

Networks of Trust and Distrust: Towards Logical Reputation Systems

W. T. Harwood, J. A. Clark, J. L. Jacob

University of York
Department of Computer Science

Abstract. We introduce the notion of a network of trust and distrust relations between individuals and take an argumentation approach to the assessment of whether one individual should trust another.

... good decision is based on knowledge and not on numbers"

Plato - Early Dialogues - Laches

1 Introduction

This paper reports ongoing work in creating a logical foundation for reasoning about trust and trustworthiness in networks of individuals that may recommend one another as trustworthy or untrustworthy. One solution is to adopt some form of voting or counting scheme as in commonly done in reputations systems [10]. But in many circumstances, when the stakes are sufficiently high, e.g. deciding to trust a root certificate or disclose confidential information, weight of numbers does not constitute a good argument. As Plato puts it "... good decision is based on knowledge and not on numbers"¹.

One of the ultimate goals of this work is to provide the foundations for a logically well founded trust management system, or *Logical Reputation System*, where 'reputation' is computed by maintaining some notion of consistency between trust assertions made by trusted individuals. This approach contrasts

¹ For those without a classical education, or more relevantly today, an Internet connection, this is part of a general argument that Plato directs against amalgamating opinions as a basis for reaching a good decision. This is actually a cornerstone of Plato's arguments against democracy. Today we take a more liberal view and regard some decisions as being appropriately arrived at by amalgamating individual opinions (such as who should rule the country, or what colour should we paint the school) and other decisions as arrived at by knowledge. It is certainly the contention of this paper that trust is best arrived at through knowledge rather than opinion.

with trust models, such as those of Coleman[6] or Marsh[12], that appeal to probabilities of trustworthiness or any similar numeric notions of degree of trustworthiness. Rather, in the approach considered here, a trust judgment is a purely logical resolution of possibly conflicting trust arguments. The intent is to use such a system to automatically make trust judgements in social network applications based on relational information gathered from users. This paper aims at setting out a logical framework based on argumentation theory to achieve this goal.

Our starting point is to consider networks of individuals that assert that they trust some individuals and distrust others.

Trust and distrust² are statements about the relationship between two individuals in relation to some action, such as, *information disclosure*, that holds in some context, such as, *today, in this building* (see, for example, Hardin's discussion in [9]). Throughout this paper we will consider the action and context as fixed so that we may talk of trust and distrust as binary relations. It should be apparent that we can put the additional dimensions back into the picture by considering families of relations parameterized by action and context.

If we only had information to the effect that certain individuals were trustworthy we would have a *web of trust* model (see, e.g. Zimmermann [13]) in which one individual trusts another if there is a trusted path between them. Here we consider how such models may be extended in the presence of additional negative assertions to the effect that certain individuals distrust one another. This allows the possibility of a trust path being undermined by a *distrust path*. Here we present a model of such systems in three stages of increasing complexity.

The first stage, *simple trust systems*, captures the idea that an individual trusts another if there is a trust path between them that is not undermined by distrust. Simple trust systems are modeled after argumentation theory[3, 4, 7]. Essentially, the approach is to assess the soundness of the argument that an individual, x_0 say, can trust an individual x_n . In our case the argument for trust is the existence of a trust path between x_0 and x_n in a network of trust relations. However, this argument may be undermined by an attack on it. An attack is an argument that some link in the chain of trust from x_0 to x_n is untrustworthy. In our case, such an argument is the existence of a path of trust from x_0 to some node y_m such that y_m *distrusts* some node connecting x_0 and x_n (including x_n itself). The existence of such an attack would make the original argument unsound, unless, of course, the attack itself was attacked in a similar manner, etc. etc.

² The relationship between trust and distrust is far from uncontroversial, see, for example, the discussions in the collection of articles [8]. We take distrust as more than the mere absence of trust. That is, distrust is not simply the complement of trust. Rather, trust and distrust are two relations that can exist between individuals and it is even possible for an individual to trust and distrust another individual simultaneously about the same topic. In such cases, although the individual is conflicted about trust, they are not logically inconsistent about trust.

