	x					x
p_2	0.6	0.15		x		x		
p_3	0.4	0.15			x	x		
p_4	0.3	0.1		x	x		x	
p_5	0.0	1.0		x	x			x

Table 2: Experimental Conditions

Condition	Description
GG	Global interaction, global reputation
AG	Ad-hoc group interaction, global reputation
AA	Ad-hoc group interaction and reputation
GGS	GG + Stereotypes
AGS	AG + Stereotypes
AAS	AA + Stereotypes

4. EVALUATION

4.1 Experimental Setup

In evaluating our approach, we employed a simulated agent society where a set of trustor agents X interact with a set of trustee agents Y over a number of rounds. Each trustee is assigned a performance profile which determines how it will behave. Each profile specifies the mean and standard deviation parameters of a Gaussian distribution from which simulated interaction outcomes will be drawn (from the range [0,1]). In addition, each profile also specifies a number of informative features shared by all agents of that profile. In this

Table 3: Experimental Parameters

Parameter	Value	Description
N_{agents}	500	Trustee agent count
$N_{trustors}$	20	Trustor agent count
N_{groups}	20	Ad-hoc group count
N_{nf}	6	No. of noise features
P_I	0.8	Interaction probability
$Ps_{trustee}$	0.1	Trustee swap probability
G_{lt}	5	Ad-hoc group lifetime
G_{size}	10	Ad-hoc group size
L	100	Learning interval

way, we define the target feature-behaviour relationships we wish our agents to identify. All features are represented as binary variables, each one signifying presence or absence of a given feature.

Since we use continuous values to represent trustee performance, and subjective logic requires frequencies of positive and negative experiences, each trustor x uses a subjective evaluation function $F_x(v,t)$ to map an observed, objective performance value v in a particular task t to a subjective binary evaluation of performance. This function could vary between trustors, so that different trustors "perceive" the same outcome differently. However, for simplicity, we assume that all agents use the same function for all tasks, based around a threshold performance value:

$$F_x(v,t) = \begin{cases} 1 & : v \geq 0.5 \\ -1 & : v < 0.5 \end{cases}$$

The test profiles used in our experiments are given in Table 1. The profile p_1 represents a completely reliable class of agents, while p_4 represents agents who will always perform poorly. Profiles p_2 and p_3 represent unreliable agents who may perform well or poorly, and p_5 represents agents with uniform performance distributions. Agents of type p_5 add noise to the stereotyping process, because their random performance confuses the identification of informative features.

In addition to the informative profile features, agents are also assigned a number of non-informative noise features (N_{nf}), which are not related to profiles. This allows us to evaluate the stereotyping algorithm's ability to deal with features which do not correlate with performance.

In each round, each trustor will decide to interact with probability P_I. If the trustor chooses to interact, it contacts the environment for a list of available agents, and uses its trust model to evaluate the result. The trustor will also obtain a list of available reputation providers from the environment and query them for evidence. The trustee with the highest rating (probability expectation value) according to the trust model will then be selected.

We simulate global population dynamism with the parameter P_L, which determines, for each trustee, the probability that it will leave the society entirely. If a trustee leaves, it is replaced immediately with a new trustee of the same profile as the first. This enables us to maintain the balance of profiles while making the task more difficult for the trustors. As we are interested in ad-hoc teams of agents, we use three parameters, N_{groups}, G_{lt}, and G_{size} which control the number of groups to be created, the group lifetime, and the group size respectively.

