Using argumentation to reason with and about trust

Simon Parsons[1,2], Elizabeth Sklar[1,2], and Peter McBurney[3]

[1] Department of Computer & Information Science, Brooklyn College,
City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210, USA
{sklar,parsons}@sci.brooklyn.cuny.edu
[2] Department of Computer Science, The Graduate Center
City University of New York, 365 5th Avenue, New York, NY 10016, USA
[3] Department of Informatics, King's College London
Strand, London WC2R 2LS, United Kingdom
peter.mcburney@kcl.ac.uk

Abstract. Trust is an approach to managing the uncertainty about autonomous entities and the information they store, and so can play an important role in any decentralized system. As a result, trust has been widely studied in multiagent systems and related fields such as the semantic web. Here we introduce a simple approach to reasoning about trust with logic, describe how it can be combined with reasoning about beliefs using logic, and demonstrate its use on an example. The example highlights a number of issues related to resolving weighted arguments.

1 Introduction

Trust is an approach to 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 become steadily more important within Computer Science [4, 18].

Thus, for example, we see work on trust in peer-to-peer networks, including the EigenTrust algorithm [22] — a variant of PageRank [34] 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 from manipulating their trust values to get preferential downloads. [52] is concerned with manipulation in mobile ad-hoc networks, and looks to prevent nodes from getting others to transmit their messages while refusing to transmit the messages of others.

The internet, as the largest distributed system of all, is naturally a target of much of the research on trust. There have been studies, for example, on the development of trust in ecommerce [31, 43, 51], on mechanisms to determine which sources to trust when faced with multiple conflicting sources [10, 39, 50], on mechanisms for identifying which individuals to trust based on their past activity [2, 20, 27], and on the manipulation of online recommendation systems [25]. The work we have just cited can be thought of as helping agents to decide who is worthy of trust. A development from a slightly different perspective — that of making it possible to trust individuals who might otherwise be deemed untrustworthy — is the idea of having individuals indemnify each other by placing some form of financial guarantee on transactions that others enter into [8, 9]. Thus I might indemnify you against a third party that I trust, thus making you feel comfortable doing business with them.

Trust is an especially important issue from the perspective of autonomous agents and multiagent systems [48]. 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 amount that 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 [45, 49], including work that identifies weaknesses in some of the major trust models [46].

In the work in this area, 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. One natural question to ask in this context is what inference is reasonable in such networks. The propagation of trust — both the transitivity of trust relations [44, 49] and more complex relationships like "co-citation" [19] — has been studied. In many cases this work has been empirically validated [19, 23, 24].

In a previous paper [37], we suggested that, given the role that provenance plays in trust [16, 17], *argumentation* — which tracks the origin of data used in reasoning — might play a role. We have developed a graph-based model to explore the relationship between argumentation and trust [47]. Here we explore a different direction, discussing how the usual approach to dealing with trust information can be captured in logic, how it can be integrated with argumentation-based reasoning about beliefs, and how it might be used in a combined system.

2 Trust

We are interested in a finite set of agents Ags and how these agents trust one another. Following the usual presentation (for example [23, 44, 49]), 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 define a *trust network* to be 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 $\mathcal{T}$ indicating that Ag_i trusts Ag_j.

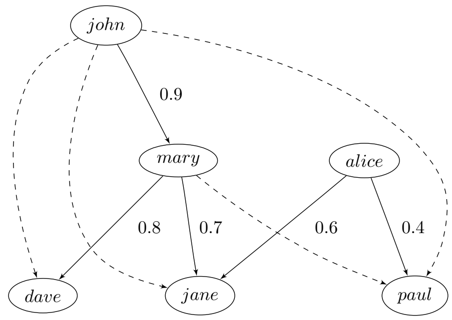

Fig. 1. An example trust graph. The solid lines represent direct trust relations, and the dashed lines represent derived trust. The link between $john$ and $jane$ and the link between $john$ and $dave$ are the result of direct propagation. The link between $mary$ and $paul$ is the result of co-citation (see below).

In this graph, the set of agents is the set of vertices, and the trust relations define the arcs. 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. The usual assumption in the literature is that we can place some measure on the trust relation, quantifying 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:

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

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. The notion of trust embodied here is exactly Jøsang's "indirect trust" or "derived trust" [21] and the process of inference is what [19] 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 [49] in using the symbol $\otimes$, to stand for this generic operation.[1] The superscript distinguishes this from a similar operation $\otimes^{bel}$ on belief values which we will meet below.

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

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

and

$$tr(Ag_i, Ag_j)' = tr(Ag_i, Ag'_{i+1}) \otimes^{tr} \ldots \otimes^{tr} tr(Ag'_{j-1}, Ag_j)$$
$$tr(Ag_i, Ag_j)'' = tr(Ag_i, Ag''_{i+1}) \otimes^{tr} \ldots \otimes^{tr} tr(Ag''_{j-1}, Ag_j)$$

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

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

Again we use the standard notation $\oplus$ for a function that combines trust measures along two paths [49]. Clearly we can extend this to handle the combination of more than two paths.

As an example of a trust graph, consider Figure 1 which shows the trust relationship between $john$, $mary$, $alice$, $jane$, $paul$ and $dave$. This is adapted from the example in [23] by normalizing the values to lie between 0 and 1 and adding $paul$. The solid lines are direct trust relationships and the dotted lines are indirect links derived from the direct links. Thus, for example, $john$ trusts $jane$ and $dave$ because he trusts $mary$ and $mary$ trusts $jane$ and $dave$.