The argumentation theory approach to resolving the set of attacks and counterattacks is to say that the original argument is sound if it is possible to partition the set, S , containing the original argument and the closure of all the attacks and counter attacks possible based on the initial argument, into two distinct sets, which we call S^+ and S^- , such that: S^+ is consistent in that no paths in S^+ attack one another; S^+ contains the original trust path; and for every path in S^- that attacks a path in S^+ , S^+ contains a path that counter attacks that path.

Although formally straightforward, simple trust systems fail to capture an important aspect of trust: that when faced with a choice over conflicting recommendations of who to trust we have preferences over the choices. This leads to the formulation of the second stage, *preferential trust systems*, which introduces the notion that individuals may rank the other individuals into a partial ordering indicating their relative efficacy at making trust or distrust recommendations. This relative ranking is then extended to a partial order on paths which is used to measure the relative strength of paths. A distrust path can only undermine another path if it is sufficiently strong when compared to the path it is attacking (up to the point of attack). This second form of system is formalized by revising the notion of attack between paths.

The final stage *asymmetric preferential trust systems* addresses the fact that, in many situations, individuals have an asymmetric attitude to trust and distrust in that they are more willing to accept an argument that leads them to distrust than they are to accept one that leads them to trust. In the approach considered here, individuals require stronger arguments to make them trust than they do to make them distrust.

In order to directly describe the relationship between individuals, individuals' efficacy assessments, trust paths and distrust paths, trust systems are described relatively concretely. Of course these systems may be considered more abstractly using Dung's abstract argumentation systems framework. The connection between trust systems and Dung's framework is sketched in section 7.

2 Trust Systems

First we set out the framework of trust systems that we use throughout the paper.

A trust system is a collection of individuals I each of which may assert some collection of propositions, P_i for $i \in I$, and two binary relations $Trust : I \Leftrightarrow I$ and $Distrust : I \Leftrightarrow I$. If an individual, say x_0 , trusts another individual, say x_n , then x_0 accepts P_n as true. If however x_0 distrusts x_n then x_0 neither accepts P_n as true nor rejects P_0 as false.

Informally, a trust system is a collection of individuals each of which may make assertions about the state of the world. In particular, each individual may assert whether or not they regard some other individuals as trustworthy or untrustworthy. If an individual i regards an individual j as trustworthy we

will say that i trusts j . If, on the other hand, i regards j as untrustworthy we will say that i distrusts j . It is also possible for i to neither trust nor distrust j . If i trusts j then i is willing to accept j 's assertions as true. In particular i accepts j 's assertions about the trustworthiness of others as true. If i accepts a trust assertion of j as true e.g. if j trusts k , then i accepts there is an argument for trusting k , specifically i trusts j and j trusts k . If, however, j distrusts k then i accepts there is an argument that k is untrustworthy i.e. j whose assertions i trusts, distrusts k . It should be clear at this point that trust arguments can be extended (i.e. if i trusts j , j trusts k and k trusts l , then there is an argument for i trusting l) but distrust arguments cannot (i.e. if i trusts j , j distrusts k and k trusts l , then, since j does not accept k 's assertions there is neither a trust argument nor a distrust argument, derivable from these facts alone, linking i and l).

Formally a trust system is a collection of individuals I and two binary relations $Trust : I \Leftrightarrow I$ and $Distrust : I \Leftrightarrow I$. Arguments for the trustworthiness and untrustworthiness of individuals will be modeled as trust paths and distrust paths between individuals. A trust path from x_0 to x_n is a sequence $\langle x_0, x_1, \dots, x_{n-1}, x_n \rangle$ such that every pair (x_i, x_{i+1}) is in $Trust$. A distrust path from x_0 to x_n is a sequence $\langle x_0, x_1, \dots, x_{n-1}, x_n \rangle$ such that every pair (x_i, x_{i+1}) for $i < n - 1$ is in $Trust$ and (x_{n-1}, x_n) is in $Distrust$. That is, the path $\langle x_0, x_1, \dots, x_{n-1} \rangle$ is a trust path and the final step $\langle x_{n-1}, x_n \rangle$ is distrusting.

