Stereotype Reputation with Limited Observability

Phillip Taylor[1⋆], Nathan Griffiths[1], and Lina Barakat[2] and Simon Miles[2]

[1] Department of Computer Science, The University of Warwick, CV4 7AL, UK
[2] Department of Informatics, Kings College London, WC2R 2LS, UK

Abstract. Assessing trust and reputation is essential in multi-agent systems where agents must decide who to interact with. Assessment typically relies on the direct experience of a trustor with a trustee agent, or on information from witnesses. Where direct or witness information is unavailable, such as when agent turnover is high, stereotypes learned from common traits and behaviour can provide this information. Such traits may be only partially or subjectively observed, with witnesses not observing traits of some trustees or interpreting their observations differently. Existing stereotype-based techniques are unable to account for such partial observability and subjectivity. In this paper we propose a method for extracting information from witness observations that enables stereotypes to be applied in partially and subjectively observable dynamic environments. Specifically, we present a mechanism for learning translations between observations made by trustor and witness agents with subjective interpretations of traits. We show through simulations that such translation is necessary for reliable reputation assessments in dynamic environments with partial and subjective observability.

1 Introduction

In multi-agent systems (MAS) agents must decide whether or not to interact with others, and can use trust and reputation to inform this decision [6, 20, 23]. Trust is the degree of belief, from the perspective of a trustor agent, that a trustee agent will act as they say they will in a given context [1, 2, 10]. A trustor with a high level of trust in a trustee is confident of a successful interaction with a good outcome. Likewise, a low level of trust in a trustee implies that the trustor agent expects a bad outcome. Whereas trust is assessed using experiences of the trustor, reputation is based on the opinions of several agents in a network.

In domains where agents join and leave with high frequencies, it can be difficult to reliably assess trust and reputation due to limited relevant experience. A trustor agent who recently joined a MAS, for instance, will have limited experience with trustees and be unable to reliably assess trust. In this case, opinions of witness agents can be used to produce a reputation assessment [9]. When a trustee agent is new to a MAS, however, no agent will have direct experience with them, preventing reliable assessments of trust and reputation.

⋆ Phillip.Taylor@warwick.ac.uk

In many domains, trustee agents exhibit traits that provide insight into their behaviour during, but can be observed prior to entering into, an interaction [2, 12, 16]. Such traits are referred to as stereotypes, and can be used to bootstrap trust and reputation assessments when experience is limited. If a trustor has observed a stereotype it can be used to assess stereotype-trust in a trustee, otherwise stereotype-reputation can be assessed using witnesses. A witness may be unable themselves to observe the trustee traits, however, and must assess those observed and reported by the trustor. When these trait observations are subjective and agents have different interpretations or observe different traits, communication of observations and assessing stereotype-reputation is problematic. In this paper we propose the Partially Observable and Subjective Stereotype Trust and Reputation (POSSTR) system, which enables agents in partially observable environments to translate observations from different subjective perspectives, and enables witnesses to provide reliable stereotype-reputation assessments. POSSTR does not replace existing reputation systems, but rather it should be used alongside them to provide a way to deal with partial observability and subjectivity. We make the following contributions:

- We propose a mechanism for learning a translation between traits observed by a trustor and a witness, and
- Using simulations, we show that our translation mechanism improves trust and reputation assessments in environments with partially and subjectively observable traits.

The remainder of this paper is structured as follows. Related work is discussed in Section 2. Section 3 describes our use case and outlines the problem with partial observability and subjective observations of traits. The POSSTR system, which overcomes challenges in such partially observable and subjective domains, is proposed in Section 4. The simulation environment used for evaluating POSSTR is outlined, and results from our investigation are discussed in Section 5. Finally, Section 6 concludes the paper.

2 Related work

In many domains, trustor agents use trust and reputation to select interaction partners from sets of trustees [6]. Trust can be assessed using direct experience gathered by a trustor interacting with trustee agents. Where direct experience is lacking, reputation assessments are gathered from witness agents [1, 6, 23]. In highly dynamic environments, where agents leave or depart regularly, relevant experience with trustees is often insufficient to produce reliable assessments. In these cases, stereotypes can be used to bootstrap trust and reputation [2, 12, 16].

Trustees often exhibit traits that are observable to trustors prior to an interaction. When these traits are related to the behaviour of trustees during interactions, the trustor can form stereotypes that can be used as a surrogate for other more relevant experience in assessing trust and reputation [2, 12, 16]. If several trustee agents exhibit the same trait and are similarly reputable, for

instance, a new trustee also exhibiting the trait may be assumed to have similar reputation. To build a stereotype-trust model, a trustor must interact with several trustees and analyse the correlations between their observable traits and reputations. If the trustor is unable to assess stereotype-trust because they lack relevant experience of the observed traits, stereotype-reputation assessments can be requested from witnesses [2, 12].

To assess reputation of a trustee, the trustor combines the following:

- *Direct-trust* based on direct experience the trustor has with the trustee;
- *Witness-reputation* based on witness reports summarising their experiences with the trustee;
- *Stereotype-trust* based on common trustee traits observed by the trustor; and
- *Stereotype-reputation* based on experience and common trustee traits observed by witnesses.

Direct-trust requires the trustor to have previously interacted with the trustee being evaluated. The same is true when witnesses compute opinions about a specific trustee, to be sent to the trustor. In combination, direct-trust and witness-reputation, make up the Beta Reputation System (BRS), as proposed by Jøsang et al. [9]. Other reputation systems that combine direct-trust and witness-reputation include FIRE [7], TRAVOS [18], BLADE [15], and HABIT [17]. TRAVOS extends BRS to cope with dishonest witnesses by discounting information provided by unreliable sources, and BLADE and HABIT both use Bayesian networks to transform opinions from witnesses that are unreliable in a consistent way.

As well as direct-trust and witness-reputation, FIRE [7] also combines two other sources of information, namely certified and role-based trust. Certified trust is based on testimonials gathered by the trustee and given to the trustor, and as a result is often optimistic of their performance in an interaction. Role-based trust can be viewed as a kind of stereotype, but the roles are defined statically by trustors and as a result it is limited compared to the observation based approach used in this paper. Stereotype-trust enables assessments of trustees with whom the trustor has not previously interacted, by assuming trustees with similar observable traits behave similarly. As with witness-reputation, stereotype-reputation is gathered from witnesses who provide their opinions.

