Argumentation-based reasoning in agents with varying degrees of trust

Simon Parsons
Brooklyn College
City University of New York

Yuqing Tang
Graduate Center
City University of New York

Elizabeth Sklar
Brooklyn College
City University of New York

Peter McBurney
Department of Informatics
King's College London

Kai Cai
Graduate Center
City University of New York

ABSTRACT

In any group of agents, trust plays an important role. The degree to which agents trust one another will inform what they believe, and, as a result the reasoning that they perform and the conclusions that they come to when that involves information from other agents. In this paper we consider a group of agents with varying degrees of trust of each other, and examine the combinations of trust with the argumentation-based reasoning that they can carry out. The question we seek to answer is "What is the relationship between the trust one agent has in another and the conclusions that it can draw using information from that agent?", and show that there are a range of answers depending upon the way that the agents deal with trust.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Coherence & co-ordination; languages & structures; multiagent systems.*

General Terms

Language, theory.

Keywords

Argumentation; Logic-based approaches and methods; Trust, reliability and reputation.

1. INTRODUCTION

Trust is an approach for measuring and managing the uncertainty about autonomous entities and the information they deal with. As a result trust can play an important role in any decentralized system. As computer systems have become increasingly distributed, and control in those systems has become more decentralized, trust has steadily become more important in computer science [5, 11].

Thus, for example, we see work on trust in peer-to-peer networks, including the EigenTrust algorithm [15] — a variant of PageRank [19] where downloads from a source play the role of outgoing hyperlinks and which is effective in excluding peers who want to disrupt the network — and the work in [1] that prevents peers manipulating their trust values to get preferential downloads. Zhong *et al.* [29] are concerned with slightly different issues in mobile ad-hoc networks, looking to prevent nodes from getting others to transmit their messages while refusing to transmit the messages of others, thus enforcing trustworthy behavior.

The internet, as the largest distributed system of all, is naturally a target of much of the research on trust. There have, for example, been studies on the development of trust in ecommerce [22], on mechanisms to determine which sources to trust when faced with multiple conflicting sources [28], and mechanisms for identifying which individuals to trust based on their past activity [2]. One interesting development is the idea of having individuals indemnify each other by placing some form of financial guarantee on transactions that others enter into [7, 8].

Trust is an especially important issue from the perspective of autonomous agents and multiagent systems [26]. The premise behind the multiagent systems field is that of developing software agents that will work in the interests of their "owners", carrying out their owners' wishes while interacting with other entities. In such interactions, agents will have to reason about the degree to which they should trust those other entities, whether they are trusting those entities to carry out some task, or whether they are trusting those entities to not misuse crucial information. As a result we find much work on trust in agent-based systems [24].

In such work it is common to assume that agents maintain a *trust network* of their acquaintances, which includes ratings of how much those acquaintances are trusted, and how much those acquaintances trust their acquaintances, and so on. An important line of inquiry in this context is what inference is reasonable in such networks, and the propagation of trust and provenance — both the transitivity of trust relations [23, 27] and more complex relationships like "co-citation" [12] have been studied, and in some cases empirically validated [12, 16, 28].

In this paper we look at the use of trust in other aspects of the reasoning that agents carry out. Argumentation [6] is a model of reasoning that seems well-suited to agent-based systems — it is robust against inconsistency, handles decision-making under uncertainty, and supports inter-agent communication. [20] suggests that argumentation is a suitable mechanism for reasoning about trust, and [18] shows how argumentation can be used to track trust in acquaintances. Here we investigate the combination of trust measures on agents and the use of argumentation for reasoning about belief, combining an existing system for reasoning about trust and an existing system of argumentation.

Cite as: Argumentation-based reasoning in agents with varying degrees of trust, S. Parsons, Y. Tang, E. Sklar, P. McBurney and K. Cai, *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. XXX-XXX.

2. FORMAL MODEL

This paper deals with combining two formal models — a model of trust and a model of argumentation — and we introduce both here. Though there is no standard for either kind of model, we built as generic a model of both trust and argumentation as we could, drawing from well-established models in the literature. As a result we have a combined model that has a number of features unspecified — in later sections we will examine various instantiations.

2.1 Trust

We are interested in a finite set of agents Ags and how these agents trust one another. Following the usual presentation (for example [16, 27, 23]), we start with a *trust relation*:

$$\tau \subseteq Ags \times Ags$$

which identifies which agents trust one another. If $\tau(Ag_i, Ag_j)$, where $Ag_i, Ag_j \in Ags$, then Ag_i trusts Ag_j. This is not a symmetric relation, so it is not necessarily the case that $\tau(Ag_i, Ag_j) \Rightarrow \tau(Ag_j, Ag_i)$. It is natural to represent this trust relation as a directed graph, and we have:

DEFINITION 1. *A* trust network *is a graph comprising, respectively, a set of nodes and a set of edges:*

$$\mathcal{T} = \langle Ags, \{\tau\}\rangle$$

where Ags is a set of agents and $\{\tau\}$ is the set of pairwise trust relations over Ags so that if $\tau(Ag_i, Ag_j)$ is in $\{\tau\}$ then $\{Ag_i, Ag_j\}$ is a directed arc from Ag_i to Ag_j in $\mathcal{T}$.

In this graph, the set of agents is the set of vertices, and the trust relations define the arcs. We are typically interested in *minimal* trust networks, which are connected — these thus capture the relationship between a set of agents all of whom, in one way or another are connected by a "web of trust". A directed path between agents in the trust network implies that one agent indirectly trusts another. For example if:

$$\langle Ag_1, Ag_2, \ldots Ag_n\rangle$$

is a path from agent Ag_1 to Ag_n, then we have:

$$\tau(Ag_1, Ag_2), \tau(Ag_2, Ag_3), \ldots, \tau(Ag_{n-1}, Ag_n)$$

and the path gives us a means to compute the trust that Ag_1 has in Ag_n. Below we will make use of the function $length(\cdot)$ which returns the number of agents in a path: $length(\langle Ag_1, Ag_2, \ldots Ag_n\rangle)$ is n.