The set of trust paths will be called TP and the set of distrust paths will be called DP .

Given a path, p , (either a trust path or a distrust path) then $range\ p$ is the set of all individuals in the path i.e. if $p = \langle x_0, \dots, x_n \rangle$ then $range\ p = \{x_0, \dots, x_n\}$. We will also say that $first\ p = x_0$ and $last\ p = x_n$, and, for later use, $front\ p = \langle x_0, \dots, x_{n-1} \rangle$.

A distrust path, q , *attacks* a path if it attacks the trust supporting the path, meaning it either attacks any point of a trust path (including its last node) or it attacks any point on the front of a distrust path (i.e. the trust path part of the distrust path).

Let $tr\ p$ be the *trust part* of a path p , defined by

$$tr\ p = \begin{cases} p & \text{if } p \in TP \\ front\ p & \text{if } p \in DP \end{cases}$$

Then *attacks* relation between paths is defined by:

$$q\ attacks\ p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range(tr\ p)$$

An attack is admissible if it satisfies an *admissibility condition* that varies between the three forms of trust system considered. For simple trust systems all attacks are admissible. For preferential trust systems an attack is admissible only if it is of adequate strength. For asymmetric trust systems the strength condition varies depending on the way the attack will affect the overall outcome. The following section illustrates the effect of the different admissibility conditions with a short example.

3 A Short Example

To illustrate the three systems consider the following network of trust and distrust.

- Alice trusts Bob,
- Bob trusts Carol,
- Carol trusts Dan,
- Dan distrusts Bob,
- Alice trusts Elizabeth,
- Elizabeth distrusts Dan.
- Does Alice trust Carol?

Under simple trust systems, where we have no other information, Dan's distrust of Bob defeats the chain of trust connecting Alice and Carol but Elizabeth's attack on Dan defeats it, and so cancels its effect, leaving Alice having trust in Carol.

If additionally we know:

- Alice rates herself, Bob and Carol strictly higher in ability to make trust judgements than she does Dan.

Then, under a preferential trust system, Alice will trust Carol because Dan's distrust of Bob will lead to an attack path that is weaker than the trust path between Alice and Carol. If, however,

- Alice rates herself, Bob, Carol and Dan equally in ability to make trust judgements

then we would need to consider the exact formulation of the preference system: does an attack from a path of equal strength defeat the attacked path or not? Below we differentiate between *conservative* systems in which attacks must be strictly stronger to defeat a path and *paranoid* systems in which attacks from paths of equal strength, or attacks from incomparable paths, can defeat the attacked path.

Finally, to illustrate asymmetric trust systems, which are *conservative* if the consequence is trust and *paranoid* if the consequence is distrust, we consider two situations

1. Alice rates everyone equally in their ability to make judgements.
2. Alice rates Elizabeth higher than everyone else in her ability to make trust judgements.

Under assumption 1, Dan's distrust of Bob will lead to Alice not trusting Carol even though Elizabeth distrusts Dan, because the distrusting outcome is favored over the trusting outcome. Whereas under assumption 2, Elizabeth's distrust of Bob can cancel the attack and lead to Alice having trust in Carol.

The rest of the paper provides the technical details of each of the systems.

4 Simple Trust Systems

As mentioned above, in simple trust systems all attacks are admissible.

We wish to define notion of a *sound* trust path, p , between two individuals as a path which is either not attacked or only attacked by distrust paths that are themselves defeated by other attacks. To do this we first define the attack closure set of the path p to be p^a , the least set closed under:

- $p \in p^a$,
- $q \in p^a$ and $r \in \text{attacks}(q) \implies r \in p^a$.

and say p is sound if and only if p^a can be partitioned into two sets S^+ and S^- such that:

- $p \in S^+$,
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+ , i.e. $S^+ \cap \text{attacks}^\exists(S^+) = \emptyset$,
- S^+ *defends itself* against S^- in that every path in S^- that attacks a path in S^+ is itself attacked by a path in S^+ , i.e. $\text{attacks}^\exists(S^+) \subseteq \text{attacks}^\exists(S^+)$.

where, for a relation R , $R^\exists X$ is the forward image of X under R i.e. $\{y \mid \exists x \in X. xRy\}$.

