Bootstrapping Trust with Partial and Subjective Observability

(Extended Abstract)

Phillip Taylor, Nathan Griffiths
The University of Warwick
Coventry, CV4 7AL, UK
{name.surname}@warwick.ac.uk

Lina Barakat, Simon Miles
Kings College London
London, WC2R 2LS, UK
{name.surname}@kcl.ac.uk

ABSTRACT
Assessment of trust and reputation typically relies on prior experiences of a trustee agent, which may not exist, e.g. especially in highly dynamic environments. In these cases stereotypes can be used, where traits of trustees can be used as an indicator of their behaviour during interactions. Communicating observations of traits to witnesses who are unable to observe them is difficult, however, when the traits are interpreted subjectively. In this paper we propose a mechanism for learning translations between such subjective observations, evaluating it in a simulated marketplace.

CCS Concepts
• Computing methodologies → Multi-agent systems;

Keywords
Trust and reputation, Stereotypes, Machine learning

1. INTRODUCTION
In multi-agent systems agents can use trust and reputation to decide which others to interact with [3,18,20]. Trust is the degree of belief, from the trustor agent’s perspective, that a trustee agent will act as they say they will in a given context [1,2,8]. Whereas trust is assessed using experiences of the trustee, reputation is based on the opinions of several agents. In assessing a trustee, the trustor combines:

Direct-trust — trustee experiences with the trustee
Witness-reputation — witness reports of the trustee
Stereotype-trust — trustee experiences and observed traits
Stereotype-reputation — witness reports of observed traits

Direct-trust requires the trustor to have previously interacted with the trustee. The same is true for witnesses when computing witness-reputation, where the trustor requests opinions about a specific trustee. In combination, direct-trust and witness-reputation, make up the Beta Reputation System (BRS), as proposed by Josang et al. [7]. Other reputation systems that combine these include FIRE [4,5], TRAVOS [16], BLADE [13], and HABIT [15].

2. POSSTR MODEL
An interaction is recorded as \( \langle tr, tc, \theta^{te}_{tc}, r^{te}_{tr} \rangle \), where \( \theta^{te}_{tc} \) are observations of \( tc \) by \( tr \) before the interaction, and \( r^{te}_{tr} \) is the
rating given by \( tr \). We assume ratings are binary, with 1 indicating a good outcome and 0 indicating otherwise.

As with STAGE \[14\] and Burnett et al. \[2\], POSSTR is based on BRS \[7\], which represents and processes opinions using subjective logic \[6\]. The overall trust value is computed as the likelihood of a successful interaction,

\[
P(r_{tr}^e = 1) = \frac{p + 2a}{p + n + 2},
\]

where \( p \) is the number of previous successful interactions, \( n \) is the number of unsuccessful ones, and \( a \) is a Bayesian prior. For direct-trust, \( p \) and \( n \) are calculated using interactions recorded by \( tr \). In witness-reputation, the interactions reported by witnesses are combined when calculating \( p \) and \( n \). In BRS \[7\], \( a = 0.5 \) to represent equal likelihood of a successful and unsuccessful outcome with no information. With stereotypes \[2,14\], the value of \( a \) is the output of models that map trustee traits to trustworthiness, \( f : P(\Theta) \rightarrow \mathbb{R} \).

For stereotype-trust, \( f_{tr} \) is learned by generating a training sample for each agent \( tr \) has previously interacted with. The input features are the observed traits, \( \theta_{tr}^w \), and the target is the direct-trust that \( tr \) has in \( te \), with a prior of \( a = 0.5 \). An M5 model tree \[11\] is then learned to map traits observed by \( tr \) to the trust in agents that express those traits. M5 recursively splits training samples using the values of the features that best discriminate the target. Whereas in typical decision trees the leaves are target values, leaves in M5 are regression models that output the target value, which is taken as the stereotype-trust prior, \( a = \alpha_{tr} = f_{tr}(\theta_{tr}^w) \).

When assessing stereotype-reputation, a witness may not have observed the traits themselves, meaning a translation function, \( f_{tr \rightarrow w} : P(\Theta) \rightarrow P(\Theta) \), is required and learned from training data generated from common observations that both the witness and the trustee have made. When requesting a stereotype assessment from a witness, either the trustee 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 trustee has observed the traits of three trustees, \( \{\theta_{tr1}^{e1}, \theta_{tr1}^{e2}, \theta_{tr1}^{e3}\} \), and a witness has observed those of two, \( \{\theta_{w1}^{e1}, \theta_{w2}^{e2}\} \). Training data can then be generated by matching up the common observations, as \( \{\theta_{tr1}^{e1} : \theta_{w1}^{e1}, \theta_{tr1}^{e2} : \theta_{w2}^{e2}\} \), where ‘:’ separates the inputs and outputs. These observations may have been made without having interacted with the trustees, such as during a reputation assessment without using the service. These common observations samples form the training data that can be input into a multi-target learning algorithm \[12\].

In this paper the binary relevance method \[17\] is used, and Naive Bayes \[12,19\] models are learned to map traits observed by the trustee to each trait that would be observed by the witness. The traits observed by the trustee are then input into each of the learned models and their outputs are combined to be the traits the witness would have observed.

The Bayesian prior for stereotype-reputation is then computed as the mean over the trustee and witnesses,

\[
a = \frac{1}{|W| + 1} \left( a_{tr} + \sum_{w \in W} a_{w} \right),
\]

and is used in Equation 1, where,

\[
a_{w} = \begin{cases} f_{w}(\theta_{w}^{e}) & \text{if witness observed trustee}, \\ f_{w}(f_{tr \rightarrow w}(\theta_{tr}^{e})) & \text{otherwise}. 
\end{cases}
\]

3. RESULTS

POSSTR was evaluated against combinations of direct-trust (T), witness-reputation (R), stereotype-trust (ST) and stereotype-reputation (STR) using a simulated marketplace over 250 rounds, consisting 100 trustee agents and 20 trustee agents. In each round, each trustee selected from 10 random trustees using reputation gathered from a random 10 witnesses (other trustee agents). The trustee with highest reputation was selected for an interaction, where the outcome utility was drawn from a normal distribution assigned to the trustee at the beginning. If the outcome utility was greater than 0.5, the interaction was rated as a success and otherwise it was rated as unsuccessful. All trustee and trustee agents left the simulation in each with a probability of 0.05, to be replaced by another.

Figure 1 shows the total utility gained by trustees under different levels of observability. The results are averaged over 50 simulations and the error bars represent the standard deviation. In general, using both direct-trust alongside witness-reputation gained higher utilities than using direct-trust only. Similarly, using stereotypes led to higher utilities than using direct-trust and witness-reputation only. In full observability, witnesses were always able to observe traits themselves when reporting for stereotype-reputation, leading to no benefit in the translation function used by POSSTR over TR+STR. With limited observability, where witnesses were able to only recall traits observed during their own previous assessments (as a trustor), the TR+STR strategy gained less utility and was outperformed by POSSTR. To limit the observability further we had the trustor conceal the identity of the trustee being assessed from witnesses, meaning that witness-reputation was not possible and stereotype-reputation assessments relied on the traits observed subjectively by the trustee. Here, TR+STR gained much less utility than with more observability, whereas POSSTR was again unaffected and performed the best.

4. CONCLUSION

In this paper we have presented the POSSTR reputation system, that is robust in subjective and partially observable environments. We found that in environments with full observability or with objectively observable traits, POSSTR performed equally well as TR+STR. With decreased observability and with subjective traits, the translation function in POSSTR allowed it to make reliable reputation assessments.
REFERENCES