The usual assumption in the literature is that we can place some measure on the trust that one agent has in another, so we have:

$$tr : Ags \times Ags \mapsto \Re$$

where tr gives a suitable trust value. In this paper, we take this value to be between 0, indicating no trust, and 1, indicating the greatest possible degree of trust. We assume that tr and τ are mutually consistent, so that:

$$tr(Ag_i, Ag_j) \neq 0 \quad \Leftrightarrow \quad (Ag_i, Ag_j) \in \tau$$
$$tr(Ag_i, Ag_j) = 0 \quad \Leftrightarrow \quad (Ag_i, Ag_j) \notin \tau$$

Now, this just deals with the direct trust relations encoded in τ. It is usual in work on trust to consider performing inference about trust by assuming that trust relations are transitive. This is easily captured in the notion of a trust network:

DEFINITION 2. *If, in the trust network $\mathcal{T}$, Ag_i is connected to Ag_j by a directed path $\langle Ag_i, Ag_{i+1}, \ldots Ag_j\rangle$ then Ag_i* trusts *Ag_j* according to $\mathcal{T}$

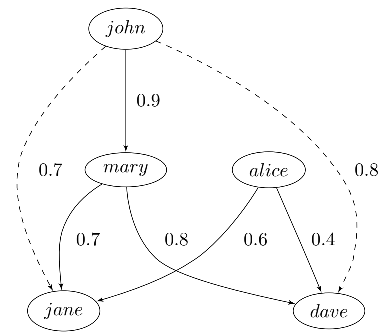

Figure 1: Example trust graph

The notion of trust embodied here is exactly Jøsang's "indirect trust" or "derived trust" [14] and the process of inference is what [12] calls "direct propagation". If we have a function tr, then we can compute:

$$tr(Ag_i, Ag_j) = tr(Ag_i, Ag_{i+1}) \otimes^{tr} tr(Ag_{i+1}, Ag_{i+2}) \otimes^{tr} \ldots \otimes^{tr} tr(Ag_{j-1}, Ag_j) \quad (1)$$

for some operation $\otimes^{tr}$. Here we follow [27] in using the symbol $\otimes$, to stand for this generic operation[1]. Sometimes it is the case that there are two or more paths through the trust network between Ag_i and Ag_j indicating that Ag_i has several opinions about the trustworthiness of Ag_j. If these two paths are

$$\langle Ag_i, Ag'_{i+1}, \ldots Ag_j\rangle \quad \text{and} \quad \langle Ag_i, Ag''_{i+1}, \ldots Ag_j\rangle$$

then the overall degree of trust that Ag_i has in Ag_j is:

$$tr(Ag_i, Ag_j) = tr(Ag_i, Ag_j)' \oplus^{tr} tr(Ag_i, Ag_j)'' \quad (2)$$

Again we use the standard notation $\oplus$ for a function that combines trust measures along two paths [27]. Clearly we can extend this to handle the combination of more than two paths.

Now, given this kind of propagation, we can define an order over the set of agents based on trust values. Since the trust measure we are using is relative to one agent, Ag_i, the order is necessarily relative that agent also. We have:

DEFINITION 3. *For an agent Ag_i, a trust network $\mathcal{T}$and a trust measure tr, we can define an order over agents $\succeq_i^{tr}$ such that $Ag_j \succeq_i^{tr} Ag_k$ iff $tr(Ag_i, Ag_j) \geq tr(Ag_i, Ag_k)$. If this is the case, we say that Ag_i* considers *Ag_j* at least as trustworthy as *Ag_k.*

We further define $=_i^{tr}$ and $\succ_1^{tr}$ in the usual way. $Ag_j =_i^{tr} Ag_k$ iff $Ag_j \succeq_i^{tr} Ag_k$ and $Ag_k \succeq_i^{tr} Ag_j$. $Ag_j \succ_i^{tr} Ag_k$ iff $Ag_j \succeq_i^{tr} Ag_k$ and $Ag_k \not\succeq_i^{tr} Ag_j$. In addition we extend all these relations to operate over a set of agents: $Ags \succeq_i^{tr} Ags'$ iff Ag_i considers every $Ag \in Ags$ at least as trustworthy as every $Ag' \in Ags'$.

As an example of a trust graph, consider Figure 1 (a) which shows the trust relationship between John, Mary, Alice, Jane and Dave. This is adapted from the example in [16] normalizing the values to lie between 0 and 1. The solid lines are direct trust relationships, the dotted lines are indirect links derived from the direct links. Thus John trusts Jane and Dave because he trusts Mary and Mary trusts Jane and Dave. However, John does not, even indirectly, trust Alice.

[1] [12, 16, 23, 27], among others, provide different possible instantiations of this operation some of which we investigate below.

2.2 Argumentation

From the many formal argumentation systems in the literature, we take as our starting point the system from [21]. An agent $Ag_i \in Ags$ maintains a knowledge base, Σ_i, containing a possibly inconsistent set of formulae of a propositional language $\mathcal{L}$. Agent i also maintains the set of its past utterances, called the "commitment store", CS_i. We refer to this as an agent's "public knowledge", since it contains information that is shared with other agents. In contrast, the contents of Σ_i are "private" to Ag_i.

Note that in the description that follows, we assume that $\vdash$ is the classical inference relation, that $\equiv$ stands for logical equivalence, and we use Δ to denote all the information available to an agent. Thus in an interaction between two agents Ag_i and Ag_j, $\Delta_i = \Sigma_i \cup CS_i \cup CS_j$, so the commitment store CS_i can be loosely thought of as a subset of Δ_i consisting of the assertions that have been made public by Ag_i. In some dialogue games, such as those in [21] anything in CS_i is either in Σ_i or can be derived from it. In other dialogue games, such as those in [4], CS_i may contain things that cannot be derived from Σ_i.