The standard approach in the literature on trust is to base the computation of derived trust values on the the trust graph, for example using a path algebra [44]. Our aim in this paper is to demonstrate how we might use logic, and in particular argumentation, to propagate trust values. In other words we want an argumentation-based approach that $john$ can use to determine that he has a reason to trust $dave$, and then use to combine this trust with his other knowledge to make decisions.

3 Reasoning about trust

We will start by considering how to capture reasoning about trust in logic. We will assume that every agent Ag_i has some collection of information about the world, which we will call Δ_i, that is expressed in logic. Δ_i is made up of a number of partitions, one of which, Δ_i^{tr}, holds information about the degree of trust Ag_i has in other agents it knows. For example, the agent $john$ from the above example might have the following collection of information:

$$\begin{array}{ll}\Delta^{tr}_{john} & (t1 : trusts(john, mary) : 0.9)\\ & (t2 : trusts(mary, jane) : 0.7)\\ & (t3 : trusts(mary, dave) : 0.8)\\ & (t4 : trusts(alice, jane) : 0.6)\\ & (t5 : trusts(alice, paul) : 0.4)\end{array}$$

[1] [19, 23, 44, 49], among others, provide different possible instantiations of this operation.

$$Ax^{tr} \quad \frac{(n : trusts(x,y) : \tilde{d}) \in \Delta_i^{tr}}{\Delta_i^{tr} \vdash_{tr} (trusts(x,y) : \{n\} : \{Ax^{tr}\} : \tilde{d})}$$

$$dp \quad \frac{\Delta_i^{tr} \vdash_{tr} (trusts(x,y) : G : R : \tilde{d}) \text{ and } \Delta_i^{tr} \vdash_{tr} (trusts(y,z) : H : S : \tilde{e})}{\Delta_i^{tr} \vdash_{tr} (trusts(x,z) : G \cup H : R \cup S \cup \{dp\} : \tilde{d} \otimes^{tr} \tilde{e})}$$

$$cc \quad \frac{\Delta_i^{tr} \vdash_{tr} (trusts(x,y) : G : R : \tilde{d}) \text{ and } \Delta_i^{tr} \vdash_{tr} (trusts(x,z) : H : S : \tilde{e}) \text{ and } \Delta_i^{tr} \vdash_{tr} (trusts(w,z) : K : T : \tilde{f})}{\Delta_i^{tr} \vdash_{tr} (trusts(w,y) : G \cup H \cup K : R \cup S \cup T \cup \{cc\} : \tilde{d} \otimes^{tr} \tilde{e} \otimes^{tr} \tilde{f})}$$

Fig. 2. Part of the tr consequence relation

where the elements of Δ^{tr}_{john} are the kind of triples that we have discussed in earlier work [35]. Each element has the form:

$$(\langle index\rangle : \langle data\rangle : \langle value\rangle)$$

The first is a means of referring to the element, the second is a formula, and here the third is the degree of trust between the individuals mentioned in the $trust$ relation.

From Δ^{tr}_{john} we can then construct arguments mirroring the trust propagation discussed above. Rules for doing this are given in Figure 2.[2] For example, using the first two rules, from Figure 2, Ax^{tr} and dp, we can construct the argument:

$$\Delta^{tr}_{john} \vdash_{tr} (trusts(john, jane) : \{t1, t2\} : \{Ax^{tr}, Ax^{tr}, dp\} : \tilde{t})$$

where all arguments in our approach take the form:

$$(\langle conclusion\rangle : \langle grounds\rangle : \langle rules\rangle : \langle value\rangle)$$

The $\langle conclusion\rangle$ is inferred from the $\langle grounds\rangle$ using the rules of inference $\langle rules\rangle$ and with degree $\langle value\rangle$. In this case the argument says $john$ trusts $jane$ with degree $\tilde{t}$ (which is $0.9 \otimes^{tr} 0.7$), through two applications of the rule Ax^{tr} and one application of the rule dp to the two facts indexed by $t1$ and $t2$.[3]

The rule Ax^{tr} says that if some agent Ag_i has a triple:

$$(t1 : trusts(john, mary) : 0.9)$$

in its Δ^{tr}_i then it can construct an argument for $trusts(john, mary)$ where the grounds are $t1$, the degree of trust is 0.9, and which records that the Ax^{tr} rule was used in its derivation.

The rule dp captures direct propagation of trust values. It says that if we can show that $trusts(x, y)$ holds with degree $\tilde{d}$ and we can show that $trusts(y, z)$ holds with degree $\tilde{e}$, then we are allowed to conclude $trusts(x, z)$ with a degree $\tilde{d} \otimes^{tr} \tilde{e}$, and that the conclusion is based on the union of the information that supported the premises, and is computed using all the rules used by both the premises.

Why is this interesting? After all, it does no more than trace paths through the trust graph.

Well, one of the strengths of argumentation, and the reason we are interested in using argumentation to handle trust, is that we want to record, in the form of the argument for some proposition, the *reasons* that it should be believed. Since information on the source of some piece of data, and the trust that an agent has in the source, is relevant, then it should be recorded in the argument. This is easier to achieve if we encode data about who trusts whom in logic.

