An abstract framework for reasoning about trust

(Extended Abstract)

Elisabetta Erriquez, Wiebe van der Hoek, Michael Wooldridge
Department of Computer Science, University of Liverpool, United Kingdom
{e.erriquez, Wiebe.Van-Der-Hoek, mjw} @liverpool.ac.uk

ABSTRACT

We present an abstract framework that allows agents to form coalitions with agents that they believe to be trustworthy. In contrast to many other models, we take the notion of *distrust* to be our key social concept. We use a graph theoretic model to capture the distrust relations within a society, and use this model to formulate several notions of mutually trusting coalitions. We then investigate principled techniques for how the information present in our distrust model can be aggregated to produce individual measures of how trustworthy an agent is considered to be by a society.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems; I.2.4 [**Knowledge representation formalisms and methods**]

General Terms

Theory

Keywords

models of trust, society models

1. INTRODUCTION

The goal of coalition formation is typically to form robust, cohesive groups that can cooperate to the mutual benefit of all the coalition members. With a relatively small number of exceptions, existing models of coalition formation do not generally consider trust [1, 5]. In more general models [6, 4], individual agents use information about reputation and trust to rank agents according to their level of trustworthiness. Therefore, if an agent decides to form a coalition, it can select those agents he reckons to be trustworthy. Or, alternatively, if an agent is asked to join a coalition, he can assess his trust in the requesting agent and decide whether or not to run the risk of joining a coalition with him.

However, we argue that these models lack a *global* view. They only consider the trust binding the agent starting the coalition and the agents receiving the request to join the coalition. In this paper, we address this limitation. We propose an abstract framework through which autonomous, self-interested agents can form coalitions based on information relating to trust. In fact, we use *distrust* as the key social concept in our work. We focus on how distrust can be used as a mechanism for modelling and reasoning about the reliability of others, and, more importantly, about how to form coalitions that satisfy some stability criteria. We present several notions of mutually trusting coalitions and define different measures to aggregate the information presented in a distrust model.

Cite as: An abstract framework for reasoning about trust (Extended Abstract), E. Erriquez, W. van der Hoek and M. Wooldridge, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1085-1086.

2. A FRAMEWORK BASED ON DISTRUST

Our approach is inspired by the abstract argumentation frameworks of Dung [2]. Essentially, Dung was interested in trying to provide a framework that would make it possible to make sense of a domain of discourse on which there were potentially conflicting views. He considered the various conflicting views to be represented in *arguments*, with an *attack relation* between arguments defining which arguments were considered to be inconsistent with each other. In our work, we use similar graph like models, but rather than arguments our graph is made up of agents, and the binary relation (which is used in determining which coalitions are acceptable), is a *distrust* relation.

A *distrust* relation between agent i and agent j is intended as agent i having none or little trust in agent j. More precisely, when saying that agent i distrusts agent j we mean that, in the context at hand, agent i has insufficient confidence in agent j to share membership with j in one and the same coalition.

The follow definitions characterize our formal model.

DEFINITION 1. *An* Abstract Trust Framework *(*ATF*), S, is a pair: $S = \langle Ag, \rightsquigarrow \rangle$ where: Ag is a finite, non-empty set of* agents*; and $\rightsquigarrow \subseteq Ag \times Ag$ is a binary* distrust *relation on Ag.*

When $i \rightsquigarrow j$ we say that agent i distrusts agent j. We assume $\rightsquigarrow$ to be irreflexive, i.e., no agent i distrusts itself. Whenever i does not distrust j, we write $i \not\rightsquigarrow j$. So, we assume $\forall i \in Ag$, $i \not\rightsquigarrow i$. Call an agent i *fully trustworthy* if for all $j \in Ag$, we have $j \not\rightsquigarrow i$. Also, i is *trustworthy* if for some $j \neq i$, $j \not\rightsquigarrow i$ holds. Conversely, call i *fully trusting* if for no j, $i \rightsquigarrow j$. And i is *trusting* if for some $j \neq i$, $i \not\rightsquigarrow j$.

In what follows, when we refer to a "coalition" it should be understood that we mean nothing other than a subset C of Ag. When forming a coalition, there are several ways to measure how much distrust there is among them, or how trustable the coalition is with respect to the overall set of agent Ag.

DEFINITION 2. *Given an* ATF $S = \langle Ag, \rightsquigarrow \rangle$*, a coalition $C \subseteq Ag$ is* distrust-free *if no member of C distrusts any other member of C. Note that the empty coalition and singleton coalitions $\{i\}$ are distrust-free: we call them trivial coalitions.*

Distrust freeness can be thought of as the most basic requirement for a *trusted* coalition of agents. It means that a set of agents has

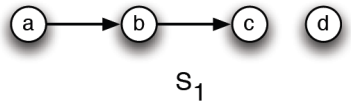

Figure 1: An ATF for four agents

no internal distrust relationships between them. Since we assume $\rightsquigarrow$ to be irreflexive, we know that for any $i \in Ag$, the coalition $\{i\}$ is distrust-free, as is the empty coalition. A distrust-free coalition for S_1 in Figure 1 is, for example, $\{a, c, d\}$. Consider ATF S_5 from Figure 1. The coalition $C_1 = \{c, d\}$ is distrust-free, but still, they are not angelic: one of their members is being distrusted by some agent in Ag, and they do not have any justification to ignore that. With this in mind, we define the following concepts.