DEFINITION 4. *An* argument A *is a pair* (S, p) *where* p *is a formula of* $\mathcal{L}$ *and* S *a subset of* Δ *such that: (i)* S *is consistent; (ii)* $S \vdash p$*; and (iii)* S *is minimal, so no proper subset of* S *satisfying both (i) and (ii) exists.*

S is called the support *of A, written S = Support(A) and p is the* conclusion *of A, written p = Conclusion(A). Thus we talk of p being* supported *by the argument* (S, p).

In general, since Δ may be inconsistent, arguments in $\mathcal{A}(\Delta)$, the set of all arguments which can be made from Δ, may conflict, and we make this idea precise with the notion of *undercutting*:

DEFINITION 5. *Let* A_1 *and* A_2 *be arguments in* $\mathcal{A}(\Delta)$. A_1 undercuts A_2 *iff there is some* $\neg p \in Support(A_2)$ *such that* $p \equiv Conclusion(A_1)$.

In other words, an argument is undercut if and only if there is another argument which has as its conclusion the negation of an element of the support for the first argument.

It will be typical for an agent Ag_i to have different degrees of belief $bel_i(\cdot)$ for the formulae in Δ_i, and in this paper we will assume that these belief values (like those in the much of the uncertainty handling literature) are between 0 and 1. Then, if there is some argument $A = (S, p)$ and $A \in \mathcal{A}(\Delta_i)$ we can compute the belief in an argument from the belief in the formulae in the support of the argument:

$$bel_i(A) = bel_i(s_1) \otimes^{bel} bel(s_2) \otimes^{bel} \ldots \otimes^{bel} bel(s_n) \qquad (3)$$

where $S = \{s_1, \ldots, s_n\}$. Where we need to establish the belief in the conclusion p of A we will set $bel_i(p)$ to be $bel_i(A)$. From these values we can then establish an order over arguments.

DEFINITION 6. *For an agent* Ag_i *and a set of belief values for arguments* $bel_i(\cdot)$*, we can define an order over arguments* $\succeq_i^{bel}$ *such that* $A_1 \succeq_i^{bel} A_2$ *iff* $bel_i(A_1) \geq bel_i(A_2)$. *If this is the case, we say that* Ag_i believes A_1 at least as much as A_2.

In addition we say that $A_1 =_i^{bel} A_2$ iff $A_1 \succeq_i^{bel} A_2$ and $A_2 \succeq_i^{bel} A_1$ and $A_1 \succ_i^{bel} A_2$ iff $A_1 \succeq_i^{bel} A_2$ and $A_2 \not\succeq_i^{bel} A_1$. As with the notion of belief on which they are grounded, we will use these relations between the conclusions of arguments when they hold for the arguments themselves.

We can now define the argumentation system we will use:

DEFINITION 7. *An* argumentation system *is a triple:*

$$\langle \mathcal{A}(\Delta_i), Undercut, \succ_i^{arg} \rangle$$

where $\mathcal{A}(\Delta)$ *is as defined as above,* $\succ_i^{arg}$ *is a preference order over arguments, and* $Undercut$ *is a binary relation collecting all pairs of arguments* A_1 *and* A_2 *such that* A_1 *undercuts* A_2.

Note that for now we don't define exactly where $\succ_i^{arg}$ comes from — later we discuss how it can be established from $\succ_i^{bel}$. We say that A_1 is *stronger than* A_2 iff $A_1 \succ_i^{arg} A_2$.

The preference order makes it possible to distinguish different types of relations between arguments:

DEFINITION 8. *Let* A_1*,* A_2 *be two arguments of* $\mathcal{A}(\Delta)$.

- *If* A_2 *undercuts* A_1 *then* A_1 defends itself *against* A_2 *iff* $A_1 \succ_i^{arg} A_2$. *Otherwise,* A_1 does not defend itself.
- *A set of arguments* $\mathcal{A}$ defends A_1 *iff for every* A_2 *that undercuts* A_1*, where* A_1 *does not defend itself against* A_2*, then there is some* $A_3 \in \mathcal{A}$ *such that* A_3 *undercuts* A_2 *and* A_2 *does not defend itself against* A_3.

If A_1 is undercut by A_2 and either does not defend itself, or is not defended by another set of arguments, we say that A_1 is *successfully undercut* and A_2 is a successful undercutter. We write $\mathcal{A}_{Undercut, \succ_i^{arg}}$ to denote the set of all arguments that are not successfully undercut (which includes those that are not undercut at all). The set $\underline{\mathcal{A}}(\Delta)$ of acceptable arguments of the argumentation system $\langle \mathcal{A}(\Delta), Undercut, \succ_i^{arg} \rangle$ is [3] the least fixpoint of a function $\mathcal{F}$:

$$\begin{aligned} \mathcal{A} &\subseteq \mathcal{A}(\Delta) \\ \mathcal{F}(\mathcal{A}) &= \{(S, p) \in \mathcal{A}(\Delta) \mid (S, p) \text{ is defended by } \mathcal{A}\} \end{aligned}$$

DEFINITION 9. *The set of* acceptable *arguments for an argumentation system* $\langle \mathcal{A}(\Delta), Undercut, \succ_i^{arg} \rangle$ *is recursively defined as:*

$$\begin{aligned} \underline{\mathcal{A}}(\Delta) &= \bigcup \mathcal{F}_{i \geq 0}(\emptyset) \\ &= A_{Undercut, \succ_i^{arg}} \cup \left[\bigcup \mathcal{F}_{i \geq 1}(A_{Undercut, \succ_i^{arg}}) \right] \end{aligned}$$

An argument is acceptable *if it is a member of the acceptable set, and a formula is* acceptable *if it is the conclusion of an acceptable argument.*

An acceptable argument is one which is, in some sense, proven since all the arguments which might undermine it are themselves undermined. If there is an acceptable argument for a formula p, then the *status* of p is *accepted*, while if there is not an acceptable argument for p, the status of p is *not accepted*.

