Arguing about the Trustworthiness of the Information Sources

Serena Villata[1], Guido Boella[1], Dov M. Gabbay[2], Leendert van der Torre[3]

[1] Dipartimento di Informatica, Universita di Torino {villata,guido}@di.unito.it
[2] King's College, London dov.gabbay@kcl.ac.uk
[3] CSC, University of Luxembourg leendert@vandertorre.com

Abstract. Trust minimizes the uncertainty in the interactions among the information sources. To express the possibly conflicting motivations about trust and distrust, we reason about trust using argumentation theory. First, we show how to model the sources and how to attack untrustworthy sources. Second, we provide a focused representation of trust about the sources in which trust concerns not only the sources but also the information items and the relation with other information.

1 Introduction

Trust is a mechanism for managing uncertain information in decision making, considering the information sources. In their interactions, the information sources have to reason whether they should trust or not the other sources, and on the extent to which they trust those other sources. This is important, for example, in medical contexts, where doctors have to inform the patient of the pro and con evidence concerning some treatment, or in decision support systems where the user is not satisfied by an answer without explanations.

In this paper, a way to deal with the conflicts about trust using Dung's abstract argumentation framework is presented. A Dung argumentation framework [5] can be instantiated by the arguments and attacks defined by a knowledge base, and the knowledge base inferences are defined in terms of the claims of the justified arguments, e.g., the ASPIC+ framework instantiates Dung frameworks with accounts of the structure of arguments, the nature of attack and the use of preferences [14]. In such a kind of framework, arguments are instantiated by sentences of a single knowledge base, without reference to the information sources. The following example presents an informal dialogue illustrating conflicts about trust among the sources and the pieces of information they provide:

- *Witness1: I suspect that the man killed his boss in Rome. (a)*
- *Witness1: But his car was broken, thus he could not reach the crime scene. (b)*
- *Witness2: Witness1 is a compulsive liar. (c)*
- *Witness3: I repaired the suspect's car at 12pm of the crime day. (d)*
- *Witness4: I believe that Witness3 is not able to repair that kind of car. (e)*
- *Witness5: The suspect has another car. (f)*
- *Witness6: Witness5 saw that the suspect parked 2 cars in my underground parking garage 3 weeks ago. (g)*

- *Witness2: Witness5 was on holidays 3 weeks ago. (h)*

To deal with the dimension of conflict in handling trust, we propose to use argumentation theory, since it is a mechanism to reason about conflicting information. The problem is that it is difficult to formalize the example above with sentences from a single knowledge base only, e.g., to model it in ASPIC+ style instantiated argumentation. We address the following research question: *How to instantiate abstract argumentation with a finite number of knowledge bases instead of a single one, in which the pieces of information are thus indexed by the source?* This breaks down into the following subquestions:

1. How to represent the information sources and attack their trustworthiness?
2. How to represent pro and con evidence, as done in Carneades [7]?
3. How to attack the sources' trustworthiness about single information items?

To answer the research question we propose meta-argumentation [8, 11, 2]. Meta-argumentation provides a way to instantiate abstract arguments, i.e., abstract arguments are treated as meta-arguments. It allows us not only to reason about arguments such as sentences from a knowledge base indexed by the information source, but also to introduce in the framework other instances like arguments about the trustworthiness of sources. The advantage is that we do not extend Dung's framework in order to introduce trust but we instantiate his theory with meta-arguments. We do not claim that argumentation is the only way to model trust, but we underline that, when the sources argue, they are strongly influenced by the trustworthiness relationships with the other sources.

The paper follows the research questions. After a brief introduction on meta-argumentation, we describe our model for representing the information sources and the focused trust relationships involving them.

2 Meta-Argumentation

A Dung-style argumentation framework AF [5] is a tuple $\langle A, \rightarrow \rangle$ where A is a set of elements called *arguments* and $\rightarrow$ is a binary relation called *attack* defined on $A \times A$. A Dung's semantics consists of a set of arguments that does not contain an argument attacking another argument in the set. For more details, see [5].

Like Baroni and Giacomin [1] 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 they 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 1. *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: $\mathcal{E}(\langle A, \rightarrow \rangle) \subseteq 2^{A}$.*

Meta-argumentation instantiates Dung's theory with meta-arguments, such that *Dung's theory is used to reason about itself* [3]. 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. Moreover, auxiliary arguments are introduced to represent, for example, attacks, so that, by being arguments themselves, they can be attacked or attack other arguments. The meta-argumentation methodology is summarized in Figure 1.

