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 trustor, reputation is based on the opinions of several agents. In assessing a trustee, the trustor combines:

Direct-trust — trustor experiences with the trustee
Witness-reputation — witness reports of the trustee
Stereotype-trust — trustor 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 Jøsang et al. [7]. Other reputation systems that combine these include FIRE [4,5], TRAVOS [16], BLADE [13], and HABIT [15].

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

Stereotype-trust enables assessments of trustees where direct experience is limited, by assuming trustees with similar traits behave similarly. Similar to witness-reputation, stereotype-reputation is gathered from witnesses. Liu et al. [9,10] proposed that trustee traits, correlated with their trustworthiness, be used to separate them into groups defined by their common characteristics. Instead of clustering trustees, Burnett et al. [2] use regression models to map traits to trustworthiness, and use predictions as base values in a probabilistic trust model. When stereotype-trust is unavailable, because of a lack of data, stereotype-reputation is gathered from witnesses. This model is extended by Şensoy et al. [14] to discount witness reports, and bootstrap this discounting using stereotypes of witnesses.

Existing reputation models require the following:

- The trustee is identified and the witness can observe its traits (i.e. trustees are *fully observable*), or
- 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 observations may be *subjective*. When trustees are *partially observable* and the witness is unable to observe the traits, the trustor must disclose their observations for the witness to process them. If a new trustee is unknown to a witness, for example, the trustor must describe their observations. If trait observations are also *subjective*, those observed by a trustor may be meaningless to a witness.

We define the set of traits in an environment as $\mathbf{\Theta}$, which may include 'airport transfer' and 'suitcase storage' for a taxi marketplace. Each trustee agent, te, exhibits a subset of these traits, $\theta^{te} \subseteq \mathbf{\Theta}$, and each trustor agent, tr, has an observation function, $\mathcal{O}_{tr} : \mathcal{P}(\mathbf{\Theta}) \rightarrow \mathcal{P}(\mathbf{\Theta})$, which determines their observations of a trustee, $\theta^{te}_{tr} = \mathcal{O}_{tr}(\theta^{te})$.

In fully observable settings, witnesses can always observe traits of trustees, $\theta^{te}_{w} = \mathcal{O}_{w}(te)$, and correctly interpret any associated stereotype. With partial observability witnesses may be unable to observe the traits of a trustee and must process those observed by the trustor, which, if objective, poses no problem. With subjectivity, however, observations made by a trustor may be different to those that a witness would make. Two customers may have different interpretations of suitable suitcase storage for a taxi, for example, leading to misunderstandings when communicating traits.

2. POSSTR MODEL

An interaction is recorded as $\langle tr, te, \theta^{te}_{tr}, r^{te}_{tr} \rangle$, where θ^{te}_{tr} are observations of te 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,

$$P(\hat{r}_{tr}^{te} = 1) = \frac{p + 2a}{p + n + 2}, \quad (1)$$

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, $f : \mathcal{P}(\mathbf{\Theta}) \rightarrow \mathcal{R}$.

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, θ_{tr}^{te}, 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, $a = a_{tr} = f_{tr}(\theta_{tr}^{te})$.

When assessing *stereotype-reputation*, a witness may not have observed the traits themselves, meaning a translation, $f_{tr \rightarrow w} : \mathcal{P}(\mathbf{\Theta}) \rightarrow \mathcal{P}(\mathbf{\Theta})$, is required and learned from training data 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 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 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.

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

$$a = \frac{1}{|W| + 1}\left(a_{tr} + \sum_{w \in W} a_w\right), \quad (2)$$

and is used in Equation 1, where,

$$a_w = \begin{cases} f_w(\theta_w^{te}) & \text{if witness observed trustee,} \\ f_w(f_{tr \rightarrow w}(\theta_{tr}^{te})) & \text{otherwise.} \end{cases} \quad (3)$$

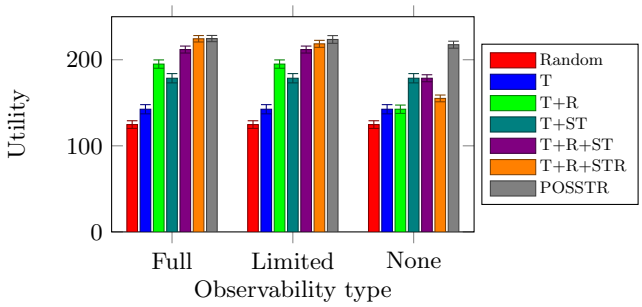

Figure 1: Utility under different observabilities.

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 trustor agents. In each round, each trustor selected from 10 random trustees using reputation gathered from a random 10 witnesses (other trustor 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 trustor 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 trustors 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 only to 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 trustor. 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

[1] D. Artz and Y. Gil. 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] C. Burnett, T. Norman, and K. Sycara. Stereotypical trust and bias in dynamic multiagent systems. *ACM Transactions on Intelligent Systems and Technology*, 4(2):26, 2013.

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

[4] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. FIRE: An integrated trust and reputation model for open multi-agent systems. In *16th European Conference on Artificial Intelligence*, pages 18–22, 2004.

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

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

[7] A. Jøsang and R. Ismail. The Beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, pages 41–55, 2002.

[8] 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.

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

[10] X. Liu, A. Datta, K. Rzadca, and E.-P. Lim. Stereotrust: A group based personalized trust model. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 7–16, New York, NY, USA, 2009. ACM.

[11] R. J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

[12] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes. Meka: A multi-label/multi-target extension to weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016.

[13] K. Regan, P. Poupart, and R. Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1206–1212, 2006.

[14] M. Şensoy, B. Yilmaz, and T. J. Norman. STAGE: Stereotypical trust assessment through graph extraction. *Computational Intelligence*, 32(1):72–101, 2016.

[15] W. L. Teacy, M. Luck, A. Rogers, and N. R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. *Artificial Intelligence*, 193:149–185, 2012.

[16] W. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Coping with inaccurate reputation sources: Experimental analysis of a probabilistic trust model. In *International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 997–1004, 2005.

[17] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1, 2007.

[18] O. Wahab, J. Bentahar, H. Otrok, and A. Mourad. A survey on trust and reputation models for web services: Single, composite, and communities. *Decision Support Systems*, 74:121 – 134, 2015.

[19] I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Jan 2011.

[20] H. Yu, Z. Shen, C. Leung, C. Miao, and V. Lesser. A survey of multi-agent trust management systems. *IEEE Access*, 1:35–50, April 2013.