3. ARGUMENTATION AND TRUST

In this paper we are concerned with the following question. If an agent makes use of information that it gets from an acquaintance, how should the degree of trust the agent has in its acquaintance inform the way it uses the information? In particular, if an agent constructs arguments using this information, what, in general terms, is it reasonable for the agent to conclude? For example, we might want to specify that if an agent is given information that it doesn't trust very highly, then it should not allow conclusions derived from this information to over-rule conclusions derived from information provided by more trustworthy sources. However it is not immediately clear how to capture principles like this in formal models we introduced above.

3.1 Combining trust and argumentation

To use our models of trust and argumentation to analyze this question, we first need to consider how to combine them. We opt for a very simple approach, adding a trust network to our existing definition of an argumentation system, so that a *trust argumentation system* is:

$$\langle Ags, \mathcal{A}(\Delta_i), \mathit{Undercut}, \succ_i^{arg}, \mathcal{T} \rangle$$

A trust argumentation system, then is specific to a given agent, Ag_i in the system above, and explicitly includes a set of agents Ags that corresponds to the trust network $\mathcal{T}$, and which are the agents whose commitment stores are combined with Σ_i to make up Δ_i.

The argumentation system from the previous section allows Ag_i to construct arguments from:

$$\Delta_i = \Sigma_i \cup \{ \bigcup_{j=1 \ldots n} CS_j \}$$

and now, thanks to the trust network, Ag_i can assign a trust value to each of the other agents[2] and hence to their commitment store. In addition, the argumentation model assumes that every formulae in Δ_i can be assigned a belief value, and that there is a preference order $\succ_i^{arg}$ over arguments that identifies the relative strength of arguments.

This model, as introduced, is deliberately vague about a number of issues, allowing us to define a whole family of trust argumentation systems, each of which includes a particular instantiation of the elements we have not specified. First, we need to know what functions to use for $\otimes^{tr}$and $\oplus^{tr}$in order to propagate trust values through the trust network in (1) and (2). Second we need to know how to use the trust value $tr(Ag_i, Ag_j)$ that Ag_i puts on Ag_j to determine the belief that i places in information from CS_j. We can express that as a function $ttb(\cdot)$ such that for some $p \in CS_j$

$$bel_i(p) = ttb(tr(Ag_i, Ag_j)) \quad (4)$$

Third, we need to specify how the belief values $bel_i(\cdot)$ are combined using (3) to establish the belief in an argument from the belief in individual formulae and hence the order $\succ_i^{bel}$. Fourth, we need to know how the preference order $\succ_i^{arg}$, which is used to determine acceptability, is established from $\succ_i^{bel}$.

The main aim of this paper is to explore some of these instantiations — different instantiations will give us different behaviors, and we will use the behaviors to evaluate the instantiations. Before we select instantiations we identify a number of desiderata which we want the instantiated trust argumentation system to adhere to.

3.2 Desirable properties

The properties we use are extracted from the literature, and our aim is to identify which make sense when used in combination with argumentation. Golbeck *et al.* [10] suggests that trust should follow the standard rules on network capacity, so that along any given path the maximum amount of trust between a source and a sink will be no larger than the smallest capacity along the path. In terms of propagating trust through a trust graph, this can be interpreted as saying that the trust that some agent Ag_i has in Ag_j is no greater than the minimum trust value along the path between them:

PROPERTY 1. *If Ag_i is connected to Ag_{i+n} by a directed path $\langle Ag_i, Ag_{i+1}, \ldots, Ag_{i+n} \rangle$ in a trust network where arcs are labelled with values $tr(\cdot)$, then:*

$$tr(Ag_i, Ag_{i+n}) \leq min_{j=0,\ldots n-1} tr(Ag_{i+j}, Ag_{i+j+1})$$

[2] If there is no directed path between the two agents, then the value is 0.

[10] also suggest that the length of the path between two agents is relevant in assessing the trust between the agents, and [13] suggests that "the weakening of trust through long transitive paths should result in a reduced confidence level". We will consider two different ways to interpret this. One says that a longer path will never lead to a stronger trust relation than a shorter path:

PROPERTY 2. *If Ag_i is connected to Ag_j and Ag_k by two directed paths in a trust network, then $tr(Ag_i, Ag_j) \leq tr(Ag_i, Ag_k)$ iff $length(Ag_i, Ag_j) \geq length(Ag_i, Ag_k)$.*

The other interpretation says that trust values are monotonically non-increasing over paths:

PROPERTY 3. *Given the directed path $\langle Ag_i, \ldots Ag_j, \ldots Ag_k \rangle$ then $tr(Ag_i, Ag_k) \leq (Ag_i, Ag_j)$*

The above properties relate to $\otimes^{tr}$. There are also properties relating to $\oplus^{tr}$. The first comes from [13] which suggests that "combination of parallel trust paths should result in an increased confidence level". In other words:

PROPERTY 4. *If Ag_i and Ag_j are linked by two paths in the trust network $\mathcal{T}$, and the trust computed along these paths are $tr(Ag_i, Ag_j)'$ and $tr(Ag_i, Ag_j)''$, then the overall trust of Ag_i in Ag_j,*

$$tr(Ag_i, Ag_j) \geq max \left(tr(Ag_i, Ag_j)', tr(Ag_i, Ag_j)'' \right)$$

The authors like to think of this as encoding the idea that having two letters of recommendation for a potential PhD student that say the student is excellent is no worse than having one. However, there is another desideratum that we might enforce here. If we have a potential PhD student with a multitude of recommendation letters that suggest they are a mediocre student, does this make them more highly recommended than a student with just a couple of letters suggesting that they are very good? The authors feel not, and so we also consider the property that combining two parallel trust paths does not cause the overall trust value to exceed the value defined by either path (which is one way to stop the many poor recommendations outweighing a few good ones for a different student).

