Talking about Trust in Heterogeneous Multi-Agent Systems*

Andrew Koster Jordi Sabater-Mir† Marco Schorlemmer†
IIIA - CSIC
Universitat Autònoma de Barcelona
Bellaterra, Spain

Abstract

In heterogeneous multi-agent systems trust is necessary to improve interactions by enabling agents to choose good partners. Most trust models work by taking, in addition to direct experiences, other agents' communicated evaluations into account. However, in an open MAS other agents may use different trust models and the evaluations they communicate are based on different principles: as such they are meaningless without some form of alignment. My doctoral research gives a formal definition of this problem and proposes two methods of achieving an alignment.

1 Introduction

Trust models are a fundamental tool for agents to achieve effective interactions in an open MAS. However, it is not as straightforward as equipping an agent with one of the available computational trust models and expecting it to function in a social environment. Using trust as a method for picking successful interaction partners relies not only on having a good trust model, but also on communication with other agents [Conte and Paolucci, 2002]. Trust is an inherently subjective concept, meaning that any computational agent functioning in a multi-agent society may base its trust on different personal values from any other agent. As such, one agent's trust may be different from another agent's trust, despite using the exact same factual evidence to support their trust evaluations. If trust evaluations - and other subjective opinions - are to be communicated effectively in an open MAS, a different set of tools is required than is used for the communication of facts [Koster, 2010].

My doctoral research focuses on studying communication about trust as a separate problem from modeling trust in intelligent agents. This problem is so far unexplored, with previous work limited to [Abdul-Rahman and Hailes, 2000] and [Regan *et al.*, 2006]. Both of these works consider the alignment as a part of the trust model, rather than a separate mechanism of interpreting incoming communication, which limits their applicability. To address this limitation, the first step of the research was to formalize *what* the problem of trust alignment is and *how* a solution can be found. This formalization is described in previous work [Koster *et al.*, 2010a]. This extended abstract summarizes this formalization as well as a practical method, based upon it. The last part will consider the future direction of the research, dealing with argumentation about trust.

*This work is supported by the Generalitat de Catalunya grant 2009-SGR-1434 and the Spanish Ministry of Education: Agreement Technologies project: CONSOLIDER CSD2007-0022, INGENIO 2010 and CBIT project: TIN2010-16306

†Supervisors

2 Formalizing Trust Alignment

Trust alignment is the process of finding a translation of the other agent's trust evaluations, based on shared evidence. Its result is a method to translate other trust evaluations from the same agent, based on non-shared evidence. With evidence I mean an objective description of some artifacts in the environment, such as interactions the agents have participated in. Shared evidence is evidence of artifacts in the environment that both agents have perceived, while non-shared evidence refers to artifacts that the receiving agent has not perceived. By using the shared evidence as a common ground, two agents can communicate their differing trust evaluations based on the same evidence and use these different evaluations of the same object as the starting point for finding a translation. This definition is grounded in Channel Theory [Barwise and Seligman, 1997], which also serves as a mathematical framework for formalizing other problems of Semantic Alignment. For the specifics see Koster et al. [2010a], in which the formalization is described in detail. The key points, however, are the realization that trust models can be considered as an abstract mapping from evidence in the environment to an evaluation of a trustee. This mapping is subjective, but the evidence in the environment can be described in objective terms. I thus do not consider the question of how trust is modeled within an agent, nor do I consider how it is used by an agent, but rather I focus on the question: if an agent communicates its subjective trust evaluation, how can the receiver interpret this?

3 Learning an Alignment

As mentioned previously, there is some previous work in the area of Trust Alignment and both these works can be considered methods of solving the alignment problem with a machine learning algorithm. Abdul-Rhaman & Hailes' work [2000] considers models which use numerical trust evaluations. Interpreting another agent's communicated trust evaluation is done by adding or subtracting a numerical bias from

this evaluation. This bias is learned by averaging the difference between the own and other's evaluations of the same agent in the past. This simple approach seems to work surprisingly well, however one thing to realize is that the evidence for the trust evaluation is not taken into account in this alignment. An objective description of the evidence which supports a trust evaluation allows agents to communicate not just the evaluation, but also the context in which it was made. Thus, alignment should improve if such evidence is taken into account: the subjective trust evaluations are explicitly linked to an objective description of the environment. BLADE [Regan *et al.*, 2006] uses a conjunction of proposition symbols to represent the evidence and a Bayesian Inference Learner to learn an alignment. However this is still very limited: a conjunction of propositions allows them to classify interactions by the concepts they represent, but it does not allow for the specification of relations between the different concepts. I therefore propose to use a first order logic to describe the evidence and an Inductive Logic Programming algorithm to learn the alignment [Koster *et al.*, 2010b]. This work has been improved on and empirically evaluated [Koster *et al.*, 2010a].