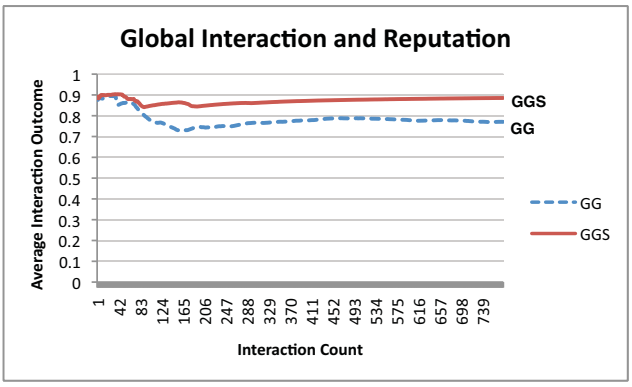

Figure 3: GG vs. GGS

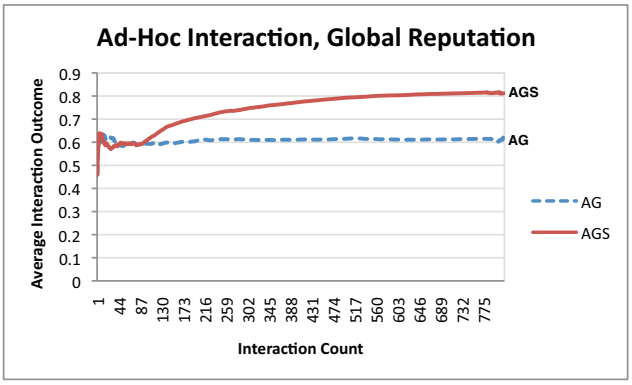

Figure 4: AG vs. AGS

We consider six experimental conditions in total, the parameters of which are summarised in Table 2.

In GG, trustors can interact with any trustee, and all other trustors can be queried for reputational evidence. In AG, only interaction within the ad-hoc group to which the trustor is assigned is allowed, but trustors can perform global reputational queries. In AA both interaction and reputational queries are constrained to the ad-hoc group. The final three cases are as above, but trustors employ stereotyping functionality.

In the ad-hoc group conditions, we create N_{groups} groups to which trustees are randomly assigned such that all groups have at least G_{size} agents. Then, trustors are randomly assigned to groups. Groups interact for G_{lt} rounds, at which point they are disbanded and the assignment process begins again. We have deliberately set the group lifetime to be short, to simulate ad-hoc group/coalition settings. Trustors therefore only have a small window of opportunity to evaluate trustees before the groups are reconfigured.

Table 3 details the parameter sets for the experiments. All parameters remain constant across conditions unless otherwise indicated for the purpose of highlighting their effects.

4.2 Results

Here we present the results of our experiments. Each run consisted of 800 interaction rounds, which was sufficient to observe the stereotype model achieve a stable performance gain in each case. 100 trustees of each profile were created. By creating an even distribution of agent profiles, we aim

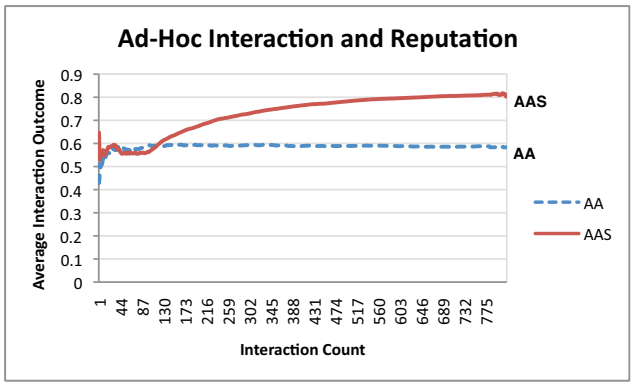

Figure 5: AA vs. AAS

to minimise any effect caused by agents being more or less likely to encounter agents of one profile than another. Also, due to the level of dynamism in the simulation, some agents may find themselves assigned to ad-hoc groups comprising only good or bad partners. As a result, the performance of individual agents may be affected by chance as well as the performance of their respective trust models. In order to minimise this effect and illustrate the model's performance, the graphs presented here plot the *mean interaction outcome* of the trustor population as a whole at each interaction.

Figure 3 shows the performance of both standard and stereotyping models in the GG condition. Both models perform well, with agents in both conditions quickly identifying trustworthy partners. The only experimental parameter which presents a challenge for the non-stereotyping model in this condition is $Ps_{trustee}$, which introduces the chance that known and trusted agents will leave the system and be replaced with new, unknown ones. In these instances, the stereotyping model is able to produce an initial evaluation about newcomers based on past experiences, even when the subjects of those experiences have left the system.

Figure 4 shows the performance of the non-stereotyping model against the stereotyping model in the AG/AGS conditions. The first thing we notice is that both models begin performing more poorly than in the GG condition. Again, once the learning interval has passed, the performance of the stereotyping agents diverges significantly from that of the non-stereotyping condition, eventually surpassing 0.8. The non-stereotyping condition settles much lower at 0.6.

Figure 5 shows the performance of both models in the most challenging case, where any interaction outside of the ad-hoc group is prohibited. Here we see a very similar pattern to that of Figure 4, although the performance of the stereotyping agents improves more slowly. This is caused by the reduced reputational evidence available to stereotyping agents during the learning interval.

