Arguing about Trust in Multiagent Systems

Serena Villata
University of Turin
villata@di.unito.it

Guido Boella
University of Turin
guido@di.unito.it

Dov M. Gabbay
King's College, London
dov.gabbay@kcl.ac.uk

Leendert van der Torre
University of Luxembourg
leendert@vandertorre.com

Abstract

Trust in multiagent systems is used for seeking to minimize the uncertainty in the interactions among the agents. In this paper, we discuss how to use argumentation to reason about trust. Using the methodology of meta-argumentation, first we represent the source of the information from which the argument is constructed in the abstract argumentation framework capturing the fact that *b is attacked* because b is from a particular source s. We show how a source of information can be attacked if it is not evaluated as trustworthy. Second, we provide a fine grained representation of the trust relationships between the information sources in which trust concerns not only the sources but also the single arguments and attack relations the sources propose. Moreover, we represent the evidences in support of the arguments which are put forward by the information sources and the agents can express arguments by referring to other agents' arguments. Meta-argumentation allows us not to extend Dung's abstract argumentation framework by introducing trust and to reuse those principles and properties defined for Dung's framework.

Introduction

Trust is a mechanism for managing uncertain information, decision making and dealing with the provenance of information. The result is that trust plays an important role in many research areas of computer science, particularly in the semantic web and multiagent systems where agents interact with other sources. In such interactions, the agents have to reason if they should trust or not the other agents and the extent to which they trust those other agents. The following illustrative example presents an informal argument exchange where several kinds of interactions between arguments and agents are reflected.

- *Witness1: I suspect the guy killed his boss in Rome. (arg a)*
- *Witness1: With a broken car he could not reach the crime scene. (arg b)*
- *Witness2: Witness1 is a compulsive liar. (arg c)*
- *Witness3: I repaired the guy's car at twelve of the crime day. (arg d)*
- *Witness4: I believe that Witness2 is not able to repair that kind of car. (arg e)*
- *Witness5: The guy has another car. (arg f)*
- *Witness6: The guy parked two cars in my underground parking garage three weeks ago. (arg g)*
- *Witness2: Witness6 was on holidays three weeks ago. (arg h)*
- *Witness7: The guy told he killed the boss. (arg i)*
- *Witness3: The guy charges himself to cover up for his wife. (arg l)*

In this informal argument exchange, different kinds of relations can be highlighted between arguments and agents. First, we have that the agents put forward the arguments and the attack relations. We will refer to these assertions by saying that the agents support their arguments and attack relations. Second, the agents can attack the trustworthiness of the other agents. These attacks are always addressed by means of arguments which attack the agent's trustworthiness itself or the trustworthiness of arguments and attack relations supported by this agent. Third, the agents can provide support to the other agents' arguments by putting forward evidences, always under the form of arguments, or by providing arguments which talk about other agents' arguments.

In this paper we argue that argumentation provides a mechanism to reason about trust handling aspects such as the origin of trust and the fine grained trust relationships. The research question addressed in the paper is:

- How to model trust in Dung's argumentation?

This breaks down into the following subquestions:

1. How to represent the information sources and the arguments they support?
2. How to represent an attack to the trustworthiness of the sources of information and a fine grained view of trust relations where trust concerns also single arguments and attacks?
3. How to represent the evidences provided in support of the arguments?
4. How to model trust when the agents express arguments concerning other agents' arguments?

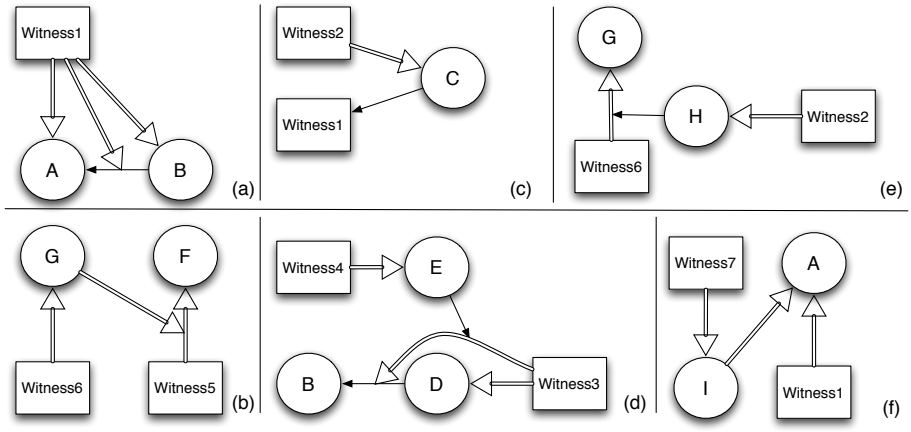

Figure 1: The patterns involving agents and arguments and the trust relationships.

To answer the research questions we propose the methodology of meta-argumentation (Boella, van der Torre, and Villata 2009; Boella et al. 2009). The advantage in using meta-argumentation is that we do not extend Dung's framework (Dung 1995) in order to introduce trust but we instantiate Dung's theory with meta-arguments. In this way we can reuse all the principles, algorithms and properties already defined for Dung's framework. In meta-argumentation, different entities besides proper arguments are introduced in the meta-level under the form of meta-arguments and the acceptable, *trusted*, meta-arguments are returned. These meta-arguments represent the arguments of the agents and their attack relations. The agents, as sources of arguments and attack relations, are introduced under the form of meta-arguments "*agent i is trustable*".