---

[2] Note that the consequence relation in Figure 2 is not intended to be comprehensive. There are many other ways to construct arguments about trust — for some examples see [36] — which could be included in the definition of $\vdash_{tr}$.

[3] There are good reasons for using the formulae themselves in the grounds and factoring the whole proof into the set of rules (as we do in [37]) to obtain structured arguments like those in [15, 41]. However, for simplicity, here we use the relevant indices.

One of the nice things that this approach allows us to do is to track the application of the rules for propagating trust. When we just use direct propagation, this is not terribly interesting (though it does allow us to distinguish between the bits of information used in the formation of arguments, which may be a criterion for preferring one argument over another [28]), but it becomes more obviously useful when we start to allow other rules for propagating trust. For example, [19] suggests a rule the authors call *co-citation*, which they describe as:

> For example, suppose i_1 trusts j_1 and j_2 and i_2 trusts j_2. Under co-citation, we would conclude that i_2 should also trust j_1.

In our example (see Figure 1), therefore, co-citation suggests that since $alice$ trusts $jane$ and $paul$, and $mary$ trusts $jane$, then $mary$ should trust $paul$. (Presumably the idea is that since $alice$ and $mary$ agree on the trustworthiness of $jane$, $mary$ should trust $alice$'s opinion about $paul$). [19] also tells us how trust values should be combined in this case — $mary$'s trust in $paul$ is just the combination of trust values along the path from $mary$ to $jane$ to $alice$ to $paul$.

This form of reasoning is captured by the rule cc in Figure 2, and the rule also takes care of the necessary bookkeeping of grounds, proof rules and trust values. Combining the application of cc with dp as before allows the construction of the argument:

$$\Delta^{tr}_{john} \vdash_{tr} (trusts(john, paul) : \{t1, t2, t4, t5\} : rules_1 : \tilde{r})$$

indicating that $john$ trusts $paul$, where $rules_1$ is:

$$\{Ax^{tr}, Ax^{tr}, Ax^{tr}, Ax^{tr}, cc, dp\}$$

and $\tilde{r}$ is $0.9 \otimes^{tr} 0.7 \otimes^{tr} 0.6 \otimes^{tr} 0.4$.

Now, when we have several rules for propagating trust, keeping track of which rule has been used in which derivation is appealing, especially since one might want to distinguish between arguments that use different rules of inference. For example, one might prefer arguments, no matter the trust value, which only make use of direct propagation over those that make use of co-citation.[4]

4 Reasoning with trust

What we have presented so far explains how agent Ag_i can reason about the trustworthiness of its acquaintances. The reason for doing this is so Ag_i can use its trust information to decide how to use information that it gets from those acquaintances. To formalize the way in which Ag_i does this, we will assume that, in addition to Δ^{tr}_i, Ag_i has a set of beliefs about the world Δ^{bel}_i (which we assume come with some measure of belief), and some information Δ^j_i provided by each of its acquaintances Ag_j, and that:

$$\Delta_i = \Delta^{tr}_i \cup \Delta^{bel}_i \cup \bigcup_j \Delta^j_i$$

[4] Though [19] shows that propagation based on co-citation matches empirical results for the way people propagate trust, our experience is that people also often find the notion of co-citation somewhat unconvincing when they are first exposed to it.

$$Ax^{bel} \quad \frac{(n : \theta : \tilde{d}) \in \Delta_i^{bel}}{\Delta_i \vdash_{bel} (\theta : G : \{Ax^{bel}\} : \tilde{d})}$$

$$\text{Trust} \quad \frac{\Delta_i^{tr} \vdash_{tr} (trusts(i,j) : G : R : \tilde{d}) \text{ and } \Delta_i^{j} \vdash_{bel} (\theta : H : S : \tilde{e})}{\Delta_i \vdash_{bel} (\theta : G \cup H : R \cup S \cup \{Trust\} : ttb(\tilde{d}) \otimes^{bel} \tilde{e})}$$

$$\wedge\text{-I} \quad \frac{\Delta_i \vdash_{bel} (\theta : G : R : \tilde{d}) \text{ and } \Delta_i \vdash_{bel} (\phi : H : S : \tilde{e})}{\Delta_i \vdash_{bel} (\theta \wedge \phi : G \cup H : R \cup S \cup \{\wedge\text{-I}\} : \tilde{d} \otimes^{bel} \tilde{e})}$$

$$\rightarrow\text{-E} \quad \frac{\Delta_i \vdash_{bel} (\theta : G : R : \tilde{d}) \text{ and } \Delta_i \vdash_{bel} (\theta \rightarrow \phi : H : S : \tilde{e})}{\Delta_i \vdash_{bel} (\phi : G \cup H : R \cup S \cup \{\rightarrow\text{-E}\}) : \tilde{d} \otimes^{bel} \tilde{e})}$$

Fig. 3. Part of the bel consequence relation

All of this information can then be used, along with the consequence relation from Figure 3, to construct arguments that combine trust and beliefs.

The proof rules in Figure 3 are based on those we introduced in [30]. The rule Ax^{bel}, as in the previous set of proof rules, bootstraps an argument from a single item of information, while the rules $\wedge$-I and $\rightarrow$-E are typical natural deduction rules — the rules for introducing a conjunction and eliminating implication — augmented with the combination of degrees of belief, and the collection of information on which data and proof rules have been used. (The full consequence relation would need an introduction rule and elimination rule for every connective in the language, and the definition of these is easy enough — we omit them here in the interest of space.)

