Computational Models of Argument
H. Prakken et al. (Eds.)

doi:10.3233/FAIA200511

Explanation Semantics for Abstract Argumentation

Beishui LIAO [a], Leendert VAN DER TORRE [b,a]
[a] *Zhejiang University*
[b] *University of Luxembourg*

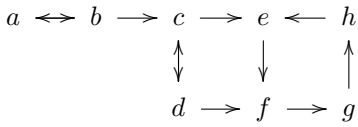
Abstract. This paper studies explanation semantics of argumentation by using a principle-based approach. In particular, we introduce and study explanation semantics associating with each accepted argument a set of such explanation arguments. We introduce various principles for explanation semantics for abstract argumentation, and list various relations among them. Then, we introduce explanation semantics based on defence graphs, and show which principles they satisfy.

Keywords. Argumentation semantics, explanation, defense graph

1. Introduction

In this paper we consider the use of formal argumentation for explainable AI [15]. According to the empirical results reported by Ye and Johnson [19], justification is the most effective type of explanation to bring about changes in user attitudes toward the system. Formal argumentation, as a formalism for representing and reasoning with inconsistent and incomplete information [1,8], provides various ways for explaining why a claim or a decision is made, in terms of justification, dialogue, and dispute trees [11]. Besides some application specific methods such as argumentation-based explanation in case-based reasoning [5] and in scientific debates [18], etc., there are some approaches for defining general theories of explanation about acceptance of arguments in terms of the notion of defense [9,20]. Along this line of work, in this paper, we study a related notion of explanation for abstract argumentation as a kind of semantics: an argument is accepted because some other arguments are accepted, and propose a new semantics, called explanation semantics.

Some basic notions of explanation semantics are illustrated by the following example. The graph below represents an argumentation framework, of which the nodes are called arguments, and the arrows represent attacks between arguments. The graph contains three strongly connected components (SCCs), $\{a, b\}$, $\{c, d\}$ and $\{e, f, g, h\}$, which represents the graph-theoretic property that there is a path from each element to each other element of the SCC. The three preferred extensions are $\{\{a, c, f, h\}, \{a, d, e, g\}, \{b, d, e, g\}\}$.



Dauphin *et al.* [6] observe that in such examples, every strongly connected component can be seen as a choice to accept some attack-free set of arguments of the SCC. For example, if argument a is chosen in the first SCC, then either c or d can be chosen in the second SCC. However, if b is chosen in the first SCC, then only d can be chosen in the second SCC. Thus, the choice in the first SCC determines the set of alternatives in the second SCC. Likewise, whatever is chosen in the first or second SCC, in the third SCC there is only one alternative.

The explanation extensions may be $\{\{a^a, c^c, f^c, h^c\}, \{a^a, d^d, e^d, g^d\}, \{b^b, d^b, e^b, g^b\}\}$. For the first choice between accepting argument a or accepting argument b, each argument is labeled by itself, which expresses that the choice does not depend on other choices. For the choice between accepting argument c and accepting argument d, it partially depends on the first choice. If accepting argument a is chosen, then the choice between accepting c or d is not restricted. Alternatively, if accepting argument b is chosen, then the only choice is to accept d. Finally, either accepting c or accepting d is chosen, the remaining choice is unique.

Thus we say that the reason that g is accepted, is because d is accepted in case a is accepted, or because b is accepted. Distinguishing this kind of explanations provides more information than only the acceptance or rejection of g. Also, we can distinguish direct from indirect reasons, and more.

The layout of this paper is as follows. In Section 2 we introduce the standard terminology of Dung's abstract argumentation and our variant of explanation semantics. In Section 3 we introduce various principles/properties of explanation semantics, and in Section 4 and 5 we introduce some concrete examples of explanation semantics.

2. Abstract Argumentation and Explanation

In this section, we recall some basic notions of abstract argumentation that are used in this paper, and then we introduce explanation semantics.