Liu et al. [12] proposed that characteristics of trustee agents, correlated with their trustworthiness, be used to separate them into groups defined by their common characteristics. When evaluating a new trustee, its observable characteristics are compared to those that define each group and the mean trustworthiness of their members is used as the stereotype and overall trust score. When a trustor is unable to determine stereotype-trust, because they lack experience with the particular characteristics, stereotype-reputation is gathered from witnesses. Similarly, Teacy et al. [17] suggest building a separate HABIT model for overlapping groups of agents defined by stereotypes, but they do not describe how such groups should be formed.

The bootstrapping model proposed by Burnett et al. [2] combines all four sources of trust and reputation. Instead of the clustering approach employed by Liu et al. [12], the trustor learns a regression model that maps observed traits to trustworthiness. Observed characteristics of trustees are then input into the model with the output used as a base reputation value in a probabilistic trust model. In this way, the base trust value has less of an impact on the overall reputation score as more direct evidence is gathered about the trustee. STAGE, proposed by Şensoy et al. [16], combines direct-trust, sterotype-trust, and witness-reputation in a similar way to Burnett et al. [2]. In STAGE, reports provided by witnesses for both witness- and stereotype-reputation are discounted based on their perceived reliability. As well as using stereotype-trust to bootstrap assessments of trustees, STAGE also learns stereotypes for witnesses to bootstrap this reliability assessment of opinions. To avoid the need for opinions, Fang et al. [3], build a stereotype-trust model that enables observations to generalise to others when experience for a particular stereotype is limited.

In these existing reputation models, witness-reputation requires that the trustee is known to the witness. This means that the trustor must be willing to identify the trustee to the witness, and the witness must have interacted with them previously. Likewise, stereotype-reputation as proposed by Burnett et al. [2] requires:

- The trustee is identified and the witness can observe its traits (i.e. trustees are *fully observable*), and
- All agents observe trustee traits in the same way (i.e. trustee traits are *objective*).

In real-world environments, however, trustees may be *partially observable* and such observations may often be *subjective*. If the trustees are only *partially observable* and the witness is unable to observe the traits, the trustor must disclose their observations of traits for the witness to provide their opinion. For example, if a new trustee is unknown to a witness, the trustor must describe their observations when requesting a stereotype-reputation assessment. If trait observations are also *subjective*, those observed by a trustor may be meaningless to a witness. In this paper, we propose the POSSTR system to overcome this issue by translating traits observed by the trustor.

3 Problem setting

To formalise these issues of partial observability and subjectivity, we define the full set of traits in an environment that agents can exhibit or observe as $\mathbf{\Theta}$. For example, taxi services can exhibit numerous traits, including 'airport transfer' and 'suitcase storage'. Each individual trustee agent, te, exhibits a subset of these traits, $\theta^{te} \subseteq \mathbf{\Theta}$, and each trustor agent, tr has an observation function, $\mathcal{O}_{tr} : \mathcal{P}(\mathbf{\Theta}) \to \mathcal{P}(\mathbf{\Theta})$. When presented with a trustee, this observation function determines how the traits of a trustee are interpreted, $\theta^{te}_{tr} = \mathcal{O}_{tr}(\theta^{te})$.

In a fully observable setting, it is valid for all agents to observe the traits of all trustees themselves. When assessing stereotype-reputation with full observability, witness agents can apply their observation function, $\theta_w^{te} = \mathcal{O}_w(te)$, and correctly interpret any associated stereotype. With partial observability the traits of some trustees may be unavailable, such as when there is a cost to making observations or if the trustees are in different locations. In such partially observable environments, traits may only be accessible when considering whether to interact with a trustee, i.e. when assessing direct-trust or stereotype-trust. An agent that has neither visited a city nor considered using a taxi there, for example, cannot use their observation function when acting as a witness for stereotype-reputation. In such cases the traits observed by the trustor must be assessed by witnesses instead.

If traits are observed objectively by agents, then observations made by a trustor are the same as those that a witness would make, i.e. $\mathcal{O}_{tr}(te) = \mathcal{O}_w(te)$. With objective observations, therefore, there is no issue with partial observability and a witness can directly assess observations made by the trustor. With subjectivity, however, agents may have no interest in a particular trait or interpret traits differently. A customer considering a taxi service for airport transfer who is carrying hand luggage only, for example, may not notice if the taxi service is able to accommodate suitcases or not. An observation of suitcase storage may then be meaningless to this customer, resulting in a poor a stereotype-reputation assessment. In another situation, two customers may have different interpretations of suitable storage for suitcases. Such subjective observations can lead to misunderstandings of stereotype-reputation assessments, and so a translation between the two subjective observations is required.

To overcome these potential misunderstandings, we propose that the trustor or witness learns to translate observations made by the trustor agent to what the witness would have observed. After the translation is made, the witness can assess the stereotype in a meaningful way and respond with their opinion. To learn such a translation function, either the trustor or witness must provide their observations of several trustees to the other. These observations do not have to be linked to a reputation assessment for the trustee, but can have been observed during other reputation assessments. Traits of trustee agents in both sets of observations can then be analysed for correlations and a translation learned. If the trustor observes 'suitcase storage' for several taxi services for which the witness has observed 'airport transfer', for example, a translation between the two observations can be learned. If the trustor observes 'suitcase storage' for an entirely new trustee, this can be translated into the witness's stereotype for 'airport transfer' when assessing stereotype-reputation.

4 The POSSTR model

In assessing trust and reputation it is typical to aggregate ratings of previous interactions. An interaction between tr and te is recorded in the tuple $\langle tr, te, \theta_{tr}^{te}, r_{tr}^{te} \rangle$, where θ_{tr}^{te} are the traits of te that were observed by tr prior

to the interaction, and r_{tr}^{te} is the rating given by tr. Without loss of generality, we assume that ratings are binary, with 1 indicating success and 0 indicating otherwise. A real-valued rating can be converted to binary by choosing a threshold, above which the interaction is deemed successful and otherwise it is unsuccessful. The aim of the reputation assessment is then to determine the likelihood of a future interaction with a trustee being successful.