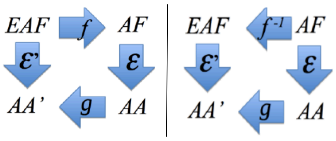

Fig. 1. The meta-argumentation methodology.

The function f assigns to each argument a in the EAF, a meta-argument "argument a is accepted" in the basic argumentation framework. The function f^{-1} instantiates an AF with an EAF. We use Dung's acceptance functions $\mathcal{E}$ to find functions $\mathcal{E}'$ between EAFs and the acceptable arguments AA' they return. The accepted arguments of the meta-argumentation framework are a function of the EAF $AA' = \mathcal{E}'(EAF)$. The transformation function consists of two parts: the function f^{-1}, transforming an AF to an EAF, and a function g which 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 the meta-argumentation approach is to define the set of EAFs. The second step consists of defining flattening algorithms as a function from this set of EAFs to the set of all basic AF: $f : EAF \to AF$. The inverse of the flattening is the instantiation of the AF. See [2, 16] for further details. We define an EAF as a set of partial argumentation frameworks of the sources $\langle A, \langle A_1, \to_1\rangle, \ldots, \langle A_n, \to_n\rangle, \to\rangle$.

Definition 2. *An extended argumentation framework* EAF *is a tuple* $\langle A, \langle A_1, \to_1\rangle, \ldots, \langle A_n, \to_n\rangle, \to\rangle$ *where for each source* $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathcal{U}$ *is a set of arguments,* $\to$ *is a binary attack relation on* $A \times A$, *and* $\to_i$ *is a binary relation on* $A_i \times A_i$. *The universe of meta-arguments is* $MU = \{acc(a) \mid a \in \mathcal{U}\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in \mathcal{U}\}$, *where* $X_{a,b}, Y_{a,b}$ *are the meta-arguments corresponding to the attack* $a \to b$. *The flattening function* f *is given by* $f(EAF) = \langle MA, \longmapsto\rangle$, *where* MA *is the set of meta-arguments and* $\longmapsto$ *is the meta-attack relation. 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 subsets of arguments* $AA \subseteq 2^{MU}$, *it is given by* $g(AA) = \{g(B) \mid B \in AA\}$.

Given an acceptance function $\mathcal{E}$ *for an* AF, *the extensions of accepted arguments of an* EAF *are given by* $\mathcal{E}'(EAF) = g(\mathcal{E}(f(EAF)))$. *The derived acceptance function* $\mathcal{E}'$ *of the* EAF *is thus* $\mathcal{E}' = \{(f^{-1}(a), g(b)) \mid (a,b) \in \mathcal{E}\}$. *We say* *that the source i provides evidence in support of argument a when $a \in A_i$, and the source supports the attack $a \rightarrow b$ when $a \rightarrow b \in \rightarrow_i$.*

Note that the union of all the A_i does not produce A because A contains also those arguments which are not supported by the sources, and are just "put on the table". Definition 3 presents the instantiation of a basic AF as a set of partial argumentation frameworks of the sources using meta-argumentation.

Definition 3. *Given an $EAF = \langle A, \langle A_1, \rightarrow_1 \rangle, \ldots, \langle A_n, \rightarrow_n \rangle \rangle$ where for each source $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathcal{U}$ is a set of arguments, $\rightarrow \subseteq A \times A$, and $\rightarrow_i \subseteq A_i \times A_i$ is a binary relation over A_i. $MA \subseteq MU$ 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 a source $1 \leq i \leq n$ such that $a, b \in A_i$ and $a \rightarrow b \in \rightarrow_i$.*

Intuitively, the $X_{a,b}$ auxiliary argument means that the attack $a \rightarrow b$ is "inactive", and the $Y_{a,b}$ auxiliary argument means that the attack is "active". An argument of an EAF is acceptable iff it is acceptable in the flattened AF.