We found that in each of the six conditions, the stereotyping model outperformed[3] the standard model after the first learning interval. In each case, the stereotyping model performs similarly to the standard model while training examples are gathered. However, once the the first learning interval has passed, stereotyping agents begin to improve, whereas non-stereotyping agents do not.

[3]The results above were found to be statistically significant by t-test with $p < 0.05$.

4.2.1 *Dynamism*

To evaluate the model's performance in scenarios with a high degree of dynamism (i.e. agents join and leave rapidly and unpredictably), we set the value of $Ps_{trustee}$ to 0.5, which means that in each round, every trustee may be replaced with probability 0.5. Such a high degree of dynamism, while unrealistic, allows us to observe how the stereotyping model performs in an extreme case. As evident from Figure 6, the non-stereotyping model cannot achieve much better than an average performance of 0.5; i.e. no better than chance. The stereotyping model, on the other hand, improves to 0.8 after L interactions.

4.2.2 *Noise features only*

In order to observe the performance of the model when no predictive features are present, we removed all the features from the profiles, so that only noise features were assigned. As these are randomly assigned, there is no hidden correlation between noise features and performance. As we would expect, Figure 7 shows that both models perform identically when no feature correlation exists.

4.2.3 *Unreliable profile features*

Until now, we have enforced a completely positive correlation between profile features and performance. However, in open MAS, features may not be so reliable as predictors of behaviour. In order to evaluate the performance of the model when this assumption is relaxed, we set the probability with which an agent will be assigned a feature from its profile to 0.8, and the probability that a non-profile feature will be assigned to 0.2. As shown in Figure 8, the model is still able to achieve a higher level of performance, albeit not by a large margin. In further trials, we found that as these two parameters are adjusted towards 0.5, the model begins to behave identically to the non-stereotyping model, because profile features are completely uninformative under those conditions.

4.2.4 *Number of Noise Features*

To test the model's resilience to noise features, we performed a series of trials with increasing numbers of noise features, but fixed numbers of informative profile features. As can be seen from Figure 9, the number of noise features generated does not significantly impact on the performance of the algorithm. This is encouraging, as agents could be described by very large feature vectors, with very few diagnostic features, if any.

5. DISCUSSION

The results we have presented show that a stereotyping mechanism based on established machine learning techniques can clearly help agents to make trust evaluations in situations where both direct and reputational evidence is not forthcoming. Under the assumption that correlations exist between trustee features and their performance, our approach can make use of this information. When this assumption does not hold, we have shown that the stereotyping model performs at least as well as the non-stereotyping model.

One possible drawback to our approach is the use of the learning interval L to control the formation of stereotypes in a simple way. The problem lies in selecting an appropriate value; if L is too small, stereotypes may be computed