The research questions ask for patterns where both agents and arguments are composed together and are related to each other by trust relationships. The patterns which emerge from the informal argument exchange are provided in Figure 1 where the arrows represent the attack relation and the double arrows represent the support relation of the agents to the arguments they built. In Figure 1.a, the representation of the information sources and the arguments they support is provided. Witness5 supports both arguments a, b and the attack relation between them. In Figure 1.b, the evidence provided by Witness6 in support to the argument of Witness5 is represented. Figure 1.c depicts the attack of Witness2 to the trustworthiness of Witness1. Note that, this agent becomes no more credible in the multiagent system because her credibility has been attacked as a whole. This is not always the case, it may be possible that the agents attack other agents' trustworthiness only concerning a particular argument or attack relation. This is described in Figure 1.d-e where Witness4 and Witness2 attack the trustworthiness of Witness3 and Witness6 respectively only concerning argument g and the attack relation $d \rightarrow b$. Finally, arguments about other agents' arguments are represented in Figure 1.f where Witness7 supports by means of his argument i Witness1's argument a.

The paper follows the research questions. After a brief introduction on the methodology of meta-argumentation, we describe how to represent the agents in an argumentation framework and we discuss how to model the patterns defined in Figure 1. Related work and conclusions end the paper.

Meta-Argumentation

Dung's theory (Dung 1995) is based on a binary *attack* relation among arguments, which are abstract entities whose role is determined only by their relation to other arguments. We restrict ourselves to *finite* argumentation frameworks, i.e., in which the set of arguments is *finite*.

Definition 1 (Argumentation framework AF) *An argumentation framework is a tuple $\langle A, \rightarrow \rangle$ where A is a finite set of elements called arguments and $\rightarrow$ is a binary relation called attack defined on $A \times A$.*

A semantics of an argumentation framework consists of a conflict-free set of arguments, i.e., a set of arguments that does not contain an argument attacking another argument in the set.

Definition 2 (Conflict-free) *Given an argumentation framework $AF = \langle A, \rightarrow \rangle$, a set $S \subseteq A$ is* conflict free*, denoted as cf(S), iff $\nexists \alpha, \beta \in S$ such that $\alpha \rightarrow \beta$.*

Like (Baroni and Giacomin 2007) we use a function $\mathcal{E}$ mapping an argumentation framework $\langle A, \rightarrow \rangle$ to its set of extensions, i.e., to a set of sets of arguments. Since Baroni and Giacomin do not give a name to the function $\mathcal{E}$, and it maps argumentation frameworks to the set of accepted arguments, we call $\mathcal{E}$ the *acceptance function*.

Definition 3 *Let $\mathcal{U}$ be the universe of arguments. An acceptance function $\mathcal{E} : 2^{\mathcal{U}} \times 2^{\mathcal{U} \times \mathcal{U}} \rightarrow 2^{2^{\mathcal{U}}}$ is a partial function which is defined for each argumentation framework $\langle A, \rightarrow \rangle$ with finite $A \subseteq \mathcal{U}$ and $\rightarrow \subseteq A \times A$, and maps an argumentation framework $\langle A, \rightarrow \rangle$ to sets of subsets of A: $\mathcal{E}(\langle A, \rightarrow \rangle) \subseteq 2^A$.*

The following definition summarizes the most widely used acceptability semantics of arguments given in the lit-

erature. Which semantics is most appropriate in which circumstances depends on the application domain of the argumentation theory.

Definition 4 (Acceptability semantics) *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework. Let $\mathcal{S} \subseteq A$. $\mathcal{S}$ defends a if $\forall b \in A$ such that $b \rightarrow a$, $\exists c \in \mathcal{S}$ such that $c \rightarrow b$. Let $D(\mathcal{S}) = \{a \mid \mathcal{S}$ defends $a\}$.*

- $\mathcal{S} \in \mathcal{E}_{admiss}(AF)$ *iff cf*$(\mathcal{S})$ *and* $\mathcal{S} \subseteq D(\mathcal{S})$.
- $\mathcal{S} \in \mathcal{E}_{compl}(AF)$ *iff cf*$(\mathcal{S})$ *and* $\mathcal{S} = D(\mathcal{S})$.
- $\mathcal{S} \in \mathcal{E}_{ground}(AF)$ *iff* $\mathcal{S}$ *is smallest in* $\mathcal{E}_{compl}(AF)$.
- $\mathcal{S} \in \mathcal{E}_{pref}(AF)$ *iff* $\mathcal{S}$ *is maximal in* $\mathcal{E}_{admiss}(AF)$.
- $\mathcal{S} \in \mathcal{E}_{skep\text{-}pref}(AF)$ *iff* $\mathcal{S} = \cap\mathcal{E}_{pref}(AF)$.
- $\mathcal{S} \in \mathcal{E}_{stable}(AF)$ *iff cf*$(\mathcal{S})$ *and* $\forall b \in A \backslash \mathcal{S}$ $\exists a \in \mathcal{S} : a \rightarrow b$.

We (Boella et al. 2009) instantiate Dung's theory with meta-arguments, *such that we use Dung's theory to reason about itself*. Meta-argumentation is a particular way to define mappings from argumentation frameworks to extended argumentation frameworks: arguments are interpreted as meta-arguments, of which some are mapped to "argument a is accepted", $acc(a)$, where a is an abstract argument from the extended argumentation framework EAF. The meta-argumentation methodology is summarized in Figure 2.

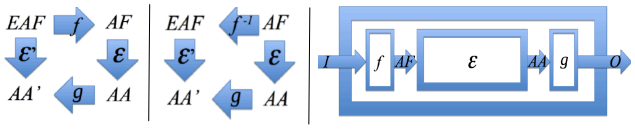

Figure 2: The meta-argumentation methodology.

The function f assigns to each argument a in the EAF, a meta-argument "argument a is accepted" in the basic argumentation framework. We use Dung's acceptance functions $\mathcal{E}$ to find functions $\mathcal{E}'$ between extended argumentation frameworks EAF and the acceptable arguments AA' they return. The accepted arguments of the meta-argumentation framework are a function of the extended argumentation framework $AA = \mathcal{E}'(EAF)$. The transformation function consists of two parts: a function f^{-1} transforms an argumentation framework AF to an extended argumentation framework EAF, and a function g transforms the acceptable arguments of the AF into acceptable arguments of the EAF. Summarizing $\mathcal{E}' = \{(f^{-1}(a), g(b)) \mid (a, b) \in \mathcal{E}\}$ and $AA' = \mathcal{E}'(EAF) = g(AA) = g(\mathcal{E}(AF)) = g(\mathcal{E}(f(EAF)))$.