We will call a set, S , a *support*, if it is consistent and defends itself (i.e. it can support some trust path p).

We say an individual x_0 *trusts* an individual x_n iff³ there is a sound trust path between x_0 and x_n .

Given a simple trust system $T = (I, \{P_i\}_{i \in I}, \text{Trust}, \text{Distrust})$ its set of sound trust paths (STP) is defined by:

$$\text{STP} = \{(x_0, x_n) \in I \times I \mid \exists p \in TP. \text{first}(p) = x_0 \wedge \text{last}(p) = x_n \wedge \text{sound}(p)\}$$

Two simple trust systems S and T are *trust equivalent* iff they have the same set of individuals and the same set of sound trust paths.

5 Preferential Trust Systems

Preferential trust systems restrict admissible attacks using a notion of relative strength between the attacked path and the attacking path. The particular notion that we use is that the strength of the path is derived from the competence, trustworthiness or reliability of the individuals in the path in making judgments about other individuals. We will settle on the neutral term *efficacy* for any of the terms competence, trustworthiness or reliability (or any other such notion).

In the above, all individuals have been regarded as of equal efficacy in rating the trustworthiness of other individuals. We will now consider what

³ Here, and throughout, we will adopt the convention of writing *iff* for *if and only if*.

happens when individuals are partially ordered by their efficacy in performing such rating. We will assume every individual i has available their own assessment of the relative efficacy of all other individuals at rating the trustworthiness of others. Formally we take this to be a family of partial orders (reflexive, anti-symmetric and transitive binary relations) over the set of individuals I , one for each member $i \in I$, denoted \succeq_i reflecting i 's view of the relative efficacy of individuals. Our goal is that, given a path p which is attacked by a path q , we wish to compare the strength of p up to the point of the attack, $last(q)$, with the strength of q . To do this we need to derive a partial ordering of paths from the partial ordering of the efficacy of the individuals in the paths.

We will call the segment of the path p up to the attack, $p \upharpoonright_{last(q)}$. If we were to use a strict total ordering to compare paths then we would say that one path, say q , was weaker than another, say p , when $range(q)$ contained an element less than any element in $range(p)$. We generalize this idea to partial orders by considering minimal elements in the ranges of the paths.

First we define an extension of a partial order over a set to a partial order over subsets of that set.

Given a partial order, \succeq , over a set S , we say that a subsets P and Q of S are comparable⁴, written $P \sim Q$ iff:

$$(\forall x \in P. \exists y \in Q. x \succeq y) \vee (\forall y \in Q. \exists x \in P. x \succeq y)$$

The set of minimal elements of a set P is defined as:

$$minimal(P) = \{x \in P \mid \forall y \in P. (y \succeq x) \implies y = x\}$$

A set, $P \subseteq S$, is at-least-as-strong-as a set, $Q \subseteq S$, written $P \sqsupseteq Q$, iff

$$P \sim Q \wedge \forall x \in minimal(P). \exists y \in minimal(Q). x \succeq y$$

A set P subset of S is stronger than a set Q subset of S , written $P \sqsupset Q$, iff

$$P \sqsupseteq Q \wedge Q \not\sqsupseteq P$$

All this amounts to is that subsets are ordered by comparing the least elements of the chains and if one of the subsets has strictly smaller elements for any of its chains (and the other does not) then it is the smaller set.

A path p is *stronger than* q , also written $p \sqsupset q$, iff

$$first(p) = first(q) \wedge last(p) = last(q) \wedge \\ range(p) \setminus \{last(p)\} \sqsupset range(q) \setminus \{last(q)\}$$

The removal of the last elements of the paths is due to the fact that we derive the efficacy of the individuals on the path that make the trust recommendations.