The key rule in Figure 3 is the rule named Trust. This says that if it is possible to construct an argument for θ from some Δ_j^i, indicating that the information comes from Ag_j, and Ag_i trusts Ag_j, then Ag_i has an argument for θ. The grounds of this argument combine all the data that was used from Δ_j^i and all the information about trust used to determine that Ag_i trusts Ag_j, and the set of rules in the argument record all the inferences needed to build this combined argument. Finally, the belief that Ag_i has in the argument is the belief in θ as it was derived from Δ_j^i combined with the trust Ag_i has in Ag_j. We carry out this last combination by first turning the trust value into a belief value using some suitable function $ttb(\cdot)$.

In other words, this rule sanctions the use of information from an agent's acquaintances, provided that the degree of belief in that piece of information is modified by the agent's trust in that acquaintance. Thus one agent can only import information from another agent if the first agent can construct a trust argument that determines it should trust the second (and so trigger the Trust rule).

5 Example

To see how this combined system might work, consider the rest of the example from [23] that goes with Figure 1 (suitably modified to provide an example of co-citation, which is not considered in the original). The trust network from [23] is based on data from the FilmTrust site[5] which features social networks centered around the exchange of information about films.

In the example, $john$ has the following information, where x is a universally quantified variable, $almodovar$ is the director Pedro Almodovar, and hce is an abbreviation for the 2002 film *Hable con ella* (Talk to her):

$$\begin{array}{ll}\Delta^{bel}_{john} & (j1 : SpanFilm(hce) : 1) \\ & (j2 : DirBy(almodovar, hce) : 1) \\ & (j3 : Comedy(x) \rightarrow \neg Watch(x) : 0.8)\end{array}$$

We take this to mean that $john$ thinks that hce is a Spanish language film, and that it is directed by Almodovar. In addition, he doesn't much like to watch comedies. $john$ also has some information from FilmTrust connections:

$$\Delta^{mary}_{john} \ (jm1 : IndFilm(hce) : 1)$$

$$\Delta^{jane}_{john} \ (jj1 : IndFilm(x) \wedge SpanFilm(x) \rightarrow \neg Watch(x) : 1)$$

$$\Delta^{dave}_{john} \ (jd1 : DirBy(x, almodovar) \rightarrow Watch(x) : 1)$$

$$\Delta^{paul}_{john} \ (jp1 : Comedy(hce) : 0.6)$$

Thus $john$ hears from $mary$ that hce is an independent film, from $jane$ that her advice is to not watch Spanish independent films, from $dave$ who says any of Almodovar's films are worth seeing, and from $paul$ who points out that he thinks hce is a comedy.

Now, we have already seen how $john$ can construct arguments for trusting $jane$ and $paul$, though we did not say what $\otimes^{tr}$was so that we could not compute the degrees of trust. For now, we follow [44] in taking $\otimes^{tr}$to be minimum, thus giving us:

$$\Delta^{tr}_{john} \vdash_{tr} (trusts(john, jane) : \{t1, t2\} : \{Ax^{tr}, Ax^{tr}, dp\} : 0.7)$$

and

$$\Delta^{tr}_{john} \vdash_{tr} (trusts(john, paul) : \{t1, t2, t4, t5\} : rules_1 : 0.4)$$

$john$ can also infer:

$$\Delta^{tr}_{john} \vdash_{tr} (trusts(john, dave) : \{t1, t3\} : \{Ax^{tr}, Ax^{tr}, dp\} : 0.7)$$

in exactly the same way as he infers trust about $jane$. He can also construct the following argument for trusting $mary$:

$$\Delta^{tr}_{mary} \vdash_{tr} (trusts(john, mary) : \{t1\} : \{Ax^{tr}\} : 0.9)$$

Each of the arguments can then be used with $\vdash_{bel}$ (Figure 3) to construct arguments that are relevant to the question of whether $john$ should watch hce. Using information from $jane$ he can determine:

$$\Delta_{john} \vdash_{bel} (\neg Watch(hce) : \{t1, t2, jj1, jm1, j1\} : rules_2 : \tilde{b})$$

[5] http://trust.mindswap.org/FilmTrust/

where

$$rules_2 = \{Ax^{tr}, Ax^{tr}, dp, Trust, Trust, Ax^{bel}, \wedge\text{-I}, \rightarrow\text{-E}\}$$

This shows that after the derivation of information about trusting $jane$, the proof of $\neg Watch(hce)$ requires the application of $Trust$ to establish a degree of belief in $jane$'s information, $Trust$ to import $jm1$ from $mary$, an application of Ax^{bel} to create an argument from $j1$, the use of $\wedge$-I to combine the data from $j1$ and $jm1$, and then $\rightarrow$-E to get the conclusion.

To establish $\tilde{b}$, we need to determine what the function $\otimes^{bel}$ is, and how to convert trust values to beliefs using $ttb(\cdot)$. For our purposes here, the choice doesn't matter greatly — we aren't arguing that any particular combination of operations for trust combination, belief combination and $ttb(\cdot)$ is best, just that if we have these operations then $john$ can use information in a way that seems to be useful. For now we handle beliefs using possibility theory [5] — which is basically equivalent to the approach adopted by [3] to handle variable strength arguments — and interpret the degree of trust in an agent to be a degree of belief that what the agent says is true [14, 32], so that $ttb(\cdot)$ is just the identity. All of this means that $\tilde{b} = 0.7$.