PROPERTY 5. *If Ag_i and Ag_j are linked by two paths in the trust network $\mathcal{T}$, and the trust computed along these paths are $tr(Ag_i, Ag_j)'$ and $tr(Ag_i, Ag_j)''$, then the overall trust of Ag_i in Ag_j,*

$$tr(Ag_i, Ag_j) \leq max \left(tr(Ag_i, Ag_j)', tr(Ag_i, Ag_j)'' \right)$$

In different situations, either of these properties may be appropriate.

We can extend several of these ideas to deal with beliefs and their role in argumentation, in essence placing constraints on the the operation $\otimes^{bel}$. Thinking of an argument as a chain of inferences that make use of formulae from Δ_i then an extension of Property 1 is that the conclusion of an argument should be believed no more than the minimum of the degrees of belief of all of the steps in the argument. This gives us:

PROPERTY 6. *If Ag_i has an argument (S, p), and the support $S = \{s_i, \ldots, s_m\}$, then:*

$$bel_i(p) \leq min_{j=1,\ldots m} bel_i(s_j)$$

We can also extend Properties 2 and 3 to argumentation. This extension suggests that an argument that requires a larger support (and so in some sense is "longer") than another is less believable, and there are two obvious ways that we might capture this:

PROPERTY 7. *If Ag_i has two arguments (S, p) and (S', p'), then $bel_i(p) \leq bel_i(p')$ iff $|S| \geq |S'|$.*

which is analogous to P2 in saying that larger support never means a greater degree of belief, and:

PROPERTY 8. *If Ag_i has two arguments (S, p) and (S', p'), then $bel_i(p) \leq bel_i(p')$ if $S \supseteq S'$.*

which is analogous to P3 in saying that adding additional formulae to a support cannot increase belief and is essentially Loui's [17] "directness" defeater.

The final property that we will consider here deals with the behavior of the combined trust and argumentation system, capturing one reading of the principle we outlined at the start of this section — the strength of an agent's arguments should reflect the trustworthiness of the agents from whom the support of those arguments was obtained. To capture this idea we need first to define:

DEFINITION 10. *Given a set of agents $Ags = \{Ag_1, \ldots Ag_n\}$ where each Ag_j has a commitment store CS_j, then a set of formulae S* corresponds *to the set of agents Ags' iff*

$$Ags' = \{Ag_j | s \in S \text{ and } s \in CS_j\}$$

so that a set of formulae corresponds to the set of agents from whose commitment stores the formulae are drawn. Then we have:

PROPERTY 9. *If Ag_i has two arguments (S, p) and (S', p'), where the supports have corresponding sets of agents Ag and Ag' then (S, p) is stronger than (S', p') only if Ag_i considers Ag to be more trustworthy than Ag'.*

If this property is obeyed, then arguments grounded in information from less trustworthy sources will not be able to defeat arguments whose grounds are drawn from more trustworthy sources. In turn this means that:

PROPOSITION 1. *In a trust argumentation system:*

$$\langle Ags, \mathcal{A}(\Delta_i), Undercut, \succ_i^{arg}, \mathcal{T} \rangle$$

If an argument (S, p), with corresponding set of agents Ag, is acceptable, then, given Property 9, a new argument (S', p') with corresponding set of agents Ag' if Ag_i cannot make (S, p) not acceptable if Ag_i considers Ag' to be less trustworthy than Ag.

PROOF. *If (S, p) is acceptable, then it is not successfully undercut, and so either (i) it is stronger than all its attackers, or (ii) it is defended by arguments that are stronger than those attackers that are stronger than it. Now consider that Ag_i learns enough information to create (S', p') which undercuts (S, p). To make (S, p) not acceptable (S', p') either has to successfully undercut (S, p) or one of (S, p)'s defenders. However, by Property 9, since (S', p')'s corresponding set of agents is less trustworthy than those of (S, p) it is not stronger than (S, p) and so cannot successfully undercut it. Furthermore, since the defenders in (ii) are also stronger than (S, p), (S', p') cannot undercut them either, and so it will fail to make (S, p) not acceptable.* $\square$

This result shows the importance of Property 9 — when it holds, it prevents arguments based on less trustworthy agents from making otherwise acceptable arguments unacceptable, and thus altering what Ag_i takes as being proven.

Note that the desiderata are not independent:

PROPOSITION 2. *Property 2 implies Property 3 and Property 7 implies Property 8.*

PROOF. *P2 requires that given paths from Ag_i to Ag_j and Ag_k, then $tr(Ag_i, Ag_j) \leq tr(Ag_i, Ag_k)$ if and only if $length(Ag_i, Ag_j)$ is greater than or equal to $length(Ag_i, Ag_k)$. If this is the case, then given a path $\langle Ag_i, \ldots Ag_j, \ldots Ag_k \rangle$ it is clear that the path from Ag_i to Ag_k is longer than the path to Ag_j and so $tr(Ag_i, Ag_k)$ will be less than or equal to $tr(Ag_i, Ag_j)$, fulfilling P3.*

Similarly, P7 requires that if Ag_i has two arguments (S, p) and (S', p'), then $bel_i(p) \leq bel_i(p')$ iff $|S| \geq |S'|$. If $S \supseteq S'$ then this will imply that $|S| \geq |S'|$ and hence $bel_i(p) \leq bel_i(p')$, fulfilling P8. $\square$

However these pairs of properties are distinct:

PROPOSITION 3. *Property 3 does not imply Property 2 and Property 8 does not imply Property 7.*

PROOF. *To prove that the first of each of these properties does not imply the second, it suffices to show a single instance where it is not the case. For P3 and P2 we do this by choosing a specific operator for $\otimes^{tr}$. If we use min, then P3 will hold for any assignment of trust values along the path $\langle Ag_i, \ldots Ag_j, \ldots Ag_k \rangle$, for example one with minimum value 0.5. However, with the same operator, we can construct a much longer path where the minimum trust value is 0.8, violating Property 2.*

The counter-example for the second pair of properties is analogous — combining beliefs with min means a small set of support can easily have a smaller belief value than a large set. $\square$

4. TRUST ARGUMENTATION

Having identified a system of trust argumentation and some desiderata for it, in this section we explore its properties.