3 Modelling trust in meta-argumentation

A number of authors have highlighted that the definition of trust is difficult to pin down precisely, thus in the literature there are numerous distinct definitions. Castelfranchi and Falcone [4] 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 [6] 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*". Common elements are a consistent degree of uncertainty and conflicting information associated with trust. In this paper, we do not refer to the actions of the sources, but we provide a model for representing the conflicts the sources have to deal with trust. We follow Liau [9] where the influence of trust on the assimilation of information into the source's mind is considered: "*if agent i believes that agent j has told him the truth on p and he trusts the judgement of j on p, then he will also believe p*". Extending the model by introducing goals to model the former two definitions is left for future work.

3.1 Information sources

The reason why abstract argumentation is not suited to model trust is that an argument, if it is not attacked by another acceptable argument, is considered acceptable. This prevents us from modeling the situation where, for an argument to be acceptable, it must be related to some sources which provide the evidence for such an argument to be accepted. Without an explicit representation of the sources, it becomes impossible to talk about trust: the argument can only be attacked by conflicting information, but it cannot be made unacceptable due to the lack of trust in the source.

Thus a challenge is how to model evidence, where sources are a particular type of evidence. Arguments needing evidence are well known in legal argumentation, where the notion of burden of proof has been introduced [7]. Meta-argumentation provides a means to model burden of proof in abstract argumentation without extending argumentation. The idea is to associate to each argument $a \in A$ put on the table, which is represented by means of meta-argument $acc(a)$, an auxiliary argument $W_{acc(a)}$ attacking it. Being auxiliary this argument is filtered out during the unflattening process. This means that without further information, just as being put on the table, argument a is not acceptable since it is attacked by the acceptable argument $W_{acc(a)}$ and there is no evidence defending it against this "default" attack, as visualized in Figure 2.a for arguments a and b. This evidence is modeled by arguments which attack auxiliary argument $W_{acc(a)}$, thus reinstating meta-argument $acc(a)$. Attacks are modeled as arguments as well. For each auxiliary argument $Y_{a,b}$, representing the activation of the attack, we associate an auxiliary argument $W_{Y_{a,b}}$.

Each argument a in the sources' mind is supported by means of an attack on $W_{acc(a)}$. Sources are introduced in the meta-argumentation framework under the form of meta-arguments "*source i is trustable*", $trust(i)$, for all the sources i. We represent the fact that one or more information sources support the same argument by letting them attack the same $W_{acc(a)}$ auxiliary argument. An example of multiple evidence is depicted in Figure 2.b. In the figures, we represent the information sources as boxes, and the arguments as circles where grey arguments are the acceptable ones. As for arguments, an attack to become active needs some trusted agent.

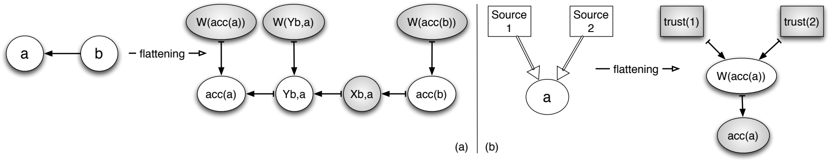

Fig. 2. (a) arguments and attack without evidence, (b) multiple evidence.

We have now to discuss which semantics we adopt for assessing the acceptability of the arguments and the sources. For example, suppose that two sources claim they are each untrustworthy. What is the extension? We adopt admissibility based semantics. We do not ask for completeness because if one wants to know whether a particular argument is acceptable, the whole model is not needed, just the part related to this particular argument is needed.

We extend the EAF proposed in Definition 2 by adding evidence provided by information sources and second-order attacks, such as attacks from an argument or attacks to another attack. For more details about second-order attacks in meta-argumentation, see [11, 2]. The unflattening function g and the acceptance function $\mathcal{E}'$ are defined as above.

Definition 4. *An EAF with second-order attacks is a tuple $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2\rangle, \rightarrow\rangle$ where for each source $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathcal{U}$ is a set of arguments, $\rightarrow \subseteq A \times A$, $\rightarrow_i$ 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 5 presents the instantiation of an EAF with second-order attacks as a set of partial frameworks of the sources using meta-argumentation.

Definition 5. *Given an $EAF = \langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2\rangle \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2\rangle, \rightarrow\rangle$, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:*

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

We say that source i is trustworthy when meta-argument $trust(i)$ is acceptable, and we say that i provides evidence in support of argument a (or attack $a \rightarrow b$) when $a \in A_i$ (when $a \rightarrow b \in \rightarrow_i$), and $trust(i) \longmapsto W_{acc(a)}$ ($trust(i) \longmapsto W_{Y_{a,b}}$).