$john$ can also construct the following arguments as a result of information from, respectively, $paul$ and $dave$, in much the same way as the argument above. First we have:

$$\Delta_{john} \vdash_{bel} (\neg Watch(hce) : \{t1, t2, t4, t5, jp1, j3\} : rules_3 : 0.4)$$

where

$$rules_3 = \{Ax^{tr}, Ax^{tr}, Ax^{tr}, Ax^{tr}, dp, cc, Trust, Ax^{bel}, \rightarrow\text{-E}\}$$

and second we have:

$$\Delta_{john} \vdash_{bel} (Watch(hce) : \{t1, t3, jd1, j1, j2\} : rules_4 : 0.6)$$

where

$$rules_4 = \{Ax^{tr}, Ax^{tr}, dp, Trust, Ax^{bel}, Ax^{bel}, \rightarrow\text{-E}\}$$

This means that $john$ has three arguments that bear on his decision about whether to watch hce, one in favor and two against.

6 Using trust values

At this point in the example, we have arguments for opposing conclusions — $john$ should watch hce and $john$ should not watch it. To reach a decision about hce, $john$ needs to choose between these conclusions. There are a number of different approaches to using the trust information to do this, and in this section we discuss some of them, showing how they affect the example. The aim here is not to provide a definitive answer but to explain some of the options — as we hope that these examples will demonstrate, it is not immediately clear which is the best approach.

6.1 Flattening

The first approach is for $john$ to proceed by combining the arguments for the formula $\neg Watch(hce)$ (what [35] calls "flattening" the arguments) and seeing if the resulting combination outweighs the argument for $Watch(hce)$. We have three arguments to consider:

$$\begin{array}{ll} A_1 & (\neg Watch(hce) : \{t1, t2, jj1, jm1, j1\} : rules_2 : 0.7) \\ A_2 & (\neg Watch(hce) : \{t1, t2, t4, t5, jp1, j3\} : rules_3 : 0.4) \\ A_3 & (Watch(hce) : \{t1, t3, jd1, j1, j2\} : rules_4 : 0.6) \end{array}$$

Flattening combines the two beliefs, 0.7 and 0.4 for $\neg Watch(hce)$, to get a combined measure. Given that we are taking the values to be possibility values, it makes sense to combine them using $\max$, thus getting a combined value of 0.7 for $\neg Watch(hce)$. This is greater than the 0.6 for $Watch(hce)$, and so under this scheme, $john$ would conclude that he should not watch hce.

Given the choice of combination operator for flattening, this approach is very simple — the choice supported by the strongest single argument will always win. It also largely ignores conflicts between the arguments. In the example so far, we just have arguments that rebut one another, and the result of flattening seems very reasonable. But what if we have more conflicts? Consider extending the example so that $john$ has additional information:

$$\begin{array}{ll} \Delta^{bel}_{john} & (j1 : SpanFilm(hce) : 1) \\ & (j2 : DirBy(almodovar, hce) : 1) \\ & (j3 : Comedy(x) \rightarrow \neg Watch(x) : 0.8) \\ & (j4 : DirBy(almodovar, x) \rightarrow \neg IndFilm(x) : 1) \end{array}$$

so $john$ is now certain that anything directed by Almodovar is not an independent film. This gives him an additional argument:

$$A_4 \qquad (\neg IndFilm(hce) : \{j2, j4\} : \{Ax^{bel}, Ax^{bel}, \rightarrow\text{-E}\} : 1)$$

Thus $john$ now has a strong argument against hce being an independent film, and this clearly conflicts with A_1 since it contradicts the information from $mary$ about hce being an independent film. A_4 however, is ignored by flattening, and this doesn't seem very reasonable.

6.2 Acceptability analysis

Of course, handling this kind of conflict is exactly what Dung's acceptability semantics [11] and subsequent variations on this theme [6, 12] are intended to do. Let's examine what they tell $john$ in this scenario. [11] starts from the position of knowing which arguments conflict, assuming a relation that specifies:

$$attacks(A_n, A_m)$$

for all conflicts between arguments. Since we are starting from a less abstract position, we need to define what constitutes this relation in our example. The notion of conflict

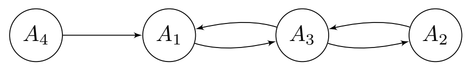

Fig. 4. The argumentation graph for the film example when the strengths of arguments are not taken into account.

between arguments used in [3] translates into our formulation of an argument as saying that $(c : G : R : v)$ attacks $(c' : G' : R' : v')$ if there is some $g \in G'$ such that $c \equiv \neg g$. That is one argument attacks another by disputing the truth of one of its grounds, "undercutting" it in the usual terminology.[6] ([3] also places some constraints on the strengths of the arguments v and v', but we will leave those for now.)