The first step of our approach is to define the set of extended argumentation frameworks. The second step consists in defining flattening algorithms as a function from this set of EAFs to the set of all basic argumentation frameworks: $f : EAF \rightarrow AF$.

Definition 5 presents the instantiation of a basic argumentation framework as a sequence of partial argumentation frameworks of the agents (Coste-Marquis et al. 2007) using meta-argumentation. A sequence of partial argumentation frameworks of the agents $\langle\langle A_1, \rightarrow_1\rangle, \ldots, \langle A_n, \rightarrow_n\rangle\rangle$ are sets composed by arguments A_i and a binary attack relation $\rightarrow_i$.

The universe of meta-arguments is $MU = \{acc(a) \mid a \in \mathcal{U}\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in \mathcal{U}\}$, and the flattening function f is given by $f(EAF) = \langle MA, \longmapsto\rangle$. For a set of arguments $B \subseteq MU$, the unflattening function g is given by $g(B) = \{a \mid acc(a) \in B\}$, and for sets of arguments $AA \subseteq 2^{MU}$, it is given by $g(AA) = \{g(B) \mid B \in AA\}$.

Definition 5 *Given an extended argumentation framework $EAF = \langle\langle A_1, \rightarrow_1\rangle, \ldots, \langle A_n, \rightarrow_n\rangle\rangle$ where for each agent $1 \leq i \leq n$, $A_i \subseteq \mathcal{U}$ is a set of arguments and $\rightarrow_i \subseteq A_i \times A_i$ is a binary relation over A_i, the set of meta-arguments $MA \subseteq MU$ is $\{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that: $acc(a) \longmapsto X_{a,b}, X_{a,b} \longmapsto Y_{a,b}, Y_{a,b} \longmapsto acc(b)$ if and only if there is an agent $1 \leq i \leq n$ such that $a, b \in A_i$ and $a \rightarrow_i b$.*

The set of acceptable arguments of a meta-argumentation framework $\langle MA, \longmapsto\rangle$ follows from $\mathcal{E}'(EAF) = g(\mathcal{E}(f(EAF)))$. For a given flattening function f, the acceptance function of the extended argumentation theory $\mathcal{E}'$ is defined using the acceptance function of the basic abstract argumentation theory $\mathcal{E}$: an argument of an EAF is acceptable *iff* it is acceptable in the flattened AF.

Modelling trust in Dung's framework

A number of authors have highlighted that the definition of trust is difficult to pin down precisely, thus in the literature there are numerous different definitions. To pick few of these definitions, (Castelfranchi and Falcone 2001) define trust as "*a mental state, a complex attitude of an agent x towards another agent y about the behaviour/action a relevant for the goal g*" while (Gambetta 1990) states that "*trust is the subjective probability by which an individual A expects that another individual B performs a given action on which its welfare depends*". The common elements are that there is a consistent degree of uncertainty associated with trust and trust is tied up with the relationships between individuals and particularly it is related to the actions of the individuals and to the effects these actions have on the others. Another approach to model trust using modal logic is proposed by (Lorini and Demolombe 2008) where they present a concept of trust that integrates the trusters goal, the trustees action ensuring the achievement of the trusters goal, and the trustees ability and intention to do this action. In this paper we does not refer to the actions of the agents and their goals but we provide a model for representing the agents' beliefs concerning the trustworthiness of the other agents. We follow the approach proposed by (Liau 2003) where the influence of trust on the assimilation of acquired information into an agent's belief is considered. (Liau 2003)'s characteristic axiom is "if agent i believes that agent j has told him the truth of p and he trusts the judgement of j on p, then he will also believe p".

Representing the information sources

Let us consider again the informal argument exchange. We have that *Witness2* has a negative opinion about the trustworthiness of *Witness1* while we can infer that all the other witnesses consider *Witness1* a reliable information source since they do not attack him. Agents are introduced in the meta-argumentation framework under the form of meta-arguments "*agent i is trustable*", $trust(i)$, for all the agents i. As in Definition 5, we add meta-arguments "*argument a is accepted*", $acc(a)$, for all arguments in A, and meta-arguments $X_{a,b}$, $Y_{a,b}$ for all arguments a and b such that $a \rightarrow b$. Each argument $a \in A$ in the mind of the agents is put forward, by means of the $Z_{acc(a)}$ meta-argument. This meta-argument attacks the meta-argument $acc(a)$ asking for an evidence in support of argument a. In the simplest case, the $Z_{acc(a)}$ meta-argument is attacked by the meta-argument $trust(i)$ which represents the agent who proposes argument a, as shown in Figure 3. More complex cases of evidences are described in the next sections. For each agent i, if $\rightarrow_i$ contains $a \rightarrow b$, such as if the agent puts forward an attack relation, then the meta-argument $trust(i)$ supports the meta-argument $Y_{a,b}$, representing the attack relation, by attacking the meta-argument $Z_{Y_{a,b}}$, as concerning the arguments. Also in this case, the attack of the agent to the Z meta-argument is an evidence in support of this attack relation.

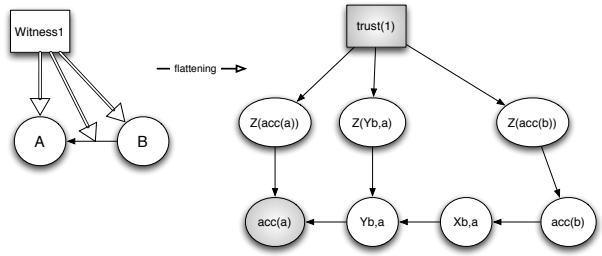

Figure 3: Introducing the agents in the framework.