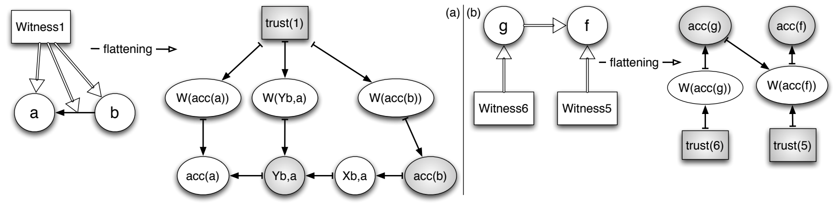

Fig. 3. Introducing (a) the sources, (b) evidence for the arguments.

Example 1. Consider the informal dialogue in the introduction. We represent the sources in the argumentation framework, as shown in Figure 3.a. Witness1 proposes a and b and the attack $a \rightarrow b$. Using the flattening function of Definition 5, we add 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. Witness1 provides evidence for these arguments, and the attack $b \rightarrow a$ by attacking the respective auxiliary arguments W. In the remainder of the paper, we model the other conflicts highlighted in the dialogue.

Let $trust(i)$ be the information source i and $acc(a)$ and $Y_{a,b}$ the argument a_i and the attack $a \rightarrow_i b$ respectively, as defined in Definitions 2 and 3. $trust(i)$ can provide evidence for $acc(a)$ and $Y_{a,b}$. Sources can attack other sources as well as their arguments and attacks. With a slight abuse of notation, we write $a \in \mathcal{E}'(EAF)$, even if the latter is a set of extensions, with the intended meaning that a is in some of the extensions of $\mathcal{E}'$. We now provide some properties of our model. Some of the proofs are omitted due to the lack of space.

Proposition 1. *Assume admissibility based semantics, if an argument $a \in A$ is not supported by evidence, i.e., $a \notin A_i$ for all i, then a is not accepted, $a \notin \mathcal{E}'(EAF)$.*

Proof. We prove the contrapositive: if argument a is accepted, then argument a is supported. Assume argument a is accepted. Then auxiliary argument $W_{acc(a)}$ is rejected due to the conflict-free principle. Meta-argument $acc(a)$ is defended, so $W_{acc(a)}$ is attacked by an accepted argument using admissible semantics. Auxiliary argument $W_{acc(a)}$ can only be attacked by meta-argument $trust(i)$. We conclude that a is supported.

Proposition 1 is strengthened to Proposition 2.

Proposition 2. *If an argument a is not supported, $a \notin A_i$, then the extensions $\mathcal{E}'(EAF)$ are precisely the same as the extensions of the $AF = \langle A, \rightarrow \rangle$ in which $a \notin A$, and the attacks on a or from a do not exist, i.e., $b \rightarrow a \notin\rightarrow$ and $a \rightarrow c \notin\rightarrow$.*

Proposition 3. *If an attack $a \rightarrow b$ is not supported, i.e., $a \rightarrow b \notin\rightarrow_i$, then the extensions $\mathcal{E}'(EAF)$ are precisely the same as the extensions of the $AF = \langle A, \rightarrow \rangle$, in which the attack does not exist, $a \rightarrow b \notin\rightarrow$.*

Proposition 4. *Assume EAF is a framework in which argument a is supported by the trustworthy source i, and there is another trustworthy source j. In that case, the extensions are the same if also j provides an evidence in support of a.*

3.2 Evidence for arguments

The evidence in favor of the arguments is evidence provided by the agents for the arguments/attacks they propose. At the meta-level, this is modeled as an attack from meta-argument $trust(i)$ to W auxiliary arguments. However, there are other cases in which more evidence is necessary to support the acceptability of an argument. Consider the case of Witness1. His trustworthiness is attacked by Witness2. What happens to the evidence provided by Witness1? Since the source is not trustworthy then it cannot provide evidence. Meta-argument $trust(1)$ becomes not acceptable and the same happens to all its arguments and attacks. What is needed to make them acceptable again is more evidence. This evidence can be provided under the form of another argument which reinstates the acceptability of these information items.