4.1 Direct-trust

In evaluating the direct-trust of a trustee, te, a trustor, tr, aggregates their relevant interaction records, $\mathbf{I}_{tr}^{te}$, with te. There are many possible aggregations, but as in existing work on stereotypes [2, 16], and BRS [9], we use one based on Subjective Logic (SL) [8]. SL is a belief calculus that can represent opinions as degrees of belief, b, disbelief, d, and uncertainty, u, in BDU triples, (b, d, u), where $b, d, u \in [0, 1]$, and $b + d + u = 1$. In SL, a completely uncertain opinion is represented as (0,0,1), and total belief is represented as (1,0,0). As evidence is accrued and the opinion changes, the degrees of belief, disbelief, and uncertainty change also.

In BRS [2, 9], the trustor computes a BDU triple by counting the number of successful interactions they have had with the trustee, $p_{tr}^{te} = |\mathbf{I}_{tr}^{te} : r_{tr}^{te} = 1|$, and the number of unsuccessful interactions, $n_{tr}^{te} = |\mathbf{I}_{tr}^{te} : r_{tr}^{te} = 0|$. A mapping from interaction records and ratings to the belief, disbelief, and uncertainty is provided by,

$$b_{tr}^{te} = \frac{p_{tr}^{te}}{p_{tr}^{te} + n_{tr}^{te} + 2}, \qquad d_{tr}^{te} = \frac{n_{tr}^{te}}{p_{tr}^{te} + n_{tr}^{te} + 2}, \qquad u_{tr}^{te} = \frac{2}{p_{tr}^{te} + n_{tr}^{te} + 2}. \tag{1}$$

If there are two ratings of 1 and one rating of 0, for example, the resulting BDU triple is $(0.4, 0.2, 0.4)$. This mapping ensures that uncertainty decreases monotonically as the evidence is accumulated. Other mappings from ratings to SL are possible, such as that proposed by Wang and Singh [21, 22] where uncertainty is affected by disagreement in ratings as well as the amount of evidence.

The likelihood that a future interaction with te will be successful, is then calculated as,

$$P(\hat{r}_{tr}^{te} = 1) = b_{tr}^{te} + a_{tr}^{te} \times u_{tr}^{te}, \tag{2}$$

where $\hat{r}_{tr}^{te}$ is the future rating being predicted and a_{tr}^{te} is the Bayesian prior. The prior in BRS [9] is $a_{tr}^{te} = 0.5$, which represents that an interaction with an unknown agent for which there is no information is equally likely to be successful or unsuccessful. A prior of greater than 0.5 means that uncertain opinions lean more to belief in success, whereas priors less than 0.5 make $P(\hat{r}_{tr}^{te} = 1)$ closer to 0. As evidence is gathered, the uncertainty reduces toward 0 and the prior has less of an effect on the likelihood of success. Stereotypes, as discussed in Sections 4.3 and 4.4, can be used to inform this prior based on observations of trustee traits.

4.2 Witness-reputation

When the trustor has insufficient ratings of a trustee, witnesses, $w \in W$, are asked to provide theirs. The witness ratings are then combined with those of the trustor using SL as described above,

$$p^{te} = p^{te}_{tr} + \sum_{w \in W} p^{te}_{w}, \quad n^{te} = n^{te}_{tr} + \sum_{w \in W} n^{te}_{w}, \tag{3}$$

where p^{te}_{w} and n^{te}_{w} are respectively the number of positive and negative interactions reported by witness, w, about te. Witness-reputation is then computed as,

$$P(\hat{r}^{te}_{tr} = 1) = b^{te} + a^{te}_{tr} \times u^{te}, \tag{4}$$

where the Bayesian prior is again $a^{te}_{tr} = 0.5$, and

$$b^{te} = \frac{p^{te}}{p^{te} + n^{te} + 2}, \quad u^{te} = \frac{2}{p^{te} + n^{te} + 2}. \tag{5}$$

4.3 Stereotype-trust

Stereotypes can be used to inform the Bayesian prior in environments where trustees that exhibit similar observable traits have performed similarly in interactions. For instance, the ratings given to interactions with known agents can be used as the prior for an unknown agent with similar traits. A stereotype model,

$$f_{tr} : \mathcal{P}(\boldsymbol{\Theta}) \rightarrow \mathbb{R}, \tag{6}$$

is learned by tr, which maps traits of a trustee agent observed by tr to a stereotype-trust value,

$$a^{te}_{tr} = f_{tr}(\theta^{te}_{tr}), \tag{7}$$

that is used as the Bayesian prior in Equations 2 and 4 when computing direct-trust or witness-reputation.

The stereotype model is learned by generating a training sample for each agent the trustor has previously interacted with. In each of these samples, the te traits observed by tr are the input features, θ^{te}_{tr}. The target, or class value, is the direct-trust that tr has in te, as outlined in Section 4.1, with a Bayesian prior of 0.5. The training data is therefore a set of samples that express observed trustee traits and their direct-trust values. A regression model is then learned to map traits observed by tr to the trust in agents that express those traits, which can be used as the Bayesian prior in Equation 2. As before, if the trustor has high uncertainty about a trustee and the stereotype model outputs a prior close to 0, the direct-trust will be low. As the trustor gains experience with trustee, the prior will have less effect on the trust value.

As in Burnett et al. [2] and Şensoy et al. [16], we learn the mapping from features of a trustee to the likelihood of a successful interaction using the M5 model tree algorithm [13]. The M5 model tree recursively splits training samples

using the values of the features that best discriminate the class labels. Whereas in typical decision trees the leaves are target values, the leaves of the M5 tree are piecewise linear regression models that output the target value. The regression models are learned using samples that were not divided in learning the tree and therefore use features not specified by the ancestors of the leaf. If all features are specified, the linear regression model defaults to outputting the mean target value of the samples in the relevant split. The splitting process stops at the level where the leaf model would have the highest accuracy on the training data. If there are many traits observed by a trustor then it may be necessary to perform feature selection to reduce their number [4].