⁴ **Warning:** For those familiar with the notation $x \parallel y$ for x incomparable with y under the partial order \preceq . The notion defined here is over subsets of the ordering, not elements of the ordering. So $P \sim Q \equiv \exists p \in P, q \in Q. \neg(p \parallel q)$.

We now modify the definition of attack to take account of the relative strength of paths. There are two possible views of relative strength that correspond to whether the individual x_0 takes a *conservative* or a *paranoid* stance with respect to attacks. If x_0 takes a conservative stance, then a path is only defeated by a strictly stronger attack. If, on the other hand, x_0 takes a paranoid stance, then a path is defeated if the attacking path is incomparable or is at-least-as-strong-as the attacked path. The paranoid position allows attacks to defeat other attacks if x_0 is not in a position to positively assert that the attacked path is the stronger of the two.

That is, if x_0 has a conservative stance, then an attack, q , on a path, p , only succeeds if $q \sqsupset p \upharpoonright_{last(q)}$:

$$q \text{ attacks}_C p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range p \wedge q \sqsupset p \upharpoonright_{last(q)}$$

and if x_0 has a paranoid stance, then an attack, q , on a path, p , only succeeds if $p \upharpoonright_{last(q)} \not\sqsupset q$:

$$q \text{ attacks}_P p \equiv q \in DP \wedge first(q) = first(p) \wedge last(q) \in range p \wedge p \upharpoonright_{last(q)} \not\sqsupset q$$

Preferential trust systems are formulated by replacing the definition of attack in simple trust systems with either the conservative or the paranoid definition of attack⁵.

6 Asymmetric Preferential Trust Systems

In practice individuals are often asymmetric in their attitude to trust and distrust. That is, they are paranoid about trust and conservative about distrust. This means that the admissibility of an attack changes according to the overall role it plays in determining the outcome, introducing an asymmetry between paths which ultimately lead to a trust decision and paths which ultimately lead to a distrust decision. We capture this asymmetry by redefining the conditions for forming S and forming the partitions S^+ and S^- :

- The attack closure set S is the least closed set of paranoid attacks based on a trust path p as above.
- S^+ is restricted to only containing the initial trust path, p , and conservative attacks.
- Since conservative attacks are a subset of paranoid attacks, S^- may contain both types of attack.

Trust path p is sound iff it is possible to form a partition of S such that:

⁵ A system may also be formulated where the stance varies from individual to individual which is essentially a simple trust system with an indexed family of *attacks* operators.

- $p \in S^+$.
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+ .
- S^+ is *conservative* in that every path in S^+ is either p or a member of $\text{attacks}_C(x)$ for some x .
- S^+ *defends itself* against S^- in that every path in S^- that attacks a path in S^+ is itself attacked by a path in S^+ .

7 Connecting Trust and Dung's Abstract Argumentation

Dung [7] defines an abstract argumentation system as a pair $(AR, Attacks)$ where AR is a set of arguments and $Attacks$ is a binary relation over AR called the *attacks* relation. We write $x Attacks y$ for x attacks y . A set, $S \subseteq D$, attacks an argument, $x \in D$, if some argument in S attacks x (we will say that $S ATTACKS x$ for $\exists y \in S. y Attacks x$).

Dung then goes on to define the notions of:

- *Conflict free*: A set of arguments $S \subseteq AR$ is *conflict free* iff there is no pair of arguments $x \in S$ and $y \in S$ such that $x Attacks y$.
- *Acceptable*: An argument $x \in AR$ is *acceptable with respect to S* iff for every argument $y \in AR$ if $y Attacks x$ then $S ATTACKS y$. Following [2] we will also say that $S defends x$ when x is acceptable with respect to S .
- *Admissible*: A set $S \subseteq AR$ is *admissible* iff S is *conflict free* and each argument in S is *acceptable with respect to S*.

Dung then goes on to discuss various notions of semantics that further restrict the notion of admissibility which are not used in our current semantics.