Definition 5 allows only the sources to directly provide evidence for the information items. As for Witness5 and Witness6 in the dialogue, sources can provide evidence also by means of other arguments. This cannot be represented using Definition 5, this is why we need to extend it with an evidence relation $\looparrowright$ representing evidence provided under the form of arguments for the information items of the other sources.

Definition 6. *An EAF with evidence* $TEAF^2 = \langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \looparrowright_1 \rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \looparrowright_n \rangle, \rightarrow \rangle$ *where* $\looparrowright_i$ *is a binary relation on* $A_i \times A_j$ *and the set of meta-arguments* MA *is* $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on* MA *such that hold the conditions of Definition 5, and:* $acc(a) \longmapsto W_{acc(b)}$ *iff* $a, b \in A_i$ *and* $a \looparrowright_i b$*, and* $W_{acc(b)} \longmapsto acc(b)$ *iff* $b \in A$ *and* $a \looparrowright_i b$*. We say that a source* j *supports the evidence provided by other sources to argument* a *when* $a \notin A_j, b \in A_j$*, and* $acc(b) \longmapsto W_{acc(a)}$*.*

The following properties hold for Definition 6.

Proposition 5. *If there are multiple arguments* $a_1 \in A_1, \ldots, a_n \in A_n$ *providing evidence for an argument* $b \in A_k$ *(or an attack), and there are no attacks on the arguments,* $c_1 \rightarrow a_1 \not\in \rightarrow_1, \ldots, c_n \rightarrow a_n \not\in \rightarrow_n$*, then* b *(or the attack) is accepted,* $b \in \mathcal{E}'(EAF)$*, iff at least one of the sources is trustworthy, i.e.,* $trust(j) \in \mathcal{E}(f(EAF))$ *with* $j \in 1, \ldots, n$*.*

Proposition 6. *Suppose two sources* i *and* j *provide evidence for the same argument* a*, i.e.,* $a \in A_i$ *and* $a \in A_j$*, then it is the same whether a source* k *supports the evidence provided by* i *or* j*, i.e.,* $b \in A_k$ *and* $acc(b) \longmapsto W_{acc(a)}$*.*

Example 2. Consider the dialogue in the introduction. Argument g by Witness6 is an evidence for argument f by Witness5. This evidence is expressed in meta-argumentation in the same way as evidence provided by the sources, such as an attack to $W_{acc(f)}$ attacking $acc(f)$. In this case, it is meta-argument $acc(g)$ which attacks $W_{acc(f)}$, as visualized in Figure 3.b.

3.3 Focused trust relationships

In our model, trust is represented *by default* as the absence of an attack towards the sources or towards the information items and as the presence of evidence in favor of the pieces of information. On the contrary, the distrust relationship is modeled as a lack of evidence in support of the information items or as a direct attack towards the sources and their pieces of information.

In the informal dialogue, Witness2 attacks the trustworthiness of Witness1 as a credible witness. In this way, she is attacking each argument and attack proposed by Witness1. Witness4, instead, is not arguing against Witness3 but she is arguing against the attack $d \rightarrow b$ as it is proposed by Witness3. Finally, for Witness2 the untrustworthiness of Witness6 is related only to the argument g. We propose a focused view of trust in which the information sources may be attacked for being untrustworthy or for being untrustworthy only concerning a particular argument or attack. Definition 7 presents an EAF in which a new relation DT between sources is given to represent distrust.

Definition 7. *A trust-based extended argumentation framework $TEAF$ is a tuple $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \looparrowright_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \looparrowright_n, DT_n\rangle, \rightarrow\rangle$ where for each source $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathcal{U}$ is a set of arguments, $\rightarrow \subseteq A \times A$, $\rightarrow_i \subseteq A_i \times A_i$ is a binary relation, $\rightarrow_i^2$ is a binary relation on $(A_i \cup \rightarrow_i) \times \rightarrow_i$, $\looparrowright_i$ is a binary relation on $A_i \times A_j$ and $DT \subseteq A_i \times \vartheta$ is a binary relation such that $\vartheta = j$ or $\vartheta \in A_j$ or $\vartheta \in \rightarrow_j$.*

Definition 8 shows how to instantiate an EAF enriched with a 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. The unflattening function g and the acceptance function $\mathcal{E}'$ are defined as above.