4.4 Stereotype-reputation

When the trustor is not confident in their stereotype-trust assessment, witnesses can be asked for their stereotype based assessment of the trustee. As with the trustor, each witness, $w \in W$, has their own stereotype model,

$$f_w : \mathcal{P}(\boldsymbol{\Theta}) \rightarrow \mathbb{R}, \tag{8}$$

learned using their own experience of trustee agents. The witness in some cases may have observed the trustee previously, in which case they are able both to provide a witness-reputation assessment as well as use the features they observed, θ_w^{te}, in their stereotype model. In other cases the witness may have not observed the trustee previously and must rely on the stereotype features observed by the trustor, who may have observed different features in different ways. This necessitates a translation function between the two observation capabilities,

$$f_{tr \rightarrow w} : \mathcal{P}(\boldsymbol{\Theta}) \rightarrow \mathcal{P}(\boldsymbol{\Theta}). \tag{9}$$

This function converts observed features of a trustee from the subjective perspective of the trustor, tr, to that of the witness, w. It is a multi-target learning problem with an input of stereotype features the trustor, tr, has observed, θ_{tr}^{te}, and an output vector of features that the witness, w, would observe, $\hat{\theta}_w^{te}$.

To learn the translation function, training data is generated from common observations that both the witness and the trustor have made. When requesting a stereotype assessment from a witness, either the trustor provides their observations of other trustee agents to the witness or vice versa. These observations, consist of the observed traits along with the trustee identifier. As an example, consider that the trustor has observed the traits of three trustees, $\{\theta_{tr}^{te_1}, \theta_{tr}^{te_2}, \theta_{tr}^{te_3}\}$, and a witness has observed those of two, $\{\theta_w^{te_1}, \theta_w^{te_2}\}$. Training data can then be generated by matching up the common observations, as $\{\theta_{tr}^{te_1} : \theta_w^{te_1}, \theta_{tr}^{te_2} : \theta_w^{te_2}\}$, where ':' separates the inputs and outputs. These observations may have been made without having interacted with the trustees, such as a potential customer observing traits of taxis during an assessment but without using their service. These common observations samples form the training data that can be input into a multi-target learning algorithm [14].

Multi-target learning algorithms learn mappings from input features to multiple targets. One simple yet powerful approach is the binary relevance method [19], where a separate model is built for each target. In this paper, a model is learned that maps traits observed by the trustor to each trait that would be observed by the witness. The traits observed by the trustor are then input into each of the learned models and their outputs are combined to be the traits the witness would have observed. As the base learning algorithm for each of the output traits we use Naïve Bayes, although any classification algorithm may be used in its place [4, 14].

If a witnesss has not observed the trustee, the trustor's observations are input into the learned translation,

$$\hat{\theta}_w^{te} = f_{tr \to w}(\theta_{tr}^{te}), \tag{10}$$

to estimate the traits that they would have observed. This output is then used in the witness stereotype model,

$$a_w^{te} = f_w(\theta_w^{te} | \hat{\theta}_w^{te}) = \begin{cases} f_w(\theta_w^{te}) & \text{if witness observed trustee,} \\ f_w(\hat{\theta}_w^{te}) & \text{if trustor provided observations,} \end{cases} \tag{11}$$

which outputs the prior from the witness perspective to be returned to the trustor. A new Bayesian prior is then computed as the mean stereotype assessment of the trustor and witnesses,

$$a^{te} = \frac{1}{|W| + 1} \left(a_{tr}^{te} + \sum_{w \in W} a_w^{te} \right). \tag{12}$$

Finally, the overall reputation score is computed as,

$$P(\hat{r}_{tr}^{te} = 1) = b^{te} + a^{te} \times u^{te}. \tag{13}$$

4.5 Subjective opinions

In many domains, witnesses cannot be assumed to rate interactions objectively or report ratings benevolently. This is the same for witness-reputation as it is for stereotype-reputation, where witnesses may be dishonest or otherwise have different opinions about a trustee or its traits. While this issue is out of the scope of this paper, there are two broad approaches to dealing with this problem. First, information provided by unreliable witnesses can be discounted, or weighted lower than more reliable information [16, 17]. In this method, opinions of a witness are compared to those of the trustor for the same trustees or traits. If there is a significant difference in opinions then the witness is deemed unreliable and their reports are discounted before being combined with others. Zhang et al. [24] evaluate the reliability of witnesses by comparing their reports to trustor ratings as well as those of other witnesses. Second, if witnesses are unreliable in a consistent way, their opinions can be reinterpreted to be from the

Profile	Description	Mean	STD	$\theta^{\mathbf{te}}$	$\mathcal{O}_{\mathbf{tr}} = \mathbf{001122}$
1	Usually good	0.9	0.05	100001	100010
2	Often good	0.6	0.15	010100	010011
3	Often poor	0.4	0.15	001100	000011
4	Usually poor	0.3	0.05	011010	010001
5	Random	0.5	1.00	011001	010010

Table 1: Objective trustee profiles. The observations of an example observation vector, $\mathcal{O}_{tr}$, are also shown.