We represent the fact that more than one information source sustains the same arguments by let them attacking by means of the $trust(i)$ meta-arguments the same $Z_{acc(a)}$ meta-argument which asks for evidences in support of meta-argument $acc(a)$. An example of multiple support of two agents regarding the same argument is depicted in Figure 4. The same solution is applied to the attack relations where we consider the meta-argument $Y_{a,b}$ instead of meta-argument $acc(a)$.

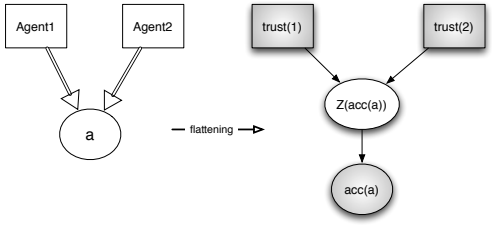

Figure 4: A multiple support to the same argument.

We extend the EAF proposed in Definition 5 by adding the information sources and second-order attacks, such as attacks from an argument or attack relation to another attack relation. For more details about second-order attacks in meta-argumentation, see (Boella et al. 2009).

The unflattening function g and the acceptance function $\mathcal{E}'$ are defined as above. In particular, the introduction of the agents in the meta-argumentation framework is defined as follows:

Definition 6 *An extended argumentation framework EAF is a tuple $\langle\langle A_1, \rightarrow_1, \rightarrow_1^2\rangle \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2\rangle\rangle$ where for each agent $1 \leq i \leq n$, $A_i \subseteq \mathcal{U}$ is a set of arguments, $\rightarrow_i$ is a binary relation on $A_i \times A_i$, $\rightarrow_i^2$ is a binary relation on $(A_i \cup \rightarrow_i) \times \rightarrow_i$.*

Definition 7 *Given an extended argumentation framework $EAF = \langle\langle A_1, \rightarrow_1, \rightarrow_1^2\rangle \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2\rangle\rangle$, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{Z_a \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:*

- $acc(a) \longmapsto X_{a,b}, X_{a,b} \longmapsto Y_{a,b}, Y_{a,b} \longmapsto acc(b)$ *iff* $a, b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $trust(i) \longmapsto Z_{acc(a)}, Z_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A_i$, *and*
- $trust(i) \longmapsto Z_{Y_{a,b}}, Z_{Y_{a,b}} \longmapsto Y_{a,b}$ *iff* $a, b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $acc(a) \longmapsto X_{a,b\rightarrow c}, X_{a,b\rightarrow c} \longmapsto Y_{a,b\rightarrow c}, Y_{a,b\rightarrow c} \longmapsto Y_{b,c}$ *iff* $a, b, c \in A_i$ *and* $a \rightarrow_i^2 (b \rightarrow_i c)$,
- $Y_{a,b} \longmapsto Y_{c,d}$ *iff* $a, b, c \in A_i$ *and* $(a \rightarrow_i b) \rightarrow_i^2 (c \rightarrow_i d)$.

Example 1 *We represent the agents in the argumentation framework as shown in Figure 3. Witness1 puts forward two arguments a and b and the attack relation between them. Using the flattening function described in Definition 7, we add the meta-argument $trust(1)$ for representing Witness1 in the framework and we add meta-arguments $acc(a)$ and $acc(b)$ for the arguments of Witness1. The attack relation is represented by means of two meta-arguments $X_{a,b}$ and $Y_{a,b}$ which stay for the inactive and active status of the attack relation $a \rightarrow b$. Witness1 provides an evidence in support to the arguments a and b and the attack relation $a \rightarrow b$ by attacking the respective meta-arguments Z.*

Representing fine grained trust relationships

In our model, trust is represented as the absence of an attack towards the agents or towards their arguments and attack relations or as the presence of an evidence in support of arguments and attack relations. On the contrary, the lack of trust, here called distrust, is modeled as a lack of evidences in support of the arguments and the attack relations or as an attack relation towards the agents and their arguments and attack relations. The three distrust relationships depicted in Figure 1.c-d-e are of different kind and must be distinguished in the framework in order to reason about trust. Notice that the notion of distrust can be associated to a different meaning from lack of trust or insufficient trust such as diffidence towards a source s. In this case, the argumentation is precisely aimed at creating distrust. Modeling this kind of distrust is left for future work.

In the informal argument exchange, Witness2 attacks the trustworthiness of Witness1 as a credible witness. In this way, he is attacking each argument and attack relation proposed by Witness1. Witness4, instead, is not arguing against Witness3 but he is arguing against the attack relation $d \rightarrow b$ as proposed by Witness3. Finally, Witness2 reasons about the trustworthiness of Witness6. The untrustworthiness of Witness6 is linked only to the precise argument g. Our proposal is a fine grained view of trust in which the sources of information may be attacked for being unreliable or for being unreliable in sustaining a particular argument or attack relation. Definition 8 presents an extended argumentation framework in which a new relation between arguments is given to represent distrust.

Definition 8 *A trust-based extended argumentation framework $TEAF$ is a tuple $\langle\langle A_1, \rightarrow_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, DT_n\rangle\rangle$ where for each agent $1 \leq i \leq n$, $A_i \subseteq \mathcal{U}$ is a set of arguments, $\rightarrow_i \subseteq A_i \times A_i$ is a binary relation and $DT \subseteq A_i \times \vartheta$ is a binary relation such that $\vartheta \in j$ or $\vartheta \in A_j$ or $\vartheta \in \rightarrow_j$.*

The extended argumentation framework $TEAF$ would need new semantics in order to compute what are the accepted arguments. In alternative, we use the meta-argumentation methodology to flatten the $TEAF$ to a meta-argumentation framework where classical Dung's semantics are used to compute the set of acceptable arguments. We define the meta-argumentation framework in the following way where the unflattening function g and the acceptance function $\mathcal{E}'$ are defined as above.