4.1 Properties of the system

We start by identifying which possible instantiations of the combined trust and argumentation model will satisfy the desiderata in the sense of guaranteeing that the properties will always hold. We begin with Properties 1–3 which depend upon the choice of $\otimes^{tr}$. Two such choices, suggested by Richardson *et al.*[23] are minimum and multiplication. We have:

PROPOSITION 4. *Combining trust values along a path in a trust network according to (1) with minimum or multiplication will satisfy Properties 1 and 3 but not Property 2.*

PROOF. *With associative operations like minimum and multiplication, combining trust values along a path in a trust network is exactly the same as combining a set of trust values. If we combine a set of trust values with minimum, then clearly the resulting value will be exactly the minimum of the values and satisfy Property 1. If we combine two sets of values S_1 and S_2 using minimum, and $S_1 \subseteq S_2$, then the minimum of S_1 will be no smaller than the minimum of S_2, and Property 3 holds. It is equally easy to prove Property 2 does not always hold. If we have two sets S_1 and S_2 and $S_1 \cap S_2 = \emptyset$, then even if S_2 is much larger than S_1, its minimum value can be larger than than that of S_1 — all the values in S_2 could be 0.8 and all those in S_1 could be 0.3.*

Combining a set of values that are no larger than 1 with multiplication will give a value that no larger than any of them, satisfying Property 1. Similarly, if we take the result of multiplying the values in S_1 and then multiply by the values in $S_2 - S_1$ for $S_1 \subseteq S_2$, the value we have won't increase, satisfying Property 3. However, with two unconnected sets S_1 and S_2 there is no necessary relationship between the product of the values in the sets and so Property 2 will not always hold. $\square$

The issue with satisfying Property 2 is that both minimum and multiplication are applied link by link so there is no way to they can meet a criterion that applies to the whole path. If we stretch the definition of computing trust values along a path to allow trust values to be combined by functions that take the whole path as arguments, then we can easily show that:

PROPOSITION 5. *Combining trust values along a path in a trust network in such a way that the trust value is inversely proportional to the length of the path will satisfy Properties 2 and 3 but not Property 1:*

PROOF. *Property 2 requires $tr(Ag_i, Ag_j) \leq tr(Ag_i, Ag_k)$ iff $length(Ag_i, Ag_j) \geq length(Ag_i, Ag_k)$ which is obviously true for this combination. By Proposition 2, Property 2 implies Property 3, so Property 3 holds as well. The last part of the result is just as easy to show — since the combination depends only on the length of the path, not on the trust values labelling the arcs, there is no reason why the trust along a path should have any particular relationship with those values.* □

The problem with this approach to propagation, and the problem with Property 2, is that it ignores the values of the individual links. As a result it is easy to construct examples which conflict with intuition — a path with very high valued links creates less trust than a marginally shorter path with very low valued links, and any attempt to bring in the values of the links creates situations in which Property 2 can easily be violated.

Now we consider options for $\oplus^{tr}$. Richardson *et al.* [23] suggest maximum and Golbeck *et al.*[10] suggest average[3], while addition seems a suitable dual operation to consider for the options we considered for $\otimes^{tr}$— addition is the dual operation to multiplication for probability theory, and some variants of possibility theory use it as a dual for minimum [9]. Considering all three of these operations, we have:

PROPOSITION 6. *Combining trust values over multiple paths in a trust network according to (2) with maximum satisfies Properties 4 and 5, combining using addition satisfies Property 4 but does not satisfy Property 5, and combining using average satisfies Property 5 but does not satisfy Property 4.*

PROOF. *Since Property 4 specifies that the combination must be greater than or equal to the maximum of the values and Property 5 specifies that it must be less than or equal to the maximum, maximum satisfies both (and will be the only operation to). Adding the two values will clearly give something no smaller than the larger, satisfying Property 4 but won't in general satisfy Property 5 (it will only satisfy it when one value is 0). Average will give something no larger than the larger value, satisfying Property 5, but will only satisfy Property 4 when the values are the same.* □

So addition meets our formulation of Jøsang's property, average obeys the property that we introduced, and maximum meets both.

The third set of properties are those for combining beliefs with $\otimes^{bel}$. In our combined trust and argumentation system, we are assuming that the belief values of propositions in Δ_i are affected by trust values (and we discuss some ways in which this could be achieved below) but to consider the properties, all we assume for now is that there is some distribution of values:

$$m_i : \Delta_i \mapsto [0, 1]$$

[3] Average is not usually considered as a binary operation, but it can be expressed in such a form, see, for example [25].

from which we can establish a belief value $bel_i(\cdot)$, between 1 and 0, for any formula in Δ_i[4]. These values are then combined to establish beliefs in the conclusions of arguments. Here we consider multiplication and minimum as possible operations for this combination, following the conjunction operations in probability theory and possibility theory respectively [9]. Given Proposition 4 and the origin of Property 1 it is no surprise to find that:

PROPOSITION 7. *Combining belief values according to (3) with minimum or multiplication will satisfy Properties 6 and 8 but not Property 7.*

PROOF. *The proof is the same as for Proposition 4.* □

In order to satisfy Property 7 we need to combine beliefs in a way that depends on the size of the set of support, for example:

PROPOSITION 8. *Consider an argument $A = (S, p)$ where $S = \{s_1, \ldots, s_n\}$. Setting $bel(p) = \frac{1}{|S|}$ will satisfy Properties 7 and 8 but not Property 6.*

PROOF. *The proof is close to that for Proposition 5. The definition of the belief computation means it clearly satisfies Property 7 and by Proposition 2, Property 8 holds as well. The last part of the result is just as easy to show — since the belief in an argument depends only on the size of the support, not on the belief values of formulae in the support, there is no reason why the overall belief should have any particular relationship with the beliefs of the formulae.* □

Thus we have ways of handling trust and belief which will satisfy the various properties we identified, but we have no set of operations that will simultaneously satisfy all the properties.