Strategy	Description	Definition
Random	No information	NA
T	Direct-trust	Eq 2, where $a_{tr}^{te} = 0.5$
TR	Direct-trust + witness-reputation	Eq 4, where $a_{tr}^{te} = 0.5$
T+ST	Direct-trust + stereotype-trust	Eq 2, where $a_{tr}^{te} = f_{tr}(\theta_{tr}^{te})$
TR+ST	Direct-trust + witness-reputation + stereotype-trust	Eq 4, where $a_{tr}^{te} = f_{tr}(\theta_{tr}^{te})$
TR+STR	Direct-trust + stereotype-trust + witness-reputation + stereotype-reputation	Eq 13, where $a_{w}^{te} = f_{w}(\theta_{w}^{te} \mid \theta_{tr}^{te}),\ \forall w \in W,$ and $a_{tr}^{te} = f_{tr}(\theta_{tr}^{te})$
POSSTR	Direct-trust + stereotype-trust + witness-reputation + stereotype-reputation (with translation)	Eq 13, where $a_{w}^{te} = f_{w}(\theta_{w}^{te} \mid \hat{\theta}_{tr}^{te}),\ \forall w \in W,$ and $a_{tr}^{te} = f_{tr}(\theta_{tr}^{te})$

Table 2: Reputation assessment strategies investigated listing their information sources and definitions.

perspective of the trustor [11, 15, 17]. It is worth noting that these translations are different to the observation translations proposed in Section 4.4, as they aim to translate a single variable (ratings) with potentially different ranges, whereas our translation is more general and aims to translate multiple observed traits. As with discounting, opinions of the witnesses and trustor are compared to learn a mapping from one to the other, but investigating either approach to subjective ratings alongside partially observable trustees and subjective stereotypes is out of the scope of this paper.

5 Evaluation and results

To evaluate POSSTR we use a simulated marketplace based on that used by Burnett et al. [2] and Şensoy et al. [16]. The simulation consists of trustor and trustee agents that interact over 250 rounds. Each trustee agent is randomly

assigned one of five profiles at the beginning of the simulation, defining a mean, standard deviation (STD), and observable traits, θ^{te}, as outlined in Table 1. The mean and STD define the Gaussian distribution from which interaction outcomes are drawn. As in Burnett et al. [2] and Şensoy et al. [16], an interaction with an outcome greater than a success threshold of 0.5 is deemed successful and given a rating of 1 by the trustor, otherwise it is rated as 0. The observable traits distinguish each of the profiles, to be used in stereotype assessments of trustees. Each element in these feature vectors can be interpreted as the trustee exhibiting a trait or not, e.g. the first trait may represent 'airport transfer'.

Each trustor and trustee agent leaves the simulation with a probability of 0.05 in each round, to be replaced by another. New trustees are assigned a profile selected uniformly at random from those in Table 1. The number of agents in the simulation is static, therefore, and in all of our simulations there were 100 trustee agents and 20 trustor agents. In each round, each trustor agent is given a random 10 available trustees from which they select the one with highest reputation as an interaction partner. Similarly, in each reputation assessment, each trustor requests witness-reputation and stereotype-reputation from 10 random witnesses.

Trustee traits are observed subjectively through trustor observation functions, $\mathcal{O}_{tr}(\theta^{te})$, defined by an observation vector, $\mathcal{O}_{tr}$, assigned to each new trustor. The observation vector is the same length as the number of traits in the network and each value corresponds to an observable trait. A value of 0 means that the trait is observed with the correct value if it is expressed by a trustee, and 1 means that the trait is never observed (or always observed as 0). A value of 2 in the vector means that the trustor always changes the value of the trait, i.e. a trustee trait of value 0 is observed as a 1 and vice versa. An example observation vector, along with the traits observed by such a trustor, is shown in the final column of Table 1. Observation vectors are sampled from a distribution defined by subjectivity parameters, s and o, which determine the likelihoods of 0, 1, or 2. A value is 1 with a probability of o, and given that its value is not 1 it has a value of 2 with probability of s. A higher o means that more traits are ignored, and a higher s increases the likelihood that a trait is interpreted incorrectly.

In our experiments we compare each of the strategies outlined in Table 2. For example, the T+ST strategy combines direct-trust and stereotype-trust information in the assessment of trustee agents, as defined by Equation 2. Similarly, TR+STR uses all four information sources in each reputation assessment, regardless of any confidence that may be derived from the number of experiences. At the end of each round, the mean overall utility gained by all agents is computed and recorded as the simulation utility. All results presented in this paper are averaged over 50 iterations of our simulation. In all settings and for all strategies the standard deviation of simulation utilities was less than 5% of the mean, and the standard error was less than 1% of the mean. Also, all significance results discussed are from an ANOVA followed by an all-pairs t-test, with multiple comparisons normalised using the Bonferroni correction.

Strategy	Mean utility	STD utility
Random	*126.73	5.36
T	*142.31	5.17
TR	*194.80	4.78
T+ST	*186.56	4.78
TR+ST	*217.07	3.98
TR+STR	226.48	3.36
POSSTR	227.60	2.14

Table 3: Fully observable trustees with objective traits. Utilities significantly smaller ($p < 0.01$) than that of POSSTR are prepended with a '*'.

5.1 Full observability and objective traits

Table 3 shows the mean utilities after 250 rounds over the 50 iterations, with fully observable trustees and objectively observable traits ($s = o = 0$). The differences between each pair of strategies, excluding TR+STR and POSSTR, was significant with $p < 0.01$. The strategy that gained the lowest utility in all cases was Random, followed by using direct-trust only (T). Using witness opinions alongside direct-trust (TR), trustor agents were able to choose better interaction partners. This extra information gathered from witnesses is clearly advantageous, given that trustor exploration of the trustee population was limited and agent turnover was high.

Trustor agents were better able to search the trustees and gain good utilities when they combined witness-reputation with either stereotype-trust or stereotype-reputation. Combining direct-trust with stereotype-trust (T+ST) was also beneficial when compared to using only direct-trust (T) although the utility gained was significantly lower than using witness-reputation (TR). The highest utilities were gained when all four kinds of trust and reputation were used (TR+STR and POSSTR). With full observability and no subjectivity the translation was not required and there was no advantage to using POSSTR over using the observed traits directly, as in TR+STR.