Definition 9 *Given a trust-based extended argumentation framework $TEAF = \langle\langle A_1, \rightarrow_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, DT_n\rangle\rangle$, see Definition 8, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{Z_a \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:*

- $acc(a) \longmapsto X_{a,b}, X_{a,b} \longmapsto Y_{a,b}, Y_{a,b} \longmapsto accept(b)$ *iff* $a, b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $trust(i) \longmapsto X_{trust(i),Z_{acc(a)}}, X_{trust(i),Z_{acc(a)}} \longmapsto Y_{trust(i),Z_{acc(a)}}, Y_{trust(i),Z_{acc(a)}} \longmapsto Z_{acc(a)}, Z_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A_i$, *and*
- $trust(i) \longmapsto X_{trust(i),Z_{Y_{a,b}}}, X_{trust(i),Z_{Y_{a,b}}} \longmapsto Y_{trust(i),Z_{Y_{a,b}}}, Y_{trust(i),Z_{Y_{a,b}}} \longmapsto Z_{Y_{a,b}}, Z_{Y_{a,b}} \longmapsto Y_{a,b}$ *iff* $a, b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $trust(i) \longmapsto Z_{acc(a)}, Z_{acc(a)} \longmapsto acc(a), acc(a) \longmapsto X_{acc(a),trust(j)}, X_{acc(a),trust(j)} \longmapsto Y_{acc(a),trust(j)}, Y_{acc(a),trust(j)} \longmapsto trust(j)$ *iff* $a \in A_i$ *and* $aDT_i trust(j)$, *and*
- $trust(i) \longmapsto Z_{acc(a)}, Z_{acc(a)} \longmapsto acc(a), acc(a) \longmapsto X_{acc(a),Y_{trust(j),Z_{acc(b)}}}, X_{acc(a),Y_{trust(j),Z_{acc(b)}}} \longmapsto Y_{acc(a),Y_{trust(j),Z_{acc(b)}}}, Y_{acc(a),Y_{trust(j),Z_{acc(b)}}} \longmapsto Y_{trust(j),Z_{acc(b)}}$ *iff* $a \in A_i, b \in A_j$ *and* $aDT_i b$, *and*
- $trust(i) \longmapsto Z_{acc(a)}, Z_{acc(a)} \longmapsto acc(a), acc(a) \longmapsto X_{acc(a),Y_{trust(j),Z_{Y_{b,c}}}}, X_{acc(a),Y_{trust(j),Z_{Y_{b,c}}}} \longmapsto Y_{acc(a),Y_{trust(j),Z_{Y_{b,c}}}}, Y_{acc(a),Y_{trust(j),Z_{Y_{b,c}}}} \longmapsto Y_{trust(j),Z_{Y_{b,c}}}$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$.

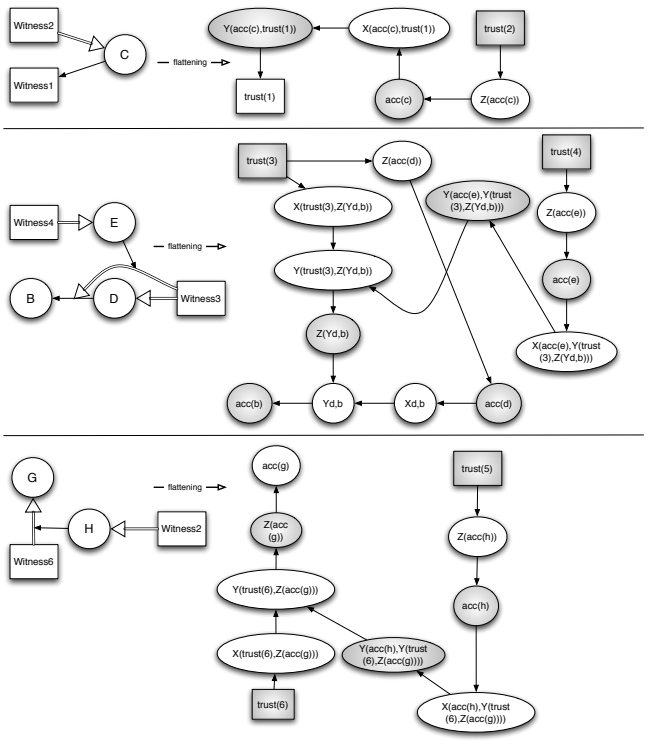

Figure 5: Fine grained trust in argumentation.

Definition 9 shows how to instantiate an extended argumentation framework composed by a set of arguments, a binary attack relation and a binary distrust relation with meta-arguments. In particular, the last three points model respectively a distrust relationship towards an agent, towards an argument and towards an attack relation.

Example 2 *In Figure 5 we highlight the three patterns where trust relations between information sources are represented. The first pattern shows that Witness2 attacks the trustworthiness of Witness1 with the argument c. In meta-argumentation, we have that $trust(2)$ proposes $acc(c)$ by attacking meta-argument $Z_{acc(c)}$ and, with meta-arguments X, Y, it attacks $trust(1)$. This means that if Witness1 is not reliable then each of his arguments and attack relations cannot be acceptable either. If we look at the extension of this pattern, we have that the set of acceptable arguments for the meta-argumentation framework is $\mathcal{E}(f(pattern1)) = \{trust(2), acc(c), Y_{acc(c),trust(1)}\}$. Consider the other two patterns of Figure 5. On the one hand, Witness4 attacks the attack relation $d \rightarrow b$ proposed by Witness3. This is achieved in meta-argumentation by an attack from meta-argument $acc(e)$, proposed by $trust(4)$, to the attack relation characterized by meta-argument $Y_{d,b}$. The set of acceptable arguments is $\mathcal{E}(f(pattern2)) = \{trust(4), trust(3), acc(d), acc(e), acc(b), Y_{acc(e),Y_{trust(3),Z_{Y_{b,d}}}}, Z_{Y_{d,b}}\}$. Witness3's attack relation $d \rightarrow b$ is evaluated as not reliable for Witness4 and it is not acceptable. On the other hand, Witness2 evaluates unreliable Witness6 concerning argument g. In meta-argumentation,*

$trust(2)$, *by means of meta-argument* $acc(h)$, *attacks meta-argument* $acc(g)$ *proposed by* $trust(6)$. *In this case, the set of acceptable arguments is* $\mathcal{E}(f(pattern3)) = \{trust(2), trust(6), acc(h), Y_{acc(h),Y_{trust(6),Z_{acc(g)}}}, Z_{acc(g)}\}$.