The final desiderata that we laid down is Property 9, which relates trust values to the conclusions of arguments. To reason about the conditions under which this will hold, we first need to decide how to convert the trust that an agent Ag_i has in agent Ag_j into the belief that Ag_i has in formulae from CS_j. In order to obtain priorities over an agent's knowledge — which is the role played by beliefs in our argumentation — [16] simply imports trust values as the priorities, and here we propose the same method, defining the function *ttb* from (4) as:

$$ttb(tr(Ag_i, Ag_j)) = tr(Ag_i, Ag_j) \cdot bel\_limit_i$$

where bel_limit_i is a scaling factor that, given belief and trust values are between 0 and 1 limits the maximum belief that a trust value can map to. There are two obvious ways to set this:

L1 $bel\_limit_i = 1$

L2 $bel\_limit_i = min_j\{bel_i(s_j) | s_j \in \Sigma_i\}$

so that we either scale the trust values compared to the maximum possible value for beliefs, so that information with a trust value of 1 is considered as believable as anything, or we scale beliefs so that everything in Σ_i is at least as believable as anything Ag_i is told by another agent.

We also need to determine how $\succ_i^{arg}$ depends on $\succ_i^{bel}$, and there are two obvious ways to do this:

O1 $(S, p) \succ_i^{arg} (S', p')$ iff $(S, p) \succ_i^{bel} (S', p')$

O2 $(S, p) \succ_i^{arg} (S', p')$ iff $(S, p) \succ_i^{bel} (S', p')$ and $Ag \succ_i^{tr} Ag'$ for all Ag corresponding to S and Ag' corresponding to S'.

[4] The reason for describing the allocation of belief values in this indirect way is that it is required by some approaches to handling uncertainty, including possibility theory [9] which we will make use of below.

With these aspects of the model instantiated, we can consider which combinations of the various features of the model satisfy Property 9. We have:

PROPOSITION 9. *A trust argumentation system that uses minimum for $\otimes^{tr}$, maximum for $\oplus^{tr}$, minimum for $\otimes^{bel}$ and adopts L2 and O1 satisfies Property 9.*

PROOF. *Property 9 requires the strength of an argument to be determined by the trust Ag_i has in the corresponding agents so that arguments with less trustworthy corresponding agents are weaker. L2 means that no formulae from any CS_j can be believed more than one from Σ_i, and using minimum to combine belief values means that the strength of any argument will be determined by the trustworthiness of the corresponding agents (a low belief from Σ_i cannot hide an argument's dependency on an untrustworthy agent).* □

Examining the proof, it is clear why we need to have bel_limit_i in the model — without it, there is nothing to stop a highly trusted source supplying information that ends up supporting a weak argument by virtue of another piece of the support which comes from Ag_i itself having a low degree of belief. This, in turn might lead to an argument supported by information from a less trusted source being stronger than an argument based on information from a more trusted source. Exactly this line of reasoning leads us to:

PROPOSITION 10. *A trust argumentation system that that uses minimum for $\otimes^{tr}$, maximum for $\oplus^{tr}$, minimum for $\otimes^{bel}$ and adopts L1 and O1 does not satisfy Property 9 unless $bel(s) = 1$ for every $s \in \Sigma_i$.*

PROOF. *Immediate from the proof of Proposition 9.* □

so not adopting L2[5] doesn't prevent a trust argumentation system meeting our benchmark of performance, Property 9, but means it can only do so under rather restricted circumstances.

Proposition 9 and Proposition 1 tell us that using possibility-style maximum and minimum operations for trust and argumentation — an instantiation of our trust-argumentation system that we will call TA_1 — can guarantee what we have argued is desirable behavior. What about using multiplication, which as we have remarked above, fits more naturally with a probabilistic interpretation of belief? It turns out that:

PROPOSITION 11. *A trust argumentation system that uses minimum for $\otimes^{tr}$, maximum for $\oplus^{tr}$, multiplication for $\otimes^{bel}$ and adopts L2 and O1 does not satisfy Property 9*

PROOF. *Since the result is only that the system does not satisfy the property, a counter example will suffice. Consider all propositions in Σ_i have belief 1. (S, p) includes just one formula that isn't from Σ_i, it comes from CS_j, and $tr(Ag_i, Ag_j) = 0.7$. $bel_i(S, p)$ is thus 0.7. (S', p') includes just two formulae that aren't from Σ_i. These formulae come from CS_k and CS_l, and $tr(Ag_i, Ag_k) = tr(Ag_i, Ag_l) = 0.8$. Thus $bel_i(p') = 0.64$ and the argument is not as strong as the argument which depends on information from a less-trusted source.* □

As the proof shows, the reason that this second trust argumentation system fails to satisfy Property 9 is because multiplying belief values will generate arguments with low beliefs and with O1 determining the order over arguments, this means weak arguments can be generated using information from highly trusted agents. One way to prevent this is to use O2 to determine the order over arguments. We have:

[5] Or, of course, some other mechanism for preventing the kind of interaction between belief and trust sketched in the proof of Proposition 9.

PROPOSITION 12. *A trust argumentation system that uses minimum for $\otimes^{tr}$, maximum for $\oplus^{tr}$, multiplication for $\otimes^{bel}$ and adopts L2 and O2 satisfies Property 9.*

PROOF. *Immediate from the definition of O2.* □

The disadvantage of adopting O2 is that it will only produce a partial order for $\succ_i^{arg}$, and given the role $\succ_i^{arg}$ plays in defining the acceptability, this will affect the reasoning the agents can carry out.

4.2 Trust thresholds

Let's look at one way we can use TA_1. Consider that Ag_i has a *trust threshold* of α, a trust value for agents below which it wishes not to use information from them. If we give arguments whose status is unaffected by information from agents whose trust value is below the threshold α the name *α-safe* then:

PROPOSITION 13. *If Ag_i has a TA_1 argumentation system:*

$$\langle Ags, \mathcal{A}(\Delta_i), Undercut, \succ_i^{arg}, \mathcal{T}\rangle$$

where all formulae in Σ_i have belief value 1, and Ag_i has a trust threshold α, then all arguments with a level of belief above α are α-safe.