To translate the above into Dung's framework we consider an individual a and the set of trust paths, P , rooted at a . For the simple trust systems:

- The set of arguments AR is the set P .
- The attacks relation between holds $q, p \in P$, i.e. $q Attacks p$, iff there exists a distrust path $d = \langle x_0, x_1, \dots, x_{n-1}, x_n \rangle$ with $q = \langle x_0, x_1, \dots, x_{n-1} \rangle$ and $d attacks p$.

Note that this definition of *Attacks* loses information by conflating multiple distinct attacks from q to different points on p .

A path p is *sound_D* iff P can be partitioned into two sets S^+ and S^- such that $p \in S^+$ and S^+ is admissible.

Clearly the above notions of consistency and conflict freeness are the same (albeit on different domains):

Proposition 1. $S \cap R^{\exists}(S) = \emptyset \equiv S \subseteq \overline{R^{\exists}(S)}$

Likewise, S defends itself and S is acceptable are essentially the same as demonstrated by the following two propositions.

First we introduce the dual of the forward image operator on binary relations over a set S : Given a binary relation $R : S \leftrightarrow S$, the function $R^{\forall} : \mathcal{P}S \rightarrow \mathcal{P}S$ is defined by:

$$R^\forall Y = \{x \mid \forall y. xRy \implies y \in Y\}$$

$R^\exists X$ is the forward image of X and $R^\forall Y$ is the set of elements in the inverse image of R that only result in elements in Y .⁶

R^\exists and R^\forall form a (covariant) galois connection, or axiomaticity, over S . This means that $R^\exists \circ R^\forall$ is an interior operator on S and $R^\forall \circ R^\exists$ is a closure operator on S . Letting \widetilde{R} represent the converse of R (i.e. $x \widetilde{R} y \equiv yRx$) then

Proposition 2. S is acceptable iff $S \subseteq (\widetilde{Attacks})^\forall (Attacks^\exists(S))$

Proof Sketch. This follows from $R^\forall X = \overline{(\widetilde{R})^\exists(\overline{X})}$ and Amgoud & Cayrol's theorem, quoted in [2], which rendered in our notation is S is acceptable iff $S \subseteq Attacks^\exists(\overline{Attacks^\exists S})$.

Proposition 3. S is acceptable iff $(\widetilde{Attacks})^\exists S \subseteq Attacks^\exists S$

Proof Sketch.

$$\begin{aligned} &\implies \\ &S \subseteq (\widetilde{Attacks})^\forall (Attacks^\exists(S)) \\ &\quad \text{by proposition 2} \\ &(\widetilde{Attacks})^\exists S \subseteq (\widetilde{Attacks})^\exists ((\widetilde{Attacks})^\forall (Attacks^\exists(S))) \\ &\quad \text{by } (\widetilde{Attacks})^\exists \text{ preserves order} \\ &(\widetilde{Attacks})^\exists S \subseteq Attacks^\exists(S) \\ &\quad \text{by } (\widetilde{Attacks})^\exists \circ (\widetilde{Attacks})^\forall \text{ being an interior operator} \\ &\longleftarrow \\ &(\widetilde{Attacks})^\forall ((\widetilde{Attacks})^\exists S) \subseteq (\widetilde{Attacks})^\forall (Attacks^\exists S) \\ &\quad \text{by } (\widetilde{Attacks})^\forall \text{ preserves order} \\ &S \subseteq (\widetilde{Attacks})^\forall (Attacks^\exists S) \\ &\quad \text{by } (\widetilde{Attacks})^\forall \circ (\widetilde{Attacks})^\exists \text{ being a closure operator} \end{aligned}$$

Proposition 4. $sound_D(p) \equiv sound(p)$

Proof Sketch. Since the definitions of S being a support and S being admissible are essentially the same between the two definitions of soundness, the major work falls on showing that the existence of a suitable partition of p^a is equivalent to the existence of a suitable partition of P .

⁶ $R^\forall Y$ is closely related to the weakest precondition operator in programming languages semantics. The exact relation depending on the particular relational theory of programs and termination used.