Representing the evidences supporting arguments

The evidences in favor of the arguments are represented, as discussed before, as a support given by the agents to the arguments at the object level. At the meta-level, this is modeled as an attack relation from meta-argument $trust(i)$ to the Z meta-arguments. However, there are also cases in which evidences are necessary to support the acceptability of an argument. Let consider the case in which the trustworthiness of an agent is attacked. What does it happen to the arguments put forward by this agent? They become not acceptable. In this case, what is needed to reinstate the acceptability of these arguments is an evidence. This evidence is provided under the form of an argument put forward by another agent.

Definition 7 presents how to instantiate an extended argumentation framework composed by a set of arguments, a binary attack relation, a binary second-order attack relation representing also the information sources. Definition 9, instead, extends Dung's framework with a distrust relation DT. In order to have an extended argumentation framework with both the relations of the EAF of Definition 7 and the $TEAF$ of Definition 9, we define an extended trust-based argumentation framework with the addition of an evidence relation $\not\rightarrow$ which represents the evidences provided in favor of the arguments of the other agents.

Definition 10 *A trust-based argumentation framework with evidences* $TEAF^2 = \langle\langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, T_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, T_n\rangle\rangle$, *see Definition 7 and Definition 9, where* $\not\rightarrow_i$ *is a binary relation on* $A_i \times A_j$ *and the set of meta-arguments* MA *is* $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{Z_a \mid a \in A_1 \cup \ldots \cup A_n\}$ *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on* MA *such that the conditions of Definition 7 and Definition 9 hold, and:*

- $acc(a) \longmapsto Z_{acc(b)}, Z_{acc(b)} \longmapsto acc(b)$ *iff* $a, b \in A_i$ *and* $a \not\rightarrow_i b$.

Example 3 *We have that argument* f *is "The guy has another car" while argument* g *by Witness6 is " The guy parked two cars in my underground parking garage three weeks ago". Argument* g *is an evidence in favor of* f. *This evidence is expressed in meta-argumentation as an attack from meta-argument* $acc(g)$ *to meta-argument* $Z_{acc(f)}$ *attacking* $acc(f)$. *This example is described in Figure 6.*

Representing arguments about other agents' arguments

The information sources may also express arguments concerning other agents' arguments as in the case of arguments i and a during the informal argument exchange. In this case we have that Witness7 proposes an argument which is based on the argument of another agent, Witness1. Moreover, we have that Witness3 introduces argument l which attacks the

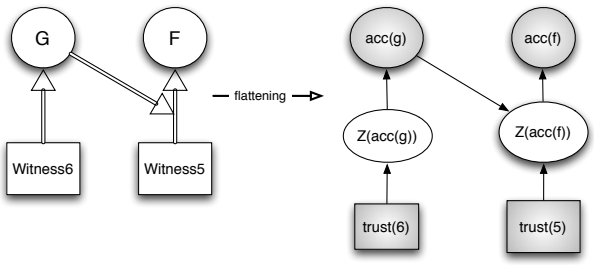

Figure 6: Introducing evidences in favor of the arguments.

support of argument i to argument a. If we add to the extended trust-based argumentation framework $TEAF^2$ a new binary relation $\dashrightarrow \subseteq A_i \times A_i$, we can model the report of other agents' arguments in the following way:

Definition 11 *Given an extended trust-based argumentation framework* $TEAF^2 = \langle\langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, T_1, \dashrightarrow_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, T_n, \dashrightarrow_n\rangle\rangle$, *the set of meta-arguments* MA *is* $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{Z_a \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{dixit_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\}$ *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on* MA *such that the conditions of Definition 7, Definition 9 and Definition 10 hold and:*

- $acc(a) \longmapsto Z_{dixit_{a,b}}, Z_{dixit_{a,b}} \longmapsto dixit_{a,b}, dixit_{a,b} \longmapsto Z_{acc(b)}, Z_{acc(b)} \longmapsto acc(b)$ *iff* $a, b \in A_i$ *and* $a \dashrightarrow_i b$ *and*
- $acc(a) \longmapsto X_{acc(a),dixit_{b,c}}, X_{acc(a),dixit_{b,c}} \longmapsto Y_{acc(a),dixit_{b,c}}, Y_{acc(a),dixit_{b,c}} \longmapsto dixit(b, c)$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \dashrightarrow_i c)$.

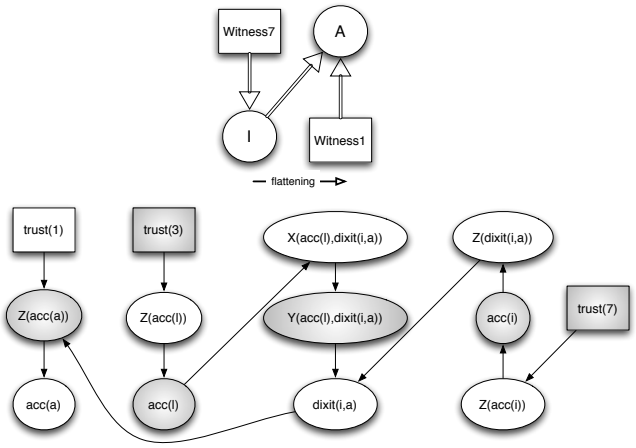

Figure 7: Introducing arguments about other agents' arguments.