5.2 Partially observable trustees

To model partial observability we first restricted observations of trustee traits to those made during previous assessments. If a witness had not previously assessed direct-trust of a trustee, they were unable to observe the traits themselves and used those observed by the trustor. This restriction on observations was also applied when generating training data for the translation function in POSSTR. Tables 4(a) and (b) show the utilities gained for strategies in this assessment-restricted partially observable setting, for different levels of subjectivity. The results in Table 4(a) are for different values of o with $s = 0$ (agents observed traits with different likelihoods) and the results in Table 4(b) are for different values of s with $o = 0$ (agents flipped values of traits with different likelihoods). As with full observability, a significant difference was observed between each

	o = 0	o = 0.25	o = 0.5	o = 0.75
Random	*126.73 (5.36)	*125.69 (4.14)	*123.67 (4.78)	*125.40 (5.31)
T	*142.31 (5.17)	*140.49 (4.76)	*141.13 (5.71)	*141.58 (4.59)
TR	*194.80 (4.78)	*195.42 (5.16)	*195.57 (5.38)	*194.22 (4.97)
T+ST	*186.56 (4.78)	*181.89 (4.56)	*177.39 (5.79)	*156.78 (6.33)
TR+ST	*217.07 (3.98)	*214.91 (3.81)	*213.26 (3.66)	*205.89 (5.07)
TR+STR	226.51 (3.27)	223.65 (2.88)	221.49 (3.57)	214.75 (4.58)
POSSTR	226.68 (3.74)	223.65 (3.98)	223.76 (3.93)	215.54 (4.52)

(a) Observed traits are not changed (s = 0)

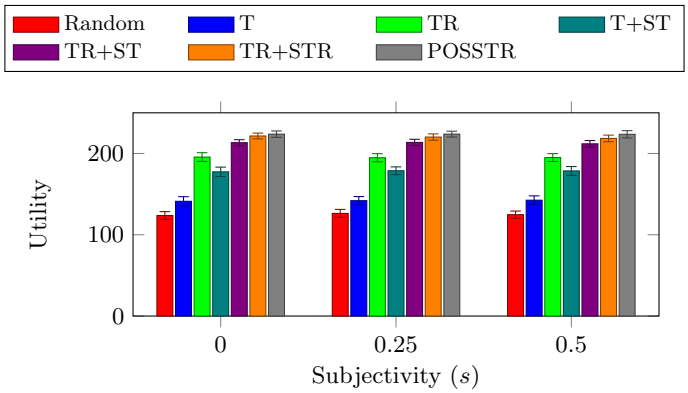
	s = 0	s = 0.25	s = 0.5	s = 0.75	s = 1
Random	*126.73 (5.36)	*125.04 (5.79)	*126.29 (5.53)	*125.40 (5.52)	*126.28 (4.55)
T	*142.31 (5.17)	*142.00 (5.74)	*141.77 (4.34)	*141.33 (5.00)	*141.05 (4.61)
TR	*194.80 (4.78)	*195.29 (6.15)	*194.71 (5.16)	*195.34 (6.06)	*195.38 (4.85)
T+ST	*186.56 (4.78)	*185.73 (4.78)	*186.25 (5.12)	*186.78 (5.03)	*186.76 (5.47)
TR+ST	*217.07 (3.98)	*217.08 (3.58)	*216.57 (3.53)	*216.46 (3.28)	*216.49 (3.65)
TR+STR	226.51 (3.27)	*223.01 (3.68)	*220.11 (3.42)	*222.46 (4.00)	226.47 (3.57)
POSSTR	226.68 (3.74)	226.94 (3.19)	226.22 (3.04)	226.44 (2.83)	226.08 (3.14)

(b) All traits are observed (o = 0)

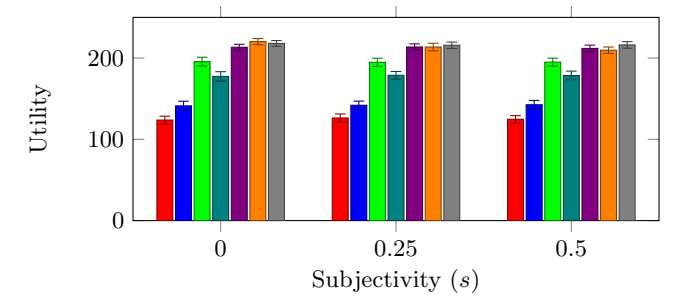
Table 4: Utilities for strategies with different levels of subjectivity. STD shown in braces after each result and results significantly smaller ($p < 0.01$) from POSSTR are prepended with '*'.

pair of strategies, other than TR+STR and POSSTR, within each subjectivity and observability condition. For all levels of subjectivity, the utilities gained by strategies that do not use stereotypes, namely Random, T, and TR, were the same as with fully observability. Similarly, with objective traits, i.e. $o = s = 0$, the utilities for TR+STR and POSSTR, which both use stereotype-reputation, were not significantly different ($p > 0.05$) from with full observability. This is because observations made by a trustor were the same as those that a witness would have made in this setting.

For all strategies that use stereotypes, higher values of o led to lower utilities, with performance being substantially lower when $o = 0.75$. This is likely due to there being fewer traits observed by the trustor agents, meaning there is less distinction between the trustee profiles. The value of s had no significant effect ($p > 0.05$) on the performance of strategies that did not use stereotype-reputation, including T+ST and TR+ST. When trustors had to communicate their observed traits to witnesses subjectively, i.e. when $0 < s < 1$, the TR+STR, which does employ stereotype-reputation, was negatively affected. POSSTR did not suffer any significant loss ($p > 0.05$) in utility gain over all values of s, as a result of it successfully translating observed traits before computing stereotype-reputation. When $0 < s < 1$, therefore, POSSTR significantly outperformed TR+STR ($p < 0.01$), again after pairwise t-tests with the Bonferroni correction. This means that POSSTR reliably assessed reputation with partially observable trustees and subjectively interpreted traits, while TR+STR did not.



(a) Assessment restricted observations



(b) Interaction restricted observations

Fig. 1: Utilities for different levels of subjectivity with $o = 0.5$, with partially observability and observations are restricted to previous (a) assessments and (b) interactions. The error bars show the STD.

Figure 1(a) shows the utilities gained by strategies for $o = 0.5$ and $s = 0$, 0.25, and 0.5. The results are similar to those found in Table 4(b), where $o = 0$, and show that POSSTR had significantly the highest performance in all cases. The results for $s = 0.75$ and $s = 1$ are omitted from this plot for clarity, as the utilities under these conditions were mirrored by those when $s = 0.25$ and $s = 0$ respectively.