Definition 8. *Given a $TEAF = \langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \looparrowright_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \looparrowright_n, DT_n\rangle, \rightarrow\rangle$, see Definition 7, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a, b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that hold the conditions of Definitions 5 and 6, and:*

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

We say that a source i is untrustworthy when there is an attack from an argument $a_j \in A_j$ to i, a_jDT_ji. We say that an argument $a_i \in A_i$ or attack $a \rightarrow_i b \in \rightarrow_i$

is untrustworthy when there is an attack from an argument $a_j \in A_j$ to a_i or $a \rightarrow_i b$, $a_j DT_j a_i$ or $a_j DT_j (a \rightarrow_i b)$.

Proposition 7. *Assume that source i is the only source providing evidence for argument $a \in A_i$ and attack $c \rightarrow b \in \rightarrow_i$, and assume admissibility based semantics. If the information source i is considered to be untrustworthy, then a and $c \rightarrow b$ are not acceptable.*

Proof. We prove the contrapositive: if the arguments and attacks supported by an information source i are acceptable then the information source i is considered to be trustworthy. Assume the source supports argument a and the attack $c \rightarrow b$ and assume that this argument and this attack are acceptable. Then auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ are rejected due to the conflict-free principle. Meta-arguments $acc(a)$ and $Y_{c,b}$ are defended, thus $W_{acc(a)}$ and $W_{Y_{c,b}}$ are attacked by an acceptable argument, using admissible semantics. We assumed that this argument and this attack have no other evidence, so auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ can only be attacked by meta-argument $trust(i)$. Since they are attacked by an acceptable argument, we conclude that the source i is acceptable.

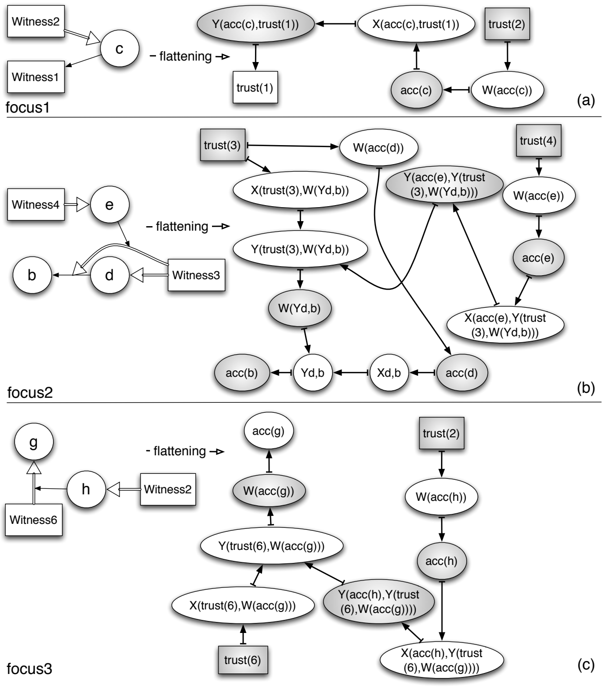

Fig. 4. Focused trust in argumentation.

Example 3. Figure 4.a shows that Witness2 attacks the trustworthiness of Witness1 by means of argument c. In meta-argumentation, we have that $trust(2)$ provides evidence for $acc(c)$ by attacking meta-argument $W_{acc(c)}$ and, with meta-arguments X, Y, it attacks $trust(1)$. This means that if Witness1 is untrustworthy then each of his arguments and attacks cannot be acceptable either, if there is no more evidence. The set of acceptable arguments for the meta-argumentation framework is $\mathcal{E}(f(focus1)) = \{trust(2), acc(c), Y_{acc(c),trust(1)}\}$. In Figure 4.b-c, instead, the attack is directed against a precise information item provided by the source. In particular, Witness4 attacks the attack $d \rightarrow b$ of Witness3. This is achieved in meta-argumentation by means of an attack from meta-argument $acc(e)$, for which $trust(4)$ provides evidence, to the attack characterized by auxiliary argument $Y_{d,b}$. The set of acceptable arguments is $\mathcal{E}(f(focus2)) = \{trust(4), trust(3), acc(d), acc(e), acc(b), Y_{acc(e),Y_{trust(3),W_{Y_{b,d}}}}, W_{Y_{d,b}}\}$. Witness3's attack $d \rightarrow b$ is evaluated as untrustworthy by Witness4 and thus it is not acceptable. Finally, Witness2 evaluates Witness6 as untrustworthy concerning argument g. In meta-argumentation, $trust(2)$, by means of meta-argument $acc(h)$, attacks meta-argument $acc(g)$ proposed by $trust(6)$. The set of acceptable arguments is $\mathcal{E}(f(focus3)) = \{trust(2), trust(6), acc(h), Y_{acc(h),Y_{trust(6),W_{acc(g)}}}, W_{acc(g)}\}$.