Example 4 *Let* $TEAF^2$ *be defined as* $A_1 = \{a, b\}$, $\rightarrow_1 = \{(b, a)\}$, $A_2 = \{c, h\}$, $DT_2 = \{(c, 1), (h, g)\}$, $A_3 = \{b, d, l\}$, $\rightarrow_3 = \{(d, b)\}$, $DT_3 = \{(l, \dashrightarrow_7 (i, a))\}$, $A_4 = \{e\}$, $DT_4 = \{(e, \rightarrow_3 (d, b))\}$, $A_5 = \{a, b, f\}$, $\rightarrow_5^2 =$

$\{(f,(b,a))\}$, $A_6 = \{a,g\}$, $\not\rightarrow_6 = \{(g,f)\}$, $A_7 = \{i\}$, $\dashrightarrow_7 = \{(i,a)\}$. *This extended argumentation framework is the model of the informal argument exchange proposed in the introduction. In Figure 7, we introduce arguments* i *and* l *under the form of meta-arguments. We have that argument* i *reports an argument from the guy which sustains argument* a *proposed by* $trust(1)$. $trust(7)$ *proposes meta-argument* $acc(i)$ *and a dixit relation* $\dashrightarrow_7 = \{(i,a)\}$ *holds between arguments* i *and* a. *This is represented in meta-argumentation by the new meta-argument* $dixit_{i,a}$ *which is sustained, by means of meta-argument* $Z_{dixit_{i,a}}$, *by* $acc(i)$. *Witness3 does not agree about the link between arguments* i *and* a *because in his opinion the guy admitted to kill the boss only to protect his wife (argument* l*). This attack has the aim to eliminate the support which argument* i *gives to argument* a. *We represent it in meta-argumentation by adding an attack relation from meta-argument* $acc(l)$ *claimed by* $trust(3)$ *to meta-argument* $dixit_{i,a}$. *In this way the attack of* $dixit(i,a)$ *to* $Z_{acc(a)}$ *is made ineffective and meta-argument* $acc(a)$ *would be acceptable only if sustained by some other meta-argument. The set of acceptable arguments of the* $TEAF^2$ *is* $\mathcal{E}'(TEAF^2) = g(\mathcal{E}(f(TEAF^2))) = \{b,c,d,e,f,h,i,l\}$ *and the trustable information sources are all the witnesses except Witness1 who has been directly attacked by Witness2.*

Related work

(Dix et al. 2009) present trust as a major issue in MAS applications concerning the research challenges for argumentation. *Which agents are trustworthy?* This is important for taking decisions and weighing arguments of other agents.

(Parsons, McBurney, and Sklar 2010) discuss why argumentation has an important role to play in reasoning about trust and highlight what are the mechanisms which need to be investigated through argumentation. The authors claim that a first problem, particularly of abstract approaches such that of (Dung 1995), is that they cannot express the provenance of trust and they cannot express the fact that b is attacked because b is proposed by agent s and there is an evidence that s is not trustworthy. In this paper, we propose a methodology which allows us to instantiate Dung's framework with meta-arguments which represent the information sources. Moreover, we show how to express trust relationships between the sources. Another problem highlighted by (Parsons, McBurney, and Sklar 2010) is the explicit expression of degrees of trust, as adopted by the prevalence of numerical measures of trust in the literature. In this paper, we do not present an explicit expression of degrees of trust and this is a topic for further research but we present a model where a fine grained view of trust relationships is provided.

(Matt, Morge, and Toni 2010) propose an extension to the Dempster-Shafer belief function. The authors allow the evaluator agent to take into account, in addition to the statistical data, a set of justified claims concerning the expected behaviour of the target agent. These claims form the basis of the evaluator's opinions and are formally represented by arguments in abstract argumentation. Two kinds of arguments are defined: forecast arguments and mitigation arguments. Forecast arguments express the trustworthiness or untrustworthiness of the target agent and mitigation arguments attack forecast arguments or other mitigation arguments because of the uncertainties of the validity of forecast arguments. Dempster-Shafer belief function is constructed both from statistical data and from these arguments. Arguments are generated by contracts. We propose an approach to the introduction of trust which is more related to modelling and no statistical problems are addressed. We show how to introduce the provenance of the information in the argumentation framework. (Matt, Morge, and Toni 2010) have focused on the computation of trust by an evaluator of a target in isolation. We propose a model in which all the trust relationships are evaluated together and we do not restrict our model to the contracts.

(Stranders, de Weerdt, and Witteveen 2007) propose an approach to trust based on argumentation in which there is a separation between the opponent modeling and decision making. The opponents' behaviour is modeled using possibilistic logic. The paper shows the results based on the ART testbed. In our approach we use Dung's abstract framework and we do not present a decision making approach to trust. We are interested in modeling fine grained trust in argumentation and we do not present experimental results.

(Prade 2007) presents a bipolar qualitative argumentative modeling of trust where a finite number of levels is assumed in a trust scale and trust and distrust are assessed independently. The author introduces the notion of reputation which is viewed as an input information used by an agent for revising or updating his trust evaluation. Reputation contributes also to provide direct arguments in favour or against a trust evaluation. There are a number of differences between (Prade 2007) and our approach. First, we does not apply a diagnostic point of view as in (Prade 2007) but we are interested in a social multiagent perspective. Second, we use a Dung's based approach while in (Prade 2007) arguments have an abductive form. Graded trust is presented also by (Lorini and Demolombe 2008) where they move beyond binary trust (i.e. either agent i trusts agent j or i does not trust j) in order to capture a concept of graded trust.

An approach related to trust in argumentation is provided by (Hunter 2008) where the author introduces a logic-based meta-level argumentation framework for evaluating arguments in terms of the appropriateness of their proponents. A further investigation of the relation between trust evaluation and proponents' appropriateness is an interesting direction for future research.