To restrict observability further we limited observations to interactions, meaning that a witness must have interacted with a trustee for their traits to be available when assessing stereotype-reputation. These results, for different values of s and $o = 0.5$, are presented in Figure 1(b), where again the strategies that do not use witness-stereotypes were unaffected by the observability of trustee traits. The TR+STR and POSSTR strategies both gained lower utilities when traits were subjectively observed in this setting than when observability was restricted to assessments. With $o = 0.5$ and $s = 0.5$, POSSTR again significantly

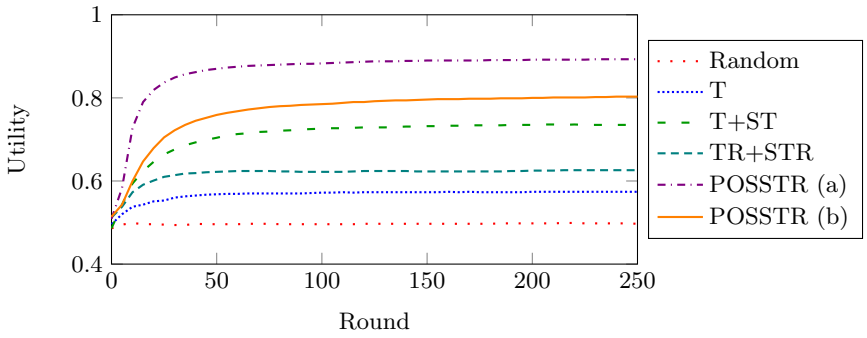


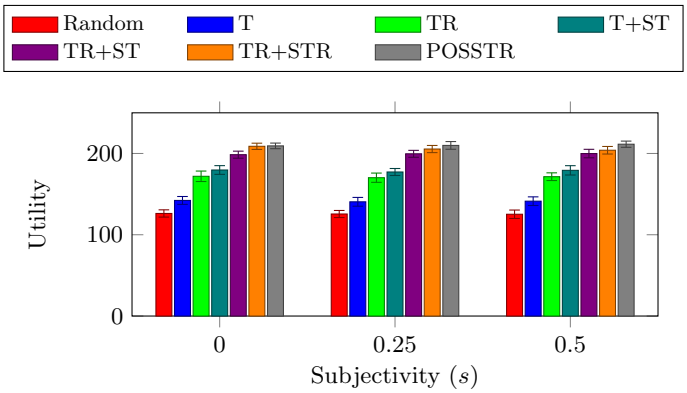
Fig. 2: Utilities for strategies with private trustee identifiers, where $o = s = 0.5$.

($p < 0.01$) outperformed all other strategies but was outperformed by TR+STR ($p < 0.05$) when $s = 0$ and traits were objective. No significant difference between TR+ST, TR+STR, and POSSTR was found ($p > 0.05$) when $s = 0.25$. These results indicate that the translation function is outputting traits as they would be observed by the witness incorrectly, possibly due to the lack of training data gathered from traits observed during interactions.

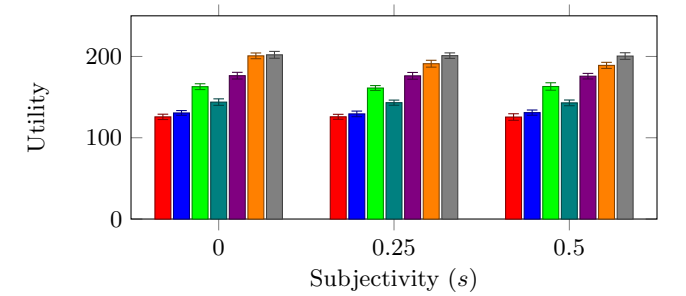
5.3 Private trustees

In this case, the identifier of the trustee agent being assessed was not disclosed to the witnesses when asking for reputation assessments. While this is extreme, a trustor may wish to keep their interest in particular trustee agents private for several reasons, including competition, embarrassment, or affects to reputation. For example, a trustor's interest in a particular doctor may reveal private health information or their interest in a particular subprovider may negatively affect their own reputation. This case is also representative of when trustees are regularly unknown to witnesses, such as when they are in a different locations. When unable to use witness-reputation, TR and TR+ST are equivalent to the T and T+ST strategies respectively, and therefore gained the same utilities. Utilities gained over simulation rounds for the five remaining strategies are presented in Figure 2, which includes utilities for POSSTR with training data for the translation function limited to (a) assessments, and (b) interactions. In all simulations, subjectivity parameters of $s = o = 0.5$ were used, and a significant difference in overall utility was observed between all pairs of strategies presented ($p < 0.01$). The TR+STR model is equivalent to using direct-trust and witness-stereotypes, or T+STR, causing its performance to drop significantly over this simulation compared to those in Sections 5.1 and 5.2.

After an initial learning phase lasting fewer than 10 rounds POSSTR (a), which learned the translation function using observations made in previous assessments, gained by far the highest utilities in each round. With less training data, POSSTR (b) gained lower utility, but still outperformed all the other



(a) Two available advisors



(b) Agents depart with probability of 0.2

Fig. 3: Utilities for strategies with different levels of subjectivity with $o = 0.5$, assessment partial observability, for (a) fewer advisors and (b) increased dynamism. The error bars show the STD.

strategies that either did not translate traits or did not use stereotypes. After around 50 rounds the utilities gained per round for all of the strategies stabilised. Interestingly, in this setting there was still a benefit to using witness-stereotypes (TR+STR) over just direct-trust (T), even though witnesses may have misinterpreted the observations made by the trustor. Without the translation function the best strategy was to use only direct evidence in the form of direct-trust and stereotype-trust (T+ST).

5.4 Fewer available advisors

The results in Figure 3(a) show the utilities gained when the number of witnesses available was reduced to two. In these results observations were limited to assessments, the subjectivity parameters were $o = 0.5$ and $s = 0$, 0.25, or 0.5. As a result, all strategies that use witness information gained less utility than with ten advisors. In these results using only direct-trust and witness-reputation

(TR) was outperformed by the combination of direct-trust and stereotype-trust, with $p < 0.01$. The extra utility gained by POSSTR compared to TR+STR when either $s = 0.25$ or $s = 0.5$, was also significantly ($p < 0.01$) more than with ten witnesses.