We will extend this notion of attack to include arguments rebutting each other, so that for our purposes $(c : G : R : v)$ attacks $(c' : G' : R' : v')$ if either $c \equiv \neg c'$ or there is some $g \in G'$ such that $c \equiv \neg g$. With this definition we have:

$$\begin{array}{l} attacks(A_1, A_3) \\ attacks(A_3, A_1) \\ attacks(A_2, A_3) \\ attacks(A_3, A_2) \\ attacks(A_4, A_1) \end{array}$$

and the argument graph is that of Figure 4. What $john$ concludes from this depends on the way that he computes which arguments are acceptable. However, none of the different approaches from [11] will help him decide what to watch. If he applies the grounded semantics, the only acceptable argument is A_4, which doesn't tell him what to watch. If he applies the complete, preferred or stable semantics, they will all tell him that A_4 is acceptable along with A_2 or A_3, but give no further guidance. As a result, while in other scenarios this analysis may suffice, in this case it leaves $john$ no wiser about whether he should watch hce or not.[7]

Since the basic acceptability analysis is not very informative, and since we have a degree of belief associated with each argument, we can incorporate the degrees of belief into the analysis. To do this, we extend our notion of $attack$ with the mechanism that [3] uses to handle strength of arguments. Broadly speaking (and counting rebutting as well as undercutting arguments), what [3] says is that $(c : G : R : v)$ attacks $(c' : G' : R' : v')$ if either $c \equiv \neg c'$ or there is some $g \in G'$ such that $c \equiv \neg g$, and $v \geq v'$. Thus if an argument has a conflict with a strictly stronger argument, that conflict is ignored in establishing the $attacks$ relation. With this definition we have:

[6] The term "undercutting" was originally used by Pollock, for example in [40], to refer to the situation in which one argument attacked an inference in another, but in the computer science community the term was rapidly co-opted to mean the kind of attack we describe here [3, 7, 42].

[7] The grounded semantics can't untangle the rebutting conflict between A_2 and A_3, while the other semantics tell $john$ that the rebutting means one of the arguments is acceptable, but they can't make a choice between the arguments. All the semantics determine that A_4 makes A_1 unacceptable, and hence unable to have any effect on the conflict between A_2 and A_3.

Fig. 5. The argumentation graph for the film example when the strengths of arguments are taken into account.

$attacks(A_1, A_3)$
$attacks(A_3, A_2)$
$attacks(A_4, A_1)$

and the argument graph is that of Figure 5. This time, any of the standard semantics from [11] tells $john$ that the acceptable arguments are A_3 and A_4, and so his conclusion using this approach is that he should watch hce.

The approaches we have discussed up to now are direct applications of existing approaches to using arguments with some form of belief value, and only use the trust information as a mechanism to establish arguments about beliefs. Our investigation is also considering three other approaches, in which we use the trust value directly. We will discuss these next.

6.3 Trust thresholds

The first of these new approaches is the use of *trust thresholds*. The formal model we are using here considers an agent to have information from a number of acquaintances, each of which has some trust rating that is applied to the information from that agent. A natural approach to using the trust rating is to specify a threshold value below which information from an agent is disregarded.

In the case of our film example, $john$ might set his trust threshold to 0.5, thus not accepting information from any acquaintance y for which he cannot infer:

$$(trusts(john, y) : G : R : v)$$

for some $v > 0.5$. (One might formulate this as an additional condition in the Trust rule in the $\vdash_{bel}$ relation.) Doing this would rule out any information from $paul$, and hence $john$ would only have A_1, A_3 and A_4. Of course, using the threshold doesn't answer $john$'s question on its own — he still has arguments for and against watching hce, so he will have to use a method like those outlined above to resolve the conflict. If, for example, $john$ chooses to use the acceptability semantics without considering the strengths of the arguments, this time he will find that all the standard semantics say that A_3 is acceptable and so he should $watch(hce)$. (The outcome of the two other approaches are not affected by the threshold, but it does mean that there are fewer arguments to consider.)

A number of questions arise about the use of thresholds. To what extent, for example, does imposing such a threshold on the information from its acquaintances protect an agent from using untrustworthy information? In other words, does excluding information from acquaintances with a trust value below some α mean that all of the agent's conclusions will be more trustworthy than α? Or are there circumstances under which less trustworthy conclusions could be reached even if data from agents below the threshold is excluded? We have shown that under some circumstances the trust threshold will give us this protection [38], but in case of our example, it won't. Imagine that the threshold is set to 0.65, ruling out data from any agent except $mary$ and $jane$, so $john$ has just A_4 and A_1 (and so no opinion about whether to watch hce because the only attack is that of A_4 on A_1 which makes A_1 unacceptable). Can this be altered by information below the threshold, say from $mary$, who is highly trusted, but maybe has some low belief information about the watchability of hce? It might. If $mary$ has information that leads to an argument A_5 with conclusion $watch(hce)$ and a belief of 0.5 say, it won't be excluded by the threshold (which only applies to $mary$ not to data from $mary$), and A_5 will be acceptable (because the attacking argument A_1 is itself attacked by A_4), giving the conclusion $watch(hce)$. Our current work is trying to establish what are reasonable levels of protection that may be provided by trust thresholds, and for which combinations of interpretations for trust and belief values the levels of protection hold.

Now, given an arbitrary threshold, there may be no arguments for or against watching hce for which the grounds are all above the threshold — meaning that $john$ has no arguments to consider — but many arguments with elements of their grounds just below the threshold — meaning that $john$ would consider them if the threshold was lower. For such cases $john$ might want to consider altering the threshold, and so we are interested in how the protection offered by the threshold is altered when the threshold moves.