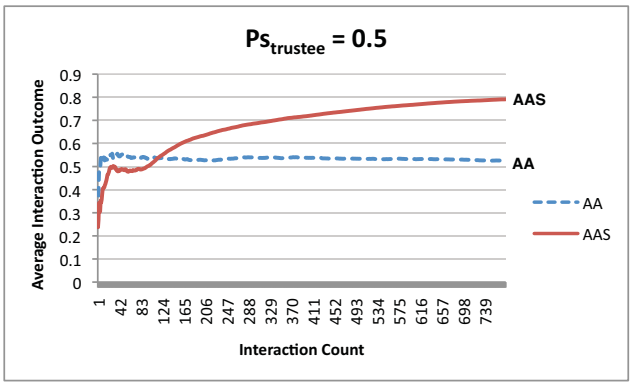

Figure 6: AA vs. AAS, $Ps_{trustee} = 0.5$

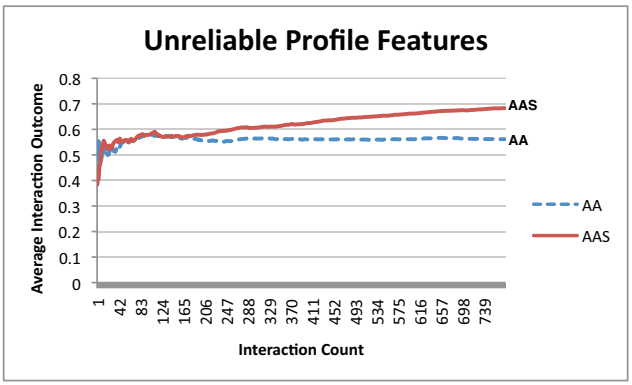

Figure 8: AA vs. AAS, unreliable profile features

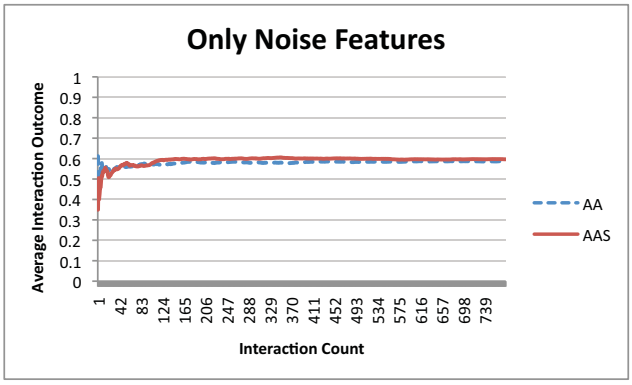

Figure 7: AA vs. AAS, noise features only

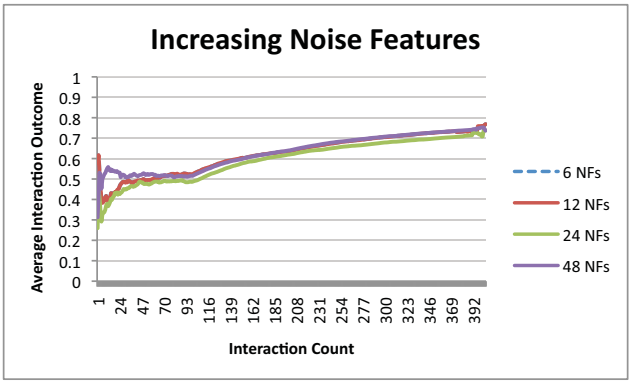

Figure 9: AAS, increasing noise feature count

too early and too frequently. Also, early stereotypes which are based on little data may result in inaccurate expectations about the trustworthiness of a stereotyped partner. Conversely, if L is too large, then the resulting model will be insensitive to changes in the stereotypical behaviours of agents. A better strategy would be to utilise statistics about the accuracy of the computed stereotype model to decide whether stereotypes should be re-computed (or disabled, if no feature-behaviour correlations are believed to exist). As can be seen in Figure 10, lower values of the L parameter result in an earlier but slower increase in performance over time. Higher values cause the model to wait longer before taking effect, but allow the stereotypes to be built from more evidence, resulting in the observed steeper increase in performance over time.

While we have referred to a number of trust evaluation models in this paper, it is worth highlighting here some related approaches which attempt to address the issues of specific interest. The FIRE [6] system employs *role-based trust* to explicitly capture relationships between agents in certain roles. In this approach, rules specify an initial, predetermined degree of trust that will be conferred on partners for whom the rules match. This means that a degree of trust may be present even when no evidence is available. In contrast with our approach, where stereotyping rules are learned from observations, FIRE rules are explicitly specified in a domain specific manner by agent owners. However, the use of such explicit role-based knowledge may be an interesting starting point for combining these two approaches.

Hermoso et al. [5] consider exploiting the organisational structure of Virtual Organisations to calculate trust approximations. They consider trust about triples $\langle B, R, I\rangle$ whose members denote an agent in a particular role engaged in a particular interaction, respectively. They use this representation to define several different types of generalisation which can be applied to obtain an approximate trust value. For example, $\langle \_, R, I\rangle$ denotes the degree of trust any agent in role R performing interaction I, whereas $\langle B, \_, I\rangle$ denotes the degree of trust in agent B, performing interaction I in any role. The aggregation of fully instantiated trust experiences is carried out using the weighted mean of experiences which match the pattern. While this approach does not deal with initial cases such as described in this paper, its query-like notation provides an intuitive way to generalise from individual trust experiences to useful approximations.

It is worth mentioning that the problem of trust evaluation that we address here is distinct from the problem of deciding *to* trust. For the purposes of this paper we have employed a model of decision most commonly found in trust literature which involves selecting the most trusted agent (Equation 1). However, this approach is not always appropriate, as it does not consider the degree of risk in a given interaction. For example, different trustees may 'cost' more to use than others. How should a trustor choose between two agents y_1 and y_2, with ratings $\omega^x_{y_1:t} = 0.8$ and $\omega^x_{y_2:t} = 0.85$, when y_2 costs ten times as much as y_1? Future work will address these issues with a richer model of trust *decision*, as well as evaluation.