4 Arguing about Trust

Rather than attempt to find a translation between the other's and own trust evaluations, an alternative approach is to attempt to convince the other *why* the own trust evaluation is correct: as such future evaluations from that agent will follow the same reasoning. Argumentation about trust thus far has focused on arguing about whether or not a communicated trust evaluation ought to be accepted or not [Pinyol and Sabater-Mir, 2010; Matt *et al.*, 2010]. For arguing about the models themselves, a new way of looking at computational trust models is necessary. Current models can be considered monolithic: evidence is provided as an input and the model performs calculations, resulting in a trust evaluation as output. To be able to argue about why one trust evaluation is better than another the first step is for an agent to be able to reason about its own trust model. Current work is moving in this direction [Pinyol *et al.*, 2008; Castelfranchi and Falcone, 2010]. I consider an agent based on multi-context systems as in Pinyol et al's work [2008] and have extended this model to allow for the specification of trust models in a reflective manner. I aim to implement this, allowing an agent to pro-actively adapt its trust model. Using this specification of the trust model, the next step is to define an argumentation protocol, that allows agents to convince each other of the correctness of their trust model, given the specific domain the agents inhabit, and to achieve alignment in this manner. Eventually the two methods might be combined, providing an agent with a number of options to interpret another agent's trust, or even use a combination of methods. In the long term I intend to explore the interplay of the various alignment methods.

5 Conclusion

The doctoral research described focuses on the problem of Trust Alignment. The first part of this research was to formalize it and the second part, which is still ongoing, is to provide different methods for agents to solve the problem of Trust Alignment. The first of these is to use machine learning techniques to learn a translation between the other's and one's own trust evaluation, and the second is using argumentation to convince the other to adopt one's own trust model. The aim is to eventually be able to use a mix of both methods, depending on the situation encountered.

References

[Abdul-Rahman and Hailes, 2000] A. Abdul-Rahman and S. Hailes. Supporting trust in virtual communities. *33rd Hawaii Conference on System Sciences*, pages 4–7, 2000.

[Barwise and Seligman, 1997] J. Barwise and J. Seligman. *Information Flow: The Logic of Distributed Systems*. Cambridge University Press, 1997.

[Castelfranchi and Falcone, 2010] C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-cognitive and Computational Model*. Wiley, 2010.

[Conte and Paolucci, 2002] R. Conte and M. Paolucci. *Reputation in Artificial Societies: Social beliefs for social order*. Kluwer, 2002.

[Koster *et al.*, 2010a] A. Koster, J. Sabater-Mir, and M. Schorlemmer. Engineering trust alignment: Theory and practice. Technical Report TR-2010-02, CSIC-IIIA, 2010.

[Koster *et al.*, 2010b] A. Koster, J. Sabater-Mir, and M. Schorlemmer. Inductively generated trust alignments based on shared interactions (extended abstract). In *AAMAS'10*, pages 1571–1572, Toronto, Canada, 2010. IFAAMAS.

[Koster, 2010] A. Koster. Why does trust need aligning? In *Proc. of 13th Workshop "Trust in Agent Societies"*, pages 125–136, Toronto, 2010. IFAAMAS.

[Matt *et al.*, 2010] P. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *AAMAS'10*, pages 209–216, Toronto, Canada, 2010. IFAAMAS.

[Pinyol and Sabater-Mir, 2010] I. Pinyol and J. Sabater-Mir. An argumentation-based protocol for social evaluations exchange. In *ECAI'10*, Lisbon, Portugal, 2010.

[Pinyol *et al.*, 2008] I. Pinyol, J. Sabater-Mir, and P. Dellunde. Cognitive social evaluations for multi-context bdi agents. In *CCIA'08*, 2008.

[Regan *et al.*, 2006] K. Regan, P. Poupart, and R. Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *AAAI'06*, pages 1206–1212, Boston, MA, USA, 2006. AAAI Press.