Another interesting question is to examine the interaction between thresholds and propagation in the trust network. What correspondence is there between imposing a trust threshold and pruning the acquaintances from the network? Clearly when we combine trust values along a path through the network using $\min$, a threshold will rule out trusting any agent downstream of an agent below the threshold, but this may not necessarily be the case when trust values are computed in different ways. Again, this is a matter that we are currently investigating.

6.4 Trust budget

The second new approach is, in some ways, an extension of the first. Using a trust threshold rules out acquaintances — or alternatively conclusions that are supported by information from those acquaintances — when the level of trust in an acquaintance drops below a particular level. Thus very untrustworthy acquaintances, and the information they provide, are ruled out. But equally, information from sources above the threshold is ruled in, along with conclusions based on it, even if a given conclusion depends upon lots of items of information that came from sources close to the threshold, and so might be considered more suspect than others based on sources further from the threshold.

The notion of a *trust budget* is intended to deal with this situation. A trust budget specifies the total amount of distrust that is permitted in the sources of data that lead to a single conclusion. In situations where trust values are, as in our example, between

0 and 1, we can compute the "cost" of Ag_i accepting information from a series of acquaintances Ag_j as:

$$\sum_j 1 - tr(i, j)$$

To illustrate this idea on the example, let us first imagine that $john$ sets the trust budget to 1. Given the levels of trust that $john$ has in his acquaintances, this allows him to accept information from at most any three of $jane$ (cost to the trust budget of 0.3), $paul$ (0.6), $dave$ (0.3) and $mary$ (0.1). For example, $john$ might spend the whole trust budget and accept information from $jane$, $paul$ and $mary$, giving him the conclusion that he should not watch hce. Or he might spend part of the budget accepting information from $jane$, $dave$ and $mary$, from which he would conclude that he should watch hce.

Given a specific budget, $john$ can identify which conclusion or conclusions that fit within the budget have the highest belief (here it is $\neg watch(hce)$). Alternatively $john$ might consider slowly increasing his trust budget from 0 until he reaches a conclusion about the question he is interested in — here he would have to "spend" at least 0.3 to get a conclusion (in this case to not watch hce, based on A_1 obtained by accepting information from $jane$). Another approach to using the trust budget would be to have $john$ establish what he needs to "spend" in order to find a conclusion he wants. In the context of the example, let's imagine he is interested in watching hce but wants to know how trusting he has to be to decide that it is a good idea. The minimum budget necessary to establish $watch(hce)$ as a conclusion is 0.6, the cost of trusting $paul$, since it only takes information from $paul$ to construct an argument for $watch(hce)$ (in more complex examples it might be necessary to trust several agents to reach an interesting conclusion).[8]

In general, the questions to ask about a trust budget are similar to those for a trust threshold, identifying how well-behaved this notion is, and what protection an agent gets by imposing such a budget. These questions are, like those for trust thresholds, subjects of our ongoing research. Furthermore, as suggested by [13], in the context of the related notion of an "inconsistency budget", and [26], in the context of optimal trust path selection, the kinds of uses we are seeking to make of the trust budget are uses that will require considerable computation. This is another topic we are considering.

6.5 Meta-argumentation

The previous two approaches are concerned with handling the values derived from the trust network. These values are then used to make decisions about which piece of information, and thus which arguments (since arguments are derived from the information) are considered by an agent. The final approach we are looking at leans more towards the kind of structural analysis described by Loui [28], where heuristic patterns of evidence and argument structure are used to decide which arguments are preferred. An example is the preference for arguments using only data from agents that are directly trusted by Ag_i over arguments that use data from agents that Ag_i trusts by co-citation. The aim of

[8] We are mainly interested in incorporating trust into planning, where the concept of establishing how much trust it "costs" to build an argument (plan) makes more sense than in the domain of the example.

this approach is to identify general heuristics for dealing with trust data, and to verify the plausibility, or otherwise, of the kinds of inference that they sanction.

7 Summary

In this paper, we have outlined work on reasoning about trust using a form of argumentation which, as the paper demonstrated, can be integrated with a system of argumentation that uses the conclusions about trust. A notable feature of the system for reasoning about trust is its flexibility — new approaches to propagating trust can easily be added (or, indeed, removed) by altering the proof rules that are used in propagation. The combined system was illustrated with an example, and current directions sketched.

Clearly the systems we have described are work in progress. Neither of the formal systems is complete as presented — both are missing much of the proof mechanism and a proper description of the syntax at the very least — and neither is rigorously evaluated. Our aim was simply to illustrate the basic ideas captured in the systems, and to illustrate the possibilities that they offer. We have also completely ignored the computational aspects of implementing a software system that employs these approaches. Our future work will, in due course, fill in the details that are missing here, more completely relate this work to approaches with similar aims, such as [29, 33], and provide an implementation. However, we believe that the work we have presented here has value in describing an area of research that we think is interesting and identifying some new approaches to handling it.

Acknowledgements

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 would like to thank the reviewers of this paper for their many interesting ideas, most of which we did not, thanks to the short time between notification and the deadline for completing revisions, have time to incorporate. We will strive to do so in future work. We would also like to thank a reviewer of an earlier paper on trust who suggested that we extend the film example with the notion that Almodovar is not really an independent film director. That reviewer was responsible for all the structural interest in the example.