DEFINITION 3. *Let* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *be given. An agent* $i \in Ag$ *is called* trustable *with respect to a coalition* $C \subseteq Ag$ *iff* $\forall y \in Ag((y \rightsquigarrow i) \Rightarrow \exists x \in C(x \rightsquigarrow y))$. *A coalition* $C \subseteq Ag$ *is a* trusted extension *of* S *iff* C *is distrust-free and every agent* $i \in C$ *is trustable with respect to* C. *A coalition* $C \subseteq Ag$ *is a* maximal trusted extension *of* S *if* C *is a trusted extension, and no superset of* C *is one.*

The concept of a *trusted extension* represents a basic and important notion for agents who want to rationally decide who to form a coalition with, basing their decisions on trust. In particular: *a trusted extension is composed of agents that have a rational basis to trust each other*.

It is possible that a particular ATF has more than one maximal trusted extension. One could assume that all the agents in the maximal trusted extensions are equally trustworthy. One way to address this is to consider how many times a particular agent occurs in the maximal trusted extensions. If one agent occurs in more than one maximal trusted extension, then we can take this as an evidence it is somehow more "trustworthy" than another agent occuring in just one.

With this in mind, we define the following concepts.

DEFINITION 4. *Let* ATF $S = \langle Ag, \rightsquigarrow \rangle$ *be given. An agent* $i \in Ag$ *is* Strongly Trusted *if it is a member of* every *maximal trusted extension. An agent* $i \in Ag$ *is* Weakly Trusted *if it is a member of at least* one *maximal trusted extension.*

The notion of strongly and weakly trusted can help agents decide in those situation where there are large maximal trusted extensions but not all the agents are required for forming a stable coalition.

3. AGGREGATE TRUST MEASURES

Abstract trust frameworks provide a social model of (dis)trust. An obvious question, however, is how the information presented in abstract trust frameworks can be *aggregated* to provide a single measure of how trustworthy (or otherwise) an individual within the society is. We present two aggregate measures of trust, which are given relative to an abstract trust framework $S = \langle Ag, \rightsquigarrow \rangle$ and an agent $i \in Ag$. Both of these trust values attempt to provide a principled way of measuring the overall trustworthiness of agent i, taking into account the information presented in S:

- *Expected trustworthiness*:

  This value is the ratio of the number of maximal trusted extensions of which i is a member to the overall number of maximal trusted extensions in the system S.To put it another way, this value is the probability that agent i would appear in a maximal trusted extension, if we picked such an extension uniformly at random from the set of all maximal trusted extensions. Formally, letting $mte(S)$ denote the set of maximal trusted extensions in $S = \langle Ag, \rightsquigarrow \rangle$, the expected trustworthiness of agent $i \in Ag$ is denoted $\mu_i(S)$, defined as:

$$\mu_i(S) = \frac{|\{C \in mte(S) \mid i \in C\}|}{|mte(S)|}.$$

- *Coalition expected trustworthiness*:

  This value attempts to measure the probability that an agent $i \in Ag$ would be trusted by an arbitrary coalition, picked from the overall set of possible coalitions in the system. To define this value, we need a little more notation. Where $R \subseteq X \times X$ is a binary relation on some set X and $C \subseteq X$, then we denote by $restr(R, C)$ the relation obtained from R by restricting it to C:

$$restr(R, C) = \{(s, s') \in R \mid \{s, s'\} \subseteq C\}.$$

  Then, where $S = \langle Ag, \rightsquigarrow \rangle$ is an abstract trust framework, and $C \subseteq Ag$, we denote by $S \downarrow C$ the abstract trust framework obtained by restricting the distrust relation $\rightsquigarrow$ to C:

$$S \downarrow C = \langle C, restr(\rightsquigarrow, C) \rangle.$$

  Given this, we can define the *coalition expected trustworthiness*, $\varepsilon_i(S)$, of an agent i in given an abstract trust framework $S = \langle Ag, \rightsquigarrow \rangle$ to be:

$$\varepsilon_i(S) = \frac{1}{2^{|Ag|-1}} \sum_{C \subseteq Ag \setminus \{i\}} \mu_i(S \downarrow C \cup \{i\}).$$

  Thus, $\varepsilon_i(S)$ measures the expected value of μ_i for a coalition $C \cup \{i\}$ where $C \subseteq Ag \setminus \{i\}$ is picked uniformly at random from the set of all such possible coalitions. There are $2^{|Ag|-1}$ coalitions not containing i, hence the first term in the definition.

These two values are related to solution concepts such as the Banzhaf index, developed in the theory of cooperative games and voting power, and indeed they are inspired by these measures [3].

4. REFERENCES

[1] Silvia Breban and Julita Vassileva. Long-term coalitions for the electronic marketplace. In *Proceedings of the E-Commerce Applications Workshop, Canadian AI Conference*, 2001.

[2] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *AI*, 77:321–357, 1995.

[3] D. S. Felsenthal and M. Machover. *The Measurement of Voting Power*. Edward Elgar: Cheltenham, UK, 1998.

[4] Nathan Griffiths and Michael Luck. Coalition formation through motivation and trust. *In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*, 2003.

[5] Guo Lei, Wang Xiaolin, and Zeng Guangzhou. Trust-based optimal workplace coalition generation. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, pages 1 – 4, 2009.

[6] Zhou Qing-hua, Wang Chong-jun, and Xie Jun-yuan. Core: A trust model for agent coalition formation. In *Natural Computation, 2009. ICNC '09. Fifth International Conference on*, volume 5, pages 541 –545, 2009.