4 Related work and conclusions

Parsons et al. [12] highlight what are the mechanisms to investigate through argumentation, first of all the provenance of trust. Tang et al. [15] present a framework to introduce the sources in argumentation and to explicitly express the degrees of trust. They connect agent-centric trust networks to argumentation networks. They do not have the possibility to attack the trustworthiness of the agents as well as the trustworthiness of single arguments and attacks. We do not express the degrees of trust. Matt et al. [10] propose to construct a belief function both from statistical data and from arguments in the context of contracts. We do not address the computation of trust by an evaluator in isolation, instead all trust relationships are evaluated together. Prade [13] presents a bipolar qualitative argumentative modeling of trust where trust and distrust are assessed independently. We do not use observed behavior and reputation to compute trust and we are interested in abstract arguments and not in arguments with an abductive format.

Trust plays an important role in many research areas of artificial intelligence, particularly in the semantic web and multiagent systems where the sources have to deal with conflicting information from other sources. We provide a model where the information sources can be introduced into the framework. In argumentation systems as ASPIC+, arguments come from a single knowledge base and they have the form $\langle\{p, p \rightarrow q\}, q\rangle$. We propose to introduce the sources, e.g., $\langle\{1 : p, 2 : p \rightarrow q\}, 2 : q\rangle$, by instantiating abstract argumentation with the different knowledge bases of the sources using meta-argumentation. In our model, arguments need to be supported in order to be accepted. Furthermore, the trustworthiness of the sources can be attacked directly, or the attack can be focused on single arguments or attacks.

We address several issues as future research. First, 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's trustworthiness, but there is feedback such that the invalidation of the data feeds back on the sources' credibility [4]. Second, we will 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

1. P. Baroni and M. Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artif. Intell.*, 171(10-15):675–700, 2007.
2. G. Boella, D. M. Gabbay, L. van der Torre, and S. Villata. Meta-argumentation modelling I: Methodology and techniques. *Studia Logica*, 93(2-3):297–355, 2009.
3. G. Boella, L. van der Torre, and S. Villata. On the acceptability of meta-arguments. In *Procs. of IAT*, pages 259–262. IEEE, 2009.
4. C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-Cognitive and Computational Model.* Wiley, 2010.
5. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–357, 1995.
6. D. Gambetta. Can we trust them? *Trust: Making and breaking cooperative relations*, pages 213–238, 1990.
7. T. F. Gordon, H. Prakken, and D. Walton. The Carneades model of argument and burden of proof. *Artif. Intell.*, 171(10-15):875–896, 2007.
8. H. Jakobovits and D. Vermeir. Robust semantics for argumentation frameworks. *J. Log. Comput.*, 9(2):215–261, 1999.
9. C.-J. Liau. Belief, information acquisition, and trust in multi-agent systems–a modal logic formulation. *Artif. Intell.*, 149(1):31–60, 2003.
10. P.-A. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *Procs. of AAMAS*, pages 209–216, 2010.
11. S. Modgil and T. Bench-Capon. Metalevel argumentation. Technical report, www.csc.liv.ac.uk/research/ techreports/techreports.html, 2009.
12. S. Parsons, P. McBurney, and E. Sklar. Reasoning about trust using argumentation: A position paper. In *Procs. of ArgMAS*, 2010.
13. H. Prade. A qualitative bipolar argumentative view of trust. In *Procs. of SUM, volume 4772 of LNCS*, pages 268–276, 2007.
14. H. Prakken. An abstract framework for argumentation with structured arguments. Technical Report UU-CS-2009-019, Utrecht University, 2009.
15. Y. Tang, K. Cai, E. Sklar, P. McBurney, and S. Parsons. A system of argumentation for reasoning about trust. In *Procs. of EUMAS*, 2010.
16. S. Villata. *Meta-Argumentation for Multiagent Systems: Coalition Formation, Merging Views, Subsumption Relation and Dependence Networks.* PhD thesis, University of Turin, 2010.