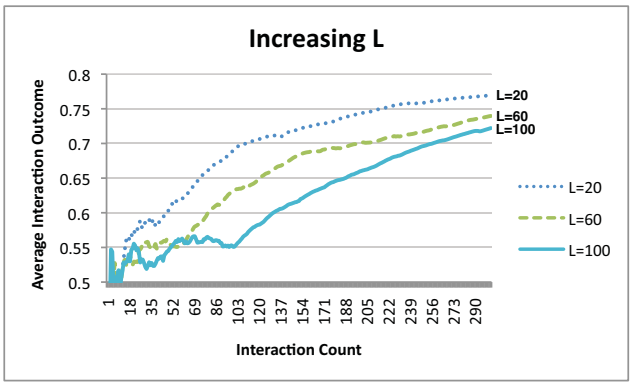

Figure 10: AAS, increasing L parameter

6. CONCLUSIONS

In open MASs, a number of situations can arise where a trust evaluation must be made, but no direct supporting evidence can be found. When a trustee is completely unknown, direct evidence can only obtained (and subsequently propagated as reputation) when a trustor takes a risk and interacts with the newcomer.

We have presented an approach for improving the performance of trust mechanisms in such initial cases when by allowing trust evaluations to be "bootstrapped" by *a priori* assumptions based on stereotypes. We have demonstrated how a stereotyping approach can be used together with a relatively straightforward probabilistic trust model in order to significantly improve performance.

Where hidden feature-behaviour correlations exist in the trustee population, our model has been shown to be robust when both interaction and reputation gathering was constrained to within ad-hoc groups. Our model performs well when the probability of agents leaving, joining or changing identity is high. It has been shown to be resilient to increasing levels of random noise in agent feature vectors. When the assumption that hidden feature-behaviour correlations exist does not hold, the stereotyping model performs no worse than the non-stereotyping model. We therefore conclude that stereotyping can assist in bootstrapping trust evaluations in the problematic initial cases addressed here.

The stereotyping approach presented here can compliment existing trust evaluation techniques. Since our model considers continuous class values, it is directly compatible with any model which reduces its dimensions to a single real measure of trust, regardless of whether the measure is probabilistic or statistical in nature.

Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

7. REFERENCES

[1] L. Breiman. *Classification and regression trees*. Chapman & Hall, 1984.

[2] C. Castelfranchi and R. Falcone. Trust is much more than subjective probability: Mental components and sources of trust. *Proceedings of the 33rd Annual Hawaii Int. Conf. on Systems Sciences*, 2000.

[3] M. Şensoy, J. Zhang, P. Yolum, and R. Cohen. Poyraz: Context-aware service selection under deception. *Computational Intelligence*, 25(4):335–364, 2009.

[4] D. Gambetta. *Trust: Making and Breaking Cooperative Relations*. Blackwell, 1990.

[5] R. Hermoso, H. Billhardt, and S. Ossowski. Integrating trust in virtual organisations. *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, pages 19–31, 2007.

[6] T. D. Huynh, N. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[7] S. Jarvenpaa and D. Leidner. Communication and trust in global virtual teams. *Organization Science*, 10(6):791–815, 1999.

[8] A. Jøsang. Artificial reasoning with subjective logic. In A. Nayak and M. Pagnucco, editors, *Proceedings of the 2nd Australian Workshop on Commonsense Reasoning*, 1997.

[9] A. Jøsang and R. Ismail. The Beta Reputation System. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

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

[11] Z. Kunda and P. Thagard. Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, 103:284–308, 1996.

[12] C. Macrae and G. Bodenhausen. Social cognition: Categorical person perception. *British Journal of Psychology*, 92:239–255, 2001.

[13] D. Meyerson, K. Weick, and R. Kramer. Swift trust and temporary groups. In R. Kramer and T. Tyler, editors, *Trust in Organizations: Frontiers of Theory and Research*. Sage Publications Inc, 1996.

[14] J. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, 1992.

[15] J. Sabater and C. Sierra. Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[16] W. Teacy, J. Patel, N. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[17] Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1551–1556, 2007.

[18] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.