Conclusions

In this paper, a way to model trust in Dung's framework is presented. We answer the research questions using the methodology of meta-argumentation where Dung's framework is used to reason about itself. Meta-argumentation (Boella, van der Torre, and Villata 2009; Boella et al. 2009) allows us to introduce the notion of trust without extending Dung's standard argumentation framework and by reusing Dung's semantics and properties.

We represent the sources of information in the abstract argumentation framework in order to link the agents to the arguments they construct. We introduce the agents as meta-arguments of the kind "agent i is trustable", $trust(i)$, and

each agent is linked to the arguments he proposes by means of meta-arguments Z_x. Meta-arguments $trust(i)$ attack meta-arguments Z_x when x is an argument or an attack relation put forward by agent i. For each agent who sustains argument/attack x, there is an attack from $trust(i)$ to Z_x. In this way, the argumentation framework keeps track of the provenance of the arguments and attack relations and it allows us to represent evidences in the framework. More than one agent can support the same argument. This is expressed in meta-argumentation in the following way. If the agents support directly an argument or attack x, then they both attack meta-argument Z_x. If the agents propose new arguments which sustain other arguments then this is expressed by an attack from meta-argument $acc(a)$ to $Z_{acc(b)}$, where argument a is an evidence of argument b.

The trustworthiness of the agents can be attacked by attacking meta-arguments $trust(i)$ representing the agents in the argumentation framework. The agents, supporting arguments against the trustworthiness of the other agents, attack the reliability of the other agents. Trust is represented as an absence of attacks on the agents' trustworthiness. The agents who are not evaluated as reliable in the framework are those whose meta-argument $trust(i)$ is not in the extension of the meta-argumentation framework. We present a fine grained view of trust relationships. The agents can express their evaluation on other agents' reliability also concerning single arguments and attack relations proposed by the unreliable agents. We express the evaluation of the untrustworthiness of arguments and attacks by means of attacks to the Y_{Z_x} meta-argument which is used by meta-argument $trust(i)$ to attack meta-argument Z_x.

If the arguments or attack relations evaluated unreliable are not supported by other evidences, such as arguments which attack the Z_x meta-argument, then they are made unacceptable in the extension of the meta-argumentation framework.

Agents can express reported arguments about arguments expressed by other agents. This is represented in meta-argumentation with meta-argument $dixit_{a,b}$ where argument a reports what is expressed by argument b. Meta-argument $acc(a)$ supports, by means of meta-argument $Z_{dixit(a,b)}$, meta-argument $dixit_{a,b}$ which supports $acc(b)$. In this way, the support of the $dixit$ meta-argument can be attacked by other arguments if the agents believe it to be an unreliable information.

Future research is addressed following different lines. First, we are studying how to express trust revision. As highlighted also by (Parsons, McBurney, and Sklar 2010), an important aspect in reasoning about trust is the need for a source to be able to revise the trust she has in another source based on experience. Second, there is a bidirectional link between the source and its input: the provided data is more or less believable on the basis of the source trustworthiness but there is a feedback such that the invalidation of the data feedbacks on the sources credibility (Castelfranchi and Falcone 2001). The investigation of this relation is left for future research. Third, we plan to investigate two dimensions of trust that have to be independently evaluated such as the sincerity/credibility of a source and the competence of a source.

References

Baroni, P., and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *Artif. Intell.* 171(10-15):675–700.

Boella, G.; Gabbay, D. M.; van der Torre, L.; and Villata, S. 2009. Meta-argumentation modelling i: Methodology and techniques. *Studia Logica* 93(2-3):297–355.

Boella, G.; van der Torre, L.; and Villata, S. 2009. On the acceptability of meta-arguments. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2009*, 259–262. IEEE.

Castelfranchi, C., and Falcone, R. 2001. Social trust: A cognitive approach. *Trust and Deception in Virtual Societies* 55–90.

Coste-Marquis, S.; Devred, C.; Konieczny, S.; Lagasquie-Schiex, M.-C.; and Marquis, P. 2007. On the merging of dung's argumentation systems. *Artif. Intell.* 171(10-15):730–753.

Dix, J.; Parsons, S.; Prakken, H.; and Simari, G. R. 2009. Research challenges for argumentation. *Computer Science - R&D* 23(1):27–34.

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

Gambetta, D. 1990. Can we trust them? *Trust: Making and breaking cooperative relations* 213–238.

Hunter, A. 2008. Reasoning about the appropriateness of proponents for arguments. In Fox, D., and Gomes, C. P., eds., *Proceedings of Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, 89–94. AAAI Press.

Liau, C.-J. 2003. Belief, information acquisition, and trust in multi-agent systems–a modal logic formulation. *Artif. Intell.* 149(1):31–60.

Lorini, E., and Demolombe, R. 2008. From binary trust to graded trust in information sources: A logical perspective. In Falcone, R.; Barber, K. S.; Sabater-Mir, J.; and Singh, M. P., eds., *AAMAS-TRUST*, volume 5396 of *Lecture Notes in Computer Science*, 205–225. Springer.

Matt, P.-A.; Morge, M.; and Toni, F. 2010. Combining statistics and arguments to compute trust. In *Ninth International Conference on Autonomous Agents and Multiagent Systems, AAMAS, In press*.

Parsons, S.; McBurney, P.; and Sklar, E. 2010. Reasoning about trust using argumentation: A position paper. In *Seventh International Workshop on Argumentation in Multi-Agent Systems, ArgMAS, In press*.

Prade, H. 2007. A qualitative bipolar argumentative view of trust. In Prade, H., and Subrahmanian, V. S., eds., *SUM*, volume 4772 of *Lecture Notes in Computer Science*, 268–276. Springer.

Stranders, R.; de Weerdt, M.; and Witteveen, C. 2007. Fuzzy argumentation for trust. In Sadri, F., and Satoh, K., eds., *CLIMA VIII*, volume 5056 of *Lecture Notes in Computer Science*, 214–230. Springer.