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*, Banff, Alberta, May 2007.
3. L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artifical Intelligence*, 34(3):197–215, 2002.
4. 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.
5. S. Benferhat, D. Dubois, and H. Prade. Representing default rules in possibilistic logic. In *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, 1992.
6. M. W. A. Caminada. Semi-stable semantics. In *Proceedings of the 1st International Conference on Computational Models of Argument*, Liverpool, UK, September 2006.
7. C. I. Chesñevar, A. G. Maguitman, and R. P. Loui. Logical models of argument. *ACM Computing Surveys*, 32(4):337–383, 2000.
8. 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.
9. D. B. DeFigueiredo and E. T. Barr. TrustDavis: A non-exploitable online reputation system. In *Proceedings of the 7th IEEE International Conference on E-Commerce Technology*, 2005.
10. X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. In *Proceedings of the 35th International Conference on Very Large Databases*, Lyon, France, August 2009.
11. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
12. P. M. Dung, P. Mancarella, and F. Toni. A dialectical procedure for sceptical, assumption-based argumentation. In *Proceedings of the 1st International Conference on Computational Models of Argument*, Liverpool, UK, September 2006.
13. P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, (in press).
14. D. Gambetta. Can we trust them? In D. Gambetta, editor, *Trust: Making and breaking cooperative relations*, pages 213–238. Blackwell, Oxford, UK, 1990.
15. A. J. García and G. Simari. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.
16. F. Geerts, A. Kementsiedtsidis, and D. Milano. Mondrian: Annotating and querying databases through colors and blocks. In *Proceedings of the 22nd International Conference on Data Engineering*, Atlanta, April 2006.
17. J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *Proceedings of the International Provenance and Annotation Workshop*, Chicago, Illinois, May 2006.
18. T. Grandison and M. Sloman. A survey of trust in internet applications. *IEEE Communications Surveys and Tutorials*, 4(4):2–16, 2000.
19. 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.
20. C-W Hang, Y. Wang, and M. P. Singh. An adaptive probabilistic trust model and its evaluation. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems*, Estoril, Portugal, 2008.
21. A. Jøsang, C. Keser, and T. Dimitrakos. Can we manage trust? In *Proceedings of the 3rd International Conference on Trust Management*, Paris, May 2005.

22. 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.
23. Y. Katz and J. Golbeck. Social network-based trust in prioritzed default logic. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.
24. Y. Kuter and J. Golbeck. SUNNY: A new algorithm for trust inference in social networks using probabilistic confidence models. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, 2007.
25. J. Lang, M. Spear, and S. F. Wu. Social manipulation of online recommender systems. In *Proceedings of the 2nd International Conference on Social Informatics*, Laxenburg, Austria, 2010.
26. G. Li, Y. Wang, and M. A. Orgun. Optimal social trust path selection in complex social networks. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, GA., 2010.
27. L. Li and Y. Wang. Subjective trust inference in composite services. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, GA., 2010.
28. R. P. Loui. Defeat among arguments: a system of defeasible inference. *Computational Intelligence*, 3(3):100–106, 1987.
29. 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*, Toronto, Canada, May 2010.
30. P. McBurney and S. Parsons. Tenacious tortoises: A formalism for argument over rules of inference. In *Proceedings of the ECAI Workshop on Computational Dialectics*, Berlin, 2000.
31. L. Mui, M. Moteashemi, and A. Halberstadt. A computational model of trust and reputation. In *Proceedings of the 35th Hawai'i International Conference on System Sciences*, 2002.
32. D. Olmedilla, O. Rana, B. Matthews, and W. Nejdl. Security and trust issues in semantic grids. In *Proceedings of the Dagstuhl Seminar, Semantic Grid: The converegance of technologies*, volume 05271, 2005.
33. N. Oren, T. Norman, and A. Preece. Subjective logic and arguing with evidence. *Artificial Intelligence*, 171(10–15):838–854, 2007.
34. 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.
35. S. Parsons. On precise and correct qualitative probabilistic reasoning. *International Journal of Approximate Reasoning*, 35:111–135, 2004.
36. S. Parsons, K. Haigh, K. Levitt, J. Rowe, M. Singh, and E. Sklar. Arguments about trust. Technical report, Collaborative Technology Alliance, 2011.
37. 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.
38. S. Parsons, Y. Tang, E. Sklar, P. McBurney, and K. Cai. Argumentation-based reasoning in agents with varying degrees of trust. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems*, Taipei, Taiwan, 2011.
39. J. Pasternak and D. Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.
40. J. Pollock. *Cognitive Carpentry*. MIT Press, Cambridge, MA, 1995.
41. H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.
42. H. Prakken and G. Sartor. Argument-based logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 1997.

43. 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.
44. M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, 2003.
45. J. Sabater and C. Sierra. Review on computational trust and reputation models. *AI Review*, 23(1):33–60, September 2005.
46. A. Salehi-Abari and T. White. Trust models and con-man agents: From mathematical to empirical analysis. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, Atlanta, Georgia, 2010.
47. Y. Tang, K. Cai, E. Sklar, P. McBurney, and S. Parsons. A system of argumentation for reasoning about trust. In *Proceedings of the 8th European Workshop on Multi-Agent Systems*, Paris, France, December 2010.
48. 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*, Estoril, Portugal, 2008.
49. 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*, Boston, MA, 2006.
50. 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.
51. B. Yu and M. Singh. Distributed reputation management for electronic commerce. *Computational Intelligence*, 18(4):535–349, 2002.
52. 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.