PROOF. *Setting the belief of all formulae in Σ_i to 1 ensures that the belief values of arguments directly reflect their trust values making the belief value equal to the threshold easy to establish[6]. If an argument A is acceptable, and has a belief value above α, then — as we recall from the proof of Proposition 1 — any undercutters that aren't weaker than A (and so may be below the trust threshold but not affecting the status of A) must, since A is acceptable, be successfully undercut by stronger arguments. Because of the way that trust is converted into belief and belief values are combined with minimum, none of these arguments can be based on information that comes from an agent trusted less than α. So not only A, but all of the arguments that determine its status, must be α-safe.*

If an argument A' is not acceptable and it is above the trust threshold, but was successfully defeated, then that defeat must have been by an argument that is above the trust threshold which (since that defeater is successful) means that in the same way as A, this defeater is α-safe, and hence so is A'. □

This result is helpful because it shows us that for TA_1 information from agents below the trust threshold has limited impact — it won't change the acceptability or otherwise of arguments above the threshold.

5. CONCLUSION

In this paper we presented a formal model that provides a simple combination of argumentation and trust. We examined some of the properties of different instantiations of the model, and showed that the system we called TA_1 has the ability to ensure that arguments grounded in information from untrustworthy agents cannot overrule arguments grounded by more trustworthy agents and under certain conditions can deal with trust thresholds.

This work is distinct from, and complementary to, other existing work on trust and argumentation. The work of Matt *et al.* [18] for example looks at constructing arguments *for* trusting other agents — it is a way to compute the tr values that we assume. In contrast, here we are concerned with computing arguments *with* trust. Similar remarks hold for [20] which looks to construct arguments about the trust that one agent has in another.

[6] The proof can be altered to deal with formulae in Σ_i having smaller belief values, it would mean replacing the trust threshold in the proof with $\alpha.min_j\{bel_i(s_j)|s_j \in \Sigma_i\}$

Though the system we define is simple, there is more to say about it. Our future work will address aspects of the system that we have not had space to discuss here. We are working on a more extensive analysis of operators for the trust argumentation systems, as well as expanding the notion of trust threshold to what we call the *trust budget* — if an agent is prepared to tolerate a certain overall amount of distrust in all the information it uses in all of its arguments, how does this affect what it finds acceptable? Other topics of interest are combining what we have here with the use of argumentation to establish trust values, and the use of more complex methods of representing trust than the simple numerical approach we adopt here.

Acknowledgement

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

We thank the reviewers for their helpful comments.

6. REFERENCES

[1] Z. Abrams, R. McGrew, and S. Plotkin. Keeping peers honest in eigentrust. In *Proceedings of the 2nd Workshop on the Economics of Peer-to-Peer Systems*, 2004.

[2] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International World Wide Web Conference*, pages 261–270, Banff, Alberta, May 2007.

[3] L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation framework. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1998.

[4] L. Amgoud, S. Parsons, and N. Maudet. Arguments, dialogue, and negotiation. In *Proceedings of the Fourteenth European Conference on Artificial Intelligence*, 2000.

[5] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2):58–71, June 2007.

[6] P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128:203–235, 2001.

[7] P. Dandekar, A. Goel, R. Govindan, and I. Post. Liquidity in credit networks: A little trust goes a long way. Technical report, Department of Management Science and Engineering, Stanford University, 2010.

[8] D. B. DeFigueiredo and E. T. Barr. TrustDavis: A non-explotable online reputation system. In *Proceedings of the 7th IEEE International Conference on E-Commerce Technology*, 2005.

[9] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York, NY, 1988.

[10] J. Golbeck, B. Parsia, and J. Hendler. Trust networks on the semantic web. In *Proceedings of the 7th International Workshop on Cooperative Information Agents*, Helsinki, August 2003.

[11] T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Communications Surveys and Tutorials*, 4(4):2–16, 2000.

[12] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on the World Wide Web*, 2004.

[13] A. Jøsang, E. Gray, and M. Kinateder. Simplification and analysis of transitive trust networks. *Web Intelligence and Agent Systems*, 4(2):139–161, 2006.

[14] A. Jøsang, C. Keser, and T. Dimitrakos. Can we manage trust? In *Proceedings of the 3rd International Conference on Trust Management*, Paris, May 2005.

[15] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th World Wide Web Conference*, May 2004.

[16] Y. Katz and J. Golbeck. Social network-based trust in prioritzed default logic. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[17] R. P. Loui. Defeat among arguments: a system of defeasible inference. *Computational Intelligence*, 3(3):100–106, 1987.

[18] P.-A. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagents Systems*, 2010.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.

[20] S. Parsons, P. McBurney, and E. Sklar. Reasoning about trust using argumentation: A position paper. In *Proceedings of the Workshop on Argumentation in Multiagent Systems*, Toronto, Canada, May 2010.

[21] S. Parsons, M. Wooldridge, and L. Amgoud. On the outcomes of formal inter-agent dialogues. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multi-Agent Systems*, 2003.

[22] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye, editor, *The Economics of the Internet and E-Commerce*, pages 127–157. Elsevier Science, Amsterdam, 2002.

[23] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, 2003.

[24] J. Sabater and C. Sierra. Review on computational trustand reputation models. *AI Review*, 23(1):33–60, September 2005.

[25] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[26] W. T. L. Teacy, G. Chalkiadakis, A. Rogers, and N. R. Jennings. Sequential decision making with untrustworthy service providers. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, 2008.

[27] Y. Wang and M. P. Singh. Trust representation and aggregation in a distributed agent system. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

[28] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the Conference on Knowledge and Data Discovery*, 2007.

[29] S. Zhong, J. Chen, and Y. R. Yang. Sprite: A simple cheat-proof, credit-based system for mobile ad-hoc networks. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, 2003.