5.5 Increased dynamism

Figure 3(b) shows results for simulations with increased dynamism, where agents departed with an increased probability of 0.2. Again, in these results observability was restricted to assessments, $o = 0.5$, and $s = 0$, 0.25, or 0.5. All strategies other than Random performed less well in this setting, and gained lower utilities than in the less dynamic scenario where agents left with a probability of 0.05. Also in this highly dynamic setting POSSTR gained much more utility than the other strategies with subjectivities of $s = 0.25$ or 0.75 ($p < 0.01$).

5.6 Summary

In summary, we found POSSTR gained significantly more utility than all other strategies, including TR+STR, in environments where partial observability was combined with subjectivity. The difference was greatest when witnesses were unable to observe traits themselves due to the trustor withholding their identities, where the performance of POSSTR was affected much less than the other strategies. In simple environments, with either full observability or objective traits, the performance of POSSTR was not significantly different to that of TR+STR, and both gained more utility than the other strategies.

6 Conclusion

In this paper we have presented the POSSTR reputation system, which combines direct-trust, witness-reputation, stereotype-trust, and stereotype-reputation, and is robust to various levels of partial and subjective observability. Using simulations we have shown that a translation function is necessary when communicating observed traits to witnesses in partially observable and subjective environments. We found that POSSTR provided significantly more reliable reputation assessments compared to other strategies in such settings.

In settings without partial observability, where witnesses were able to observe all trustee traits themselves, the utilities gained by POSSTR and TR+STR, which both use direct-trust, witness-reputation, direct-stereotypes, and witness-stereotypes, were not significantly different. This was because the translation function employed in POSSTR has no effect when agents can observe the traits of all trustee agents. With no observability, where trustors concealed the identities of trustees they were assessing, using witness stereotypes without translation provided lower utilities than using only direct evidence. With translations, however, POSSTR was able to retain much of the performance observed in much less restricted settings with full observability.

Investigating subjectivity and dishonesty in interaction ratings is left as future work, but could be solved using a strategy such as TRAVOS [18] or HABIT [17]. Either of these strategies can be applied directly to witness-reputation described in this paper, but applying them to subjective-reputation may require some alterations. Another approach is to learn a mapping, akin to the translation function for observed traits, to translate reputation-assessments from one perspective to the other.

Another limitation is that concept drift, where the profile parameters or traits change over time, is not considered. To overcome such drift a learning window is often sufficient, but determining an appropriate window size is non-trivial. Another approach may be to apply techniques from the concept drift literature [5], to both detect when a change has occurred in the underlying profiles and adapt the model accordingly.

References

[1] Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (June 2007)

[2] Burnett, C., Norman, T., Sycara, K.: Stereotypical trust and bias in dynamic multiagent systems. ACM Transactions on Intelligent Systems and Technology 4(2) (March 2013)

[3] Fang, H., Zhang, J., Şensoy, M., Thalmann, N.: A generalized stereotypical trust model. In: International Conference on Trust, Security and Privacy in Computing and Communications. pp. 698–705 (June 2012)

[4] Frank, E., Hall, M., Witten, I.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 4th edn. (2016)

[5] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Computing Survey 46(4), 44:1–44:37 (April 2014)

[6] Hendrikx, F., Bubendorfer, K., Chard, R.: Reputation systems: A survey and taxonomy. Journal of Parallel and Distributed Computing 75, 184–197 (January 2015)

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

[8] Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 09(03), 279–311 (June 2001)

[9] Jøsang, A., Ismail, R.: The Beta reputation system. In: Electronic Commerce Conference. pp. 41–55 (2002)

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

[11] Koster, A., Schorlemmer, M., Sabater-Mir, J.: Engineering trust alignment: Theory, method and experimentation. International Journal of Human-Computer Studies 70(6), 450–473 (June 2012)

[12] Liu, X., Datta, A., Rzadca, K.: Trust beyond reputation: A computational trust model based on stereotypes. Electronic Commerce Research and Applications 12(1), 24–39 (January 2013)

[13] Quinlan, R.: Learning with continuous classes. In: Australian Joint Conference on Artificial Intelligence. pp. 343–348. World Scientific, Singapore (1992)
[14] Read, J., Reutemann, P., Pfahringer, B., Holmes, G.: Meka: A multi-label/multi-target extension to weka. Journal of Machine Learning Research 17(21), 1–5 (January 2016)
[15] Regan, K., Poupart, P., Cohen, R.: Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In: National Conference on Artificial Intelligence. vol. 2, pp. 1206–1212 (July 2006)
[16] Şensoy, M., Yilmaz, B.and Norman, T.: STAGE: Stereotypical trust assessment through graph extraction. Computational Intelligence 32(1), 72–101 (July 2016)
[17] Teacy, L., Luck, M., Rogers, A., Jennings, N.: An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. Artificial Intelligence 193, 149–185 (December 2012)
[18] Teacy, L., Patel, J., Jennings, N., Luck, M.: Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model. In: International Joint Conference on Autonomous Agents and Multiagent Systems. pp. 997–1004 (2005)
[19] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3(3), 64–74 (2007)
[20] Wahab, O., Bentahar, J., Otrok, H., Mourad, A.: A survey on trust and reputation models for web services: Single, composite, and communities. Decision Support Systems 74, 121–134 (June 2015)
[21] Wang, Y., Singh, M.: Formal trust model for multiagent systems. In: International Joint Conference on Artifical Intelligence. pp. 1551–1556. Morgan Kaufmann, San Francisco, CA, USA (January 2007)
[22] Wang, Y., Singh, M.: Evidence-based trust: A mathematical model geared for multiagent systems. ACM Transactions Autonomous Adaptive Systems 5(4), 14:1–14:28 (November 2010)
[23] Yu, H., Shen, Z., Leung, C., Miao, C., Lesser, V.: A survey of multi-agent trust management systems. IEEE Access 1, 35–50 (April 2013)
[24] Zhang, J., Cohen, R.: Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. Electronic Commerce Research and Applications 7(3), 330–340 (November 2008)