Recall

$$tr(p) = \begin{cases} p & \text{if } p \in TP \\ front(p) & \text{if } p \in DP \end{cases}$$

Let p be a trust path, then $tr^{\exists}(p^a) \subseteq P$. Assume the pair S^+, S^- form a suitable partition of p^a then the pair $tr^{\exists}(S^+), P \setminus tr^{\exists}(S^+)$ form a suitable partition of P .

Conversely, if the pair S^+, S^- form a suitable partition of P then the pair $\widetilde{tr}^{\exists}(S^+) \cap p^a, p^a \setminus (\widetilde{tr}^{\exists}(S^+) \cap p^a)$ form a suitable partition of p^a .

To obtain the corresponding Dungian systems for preferential and asymmetric trust systems we modify the definition of the *Attacks* relation. Given the conflation of attacks mentioned above we must ensure that the potential multiplicity of attacks is correctly dealt with when comparing strength.

For two paths $p, q \in P$ such that q Attacks p we define:

$$q \sqsupset_C p \equiv \forall x \in range(p). (last(q), x) \in Distrust \implies q \sqsupset p|_x$$

$$q \sqsupset_P p \equiv \forall x \in range(p). (last(q), x) \in Distrust \implies p|_x \not\sqsupset q$$

And given a partition of P into S and $\bar{S} (= P \setminus S)$:

$$\sqsupset_A^S \equiv ((S \times \bar{S}) \cap \sqsupset_C) \cup ((\bar{S} \times S) \cap \sqsupset_P)$$

Using these orderings we define the three corresponding attacks relations as:

- Conservative Preferential Trust: $Attacks_C = Attacks \cap \sqsupset_C$.
- Paranoid Preferential Trust: $Attacks_P = Attacks \cap \sqsupset_P$.
- Asymmetric Trust: $Attacks_A^S = Attacks \cap \sqsupset_A^S$.

Finally we demonstrate that the asymmetric trust systems have a pleasing simplification of the acceptability condition in that $Attacks_A^S$ factors into $Attacks_P$ and $Attacks_C$ on either side of the acceptability condition.

Proposition 5. $Attacks_A^S = ((S \times \bar{S}) \cap Attacks_C) \cup ((\bar{S} \times S) \cap Attacks_P)$

Proof Sketch. by boolean algebra

Proposition 6. An set, S , is acceptable in the asymmetric trust system iff $(Attacks_P)^{\exists} S \subseteq (Attacks_C)^{\exists} S$

Proof Sketch.

$$\begin{aligned}
& \widetilde{(Attacks_p^S)}^{\exists} S \subseteq Attacks_C^S{}^{\exists} S \\
& \quad \text{by proposition 2} \\
& (((S \times \bar{S}) \cap Attacks_C) \cup ((\bar{S} \times S) \cap Attacks_p)^{\exists}) S \subseteq \\
& \quad \quad \quad (((S \times \bar{S}) \cap Attacks_C) \cup ((\bar{S} \times S) \cap Attacks_p)^{\exists}) S \\
& \quad \text{by proposition 5} \\
& ((S \times \bar{S}) \cap Attacks_C)^{\exists} S \cup ((\bar{S} \times S) \cap Attacks_p)^{\exists} S \subseteq \\
& \quad \quad \quad ((S \times \bar{S}) \cap Attacks_C)^{\exists} S \cup ((\bar{S} \times S) \cap Attacks_p)^{\exists} S \\
& \quad \text{by distribute } (_)^{\exists} \text{ over union} \\
& \widetilde{(Attacks_p)}^{\exists} S \subseteq (Attacks_C)^{\exists} \\
& \quad \text{by domain restrictions}
\end{aligned}$$

8 Conclusions

For us at least, the idea of using argumentation to reason about networks of trust, and distrust, is in its infancy. The work presented here raises more questions than it answers, some of which we raise below (and there are many more than raised here).

Trust systems as outlined above offer a logically well founded approach to reasoning about trust based on minimal information gathered from individuals i.e. the individuals relative assessment of the efficacy of the judgements of others and a map of immediate trust and distrust relations between individuals. The natural next step is to investigate this in practice in an actual social network application.

The asymmetric preferential trust systems above rely on the fact that conservative attacks are a subset of paranoid attacks. Clearly it is possible to generalize further and define relevant attacks and acceptable rebuttals to relevant attacks. Given a set of attacks, we classify some attacks as relevant, some as acceptable rebuttals of relevant attacks, and some as neither. S is built as the closure of attacks on a trust path p as above and we define S^+ and S^- by:

- $p \in S^+$,
- S^+ is *consistent* in that no path in S^+ attacks any other path in S^+ ,
- S^+ is a *rebuttal set* in that every path in S^+ is either p or a rebuttal attack,
- S^+ *defends against relevant attacks* from S^- in that every relevant attack in S^- that attacks a path in S^+ is itself attacked by a path in S^+ .

This generalization opens up the possibility of considering richer asymmetries between trust and distrust arguments. For example, if we drop the use of the extended order relation and consider using a labeling of the individuals in paths. Consider, as illustration, a sensor network based on three kinds of individual sensor: electronic sensing and people that perform either casual or detailed inspections. We may trust an individual because we have a mixed

trust path to it but relevant attacks may be limited to chains that exclude electronic sensors and rebuttals may be limited to chains of people who perform detailed inspection⁷. This approach will be the subject of further investigation.

The relation to Dungian argumentation outlined in section 7 uses only the most basic semantic notion of admissibility. This raises the question whether or not the other possible semantics have a useful meaning for trust (and distrust) relations. The question is why we would *want* a richer set of arguments than that required to support the sounds of a particular trust path? Perhaps there is a useful notion of sets of individuals you can consistently trust corresponding to the other possible semantics. It certainly is worth investigating.

During the revision of this paper the authors encountered the work of Cayrol and Lagasque-Schiex on Bipolar Argumentation [5] systems, and of Kaci and Torre , and Amgoud, Dimopoulos and Moraitis Preference Based Argumentation (se e.g. [11] and [1] respectively). Both seem to overlap on the intent pursued here and offer interesting directions for future investigation.

9 Reviewer Acknowledgements

The authors would like to thank the reviewers for their constructive criticism, knowledgeable comments and insightful questions. In particular we would like to thank two of the reviews for their detailed comments and references to the work of other authors. This latter has been particularly useful to the ongoing work, even though not adequately reflected in this paper. In addition to minor corrections of the text the comments have been addressed with additional footnotes and the addition of Section 7 (in response to the reviewer plea for more mathematics).

10 Sponsorship Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

⁷ Admittedly, this example can be done using order relations, but it seems conceptual simpler as a predicate on the acceptable sets of attacks and counter attacks.

Bibliography

- [1] Leila Amgoud, Yannis Dimopoulos, and Pavlos Moraitis. Making decisions through preference-based argumentation. In Gerhard Brewka and Jérôme Lang, editors, *KR*, pages 113–123. AAAI Press, 2008.
- [2] Philippe Besnard and Sylvie Doutre. Characterization of semantics for argument systems. In Didier Dubois, Christopher A. Welty, and Mary-Anne Williams, editors, *KR*, pages 183–193. AAAI Press, 2004.
- [3] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, 2008.
- [4] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.
- [5] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In Lluís Godo, editor, *ECSQARU*, volume 3571 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2005.
- [6] James Coleman. *Foundations of Social theory*. Belknap Press, 1990.
- [7] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.
- [8] Russell Hardin, editor. *Trust & Distrust*. Russell Sage Foundation, 2004.
- [9] Russell Hardin. *Trust*. Polity Press, 2006.
- [10] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.
- [11] Souhila Kaci and Leendert van der Torre. Preference-based argumentation: Arguments supporting multiple values. *Int. J. Approx. Reasoning*, 48(3):730–751, 2008.
- [12] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, Dept. of Computing and Mathematics, University of Stirling, 1994.
- [13] Phil Zimmermann. *PGP User Guide*. MIT Press, 1994.