2.1. Traditional semantics

All notions in this paper are defined on abstract argumentation frameworks, which is a directed graph in which nodes are called arguments and arrows represent attacks between arguments. As usual, we write a^- for the set of attackers of a, and a^+ for the set of arguments a attacks.

Definition 1 (Argumentation framework) *An argumentation framework is a pair $F = (A, \rightarrow)$ where A is a set of arguments and $\rightarrow \subseteq A \times A$ is a binary relation over A, called attacks. An argument a attacking an argument b is written as $a \rightarrow b$. A set of arguments B attacks a, written as $B \rightarrow a$, if there exists $b \in B$ such that $b \rightarrow a$. Given $a \in A$, we define $a_F^- = \{b \in A \mid b \rightarrow a\}$ and $a_F^+ = \{b \in A \mid a \rightarrow b\}$. When $a_F^- = \emptyset$, we say that a is unattacked, or a is an initial argument. When the context is clear, we also write a^+ and a^- for a_F^+ and a_F^- respectively.*

Definition 2 (Traditional argumentation semantics) *Let $\mathcal{F}$ be the set of all argumentation frameworks $F = (A, \rightarrow)$. Let an extension of F be a subset of A. Traditional argumentation semantics is a function σ from $\mathcal{F}$ to sets of their extensions, associating with each argumentation framework F a subset of 2^A, denoted as $\sigma(F)$.*

Given an argumentation framework $F = (A, \rightarrow)$, various types of argument extensions of F can be defined as follows.

Definition 3 (Dung's argumentation semantics) *Let $F = (A, \rightarrow)$ be an argumentation framework, $\mathcal{E} \subseteq A$ be a set of arguments, and $a \in A$ be an argument. $\mathcal{E}$ is conflict-free if and only if there exist no $a, b \in A$ such that $a \rightarrow b$. $\mathcal{E}$ defends a if and only if for each $b \in a_F^-$, $\mathcal{E} \rightarrow b$. $\mathcal{E}$ is admissible if and only if $\mathcal{E}$ is conflict-free, and each argument in $\mathcal{E}$ is defended by $\mathcal{E}$.*

- *$\mathcal{E}$ is a complete extension if and only if $\mathcal{E}$ is admissible, and each argument in A that is defended by $\mathcal{E}$ is in $\mathcal{E}$.*
- *$\mathcal{E}$ is the grounded extension if and only if $\mathcal{E}$ is the minimal (with respect to set-inclusion) complete extension.*
- *$\mathcal{E}$ is a preferred extension if and only if $\mathcal{E}$ is a maximal (with respect to set-inclusion) complete extension.*
- *$\mathcal{E}$ is a stable extension if and only if $\mathcal{E}$ is conflict-free and $\mathcal{E}$ attacks each argument that is not in $\mathcal{E}$.*

2.2. Explanation semantics

In the following definition, an explanation semantics is a function from graphs to sets of explanation extensions, where each explanation extension is a set of explanation arguments, where each explanation is a set of arguments. We use the letter E for extension, and we use the letter R for explanation, which expresses that the explanation is the reason the argument is accepted.

Definition 4 (Explanation semantics) *Let an explanation of each argument in F be a subset of A, and let an explanation extension of F be a subset of A, of which each argument is labeled with an explanation. Explanation semantics is a function Σ from $\mathcal{F}$ to sets of their explanation extensions, denoted as $\Sigma(F)$. We write a^R for the argument a with explanation R. When R contains a single argument (say, b), a^R is also written as a^b for conciseness.*

Each explanation semantics induces a traditional semantics, simply by stripping the labels. In such a case, we say that the explanation semantics explains the traditional semantics.

Definition 5 (Explaining argumentation semantics) *Explanation semantics Σ explains traditional semantics σ iff for all F, we have $\sigma(F) = \{\{x \mid x^R \in E\} \mid E \in \Sigma(F)\}$.*

Definition 6 (Explainable semantics) *A traditional semantics is explainable with respect to properties X iff there is an explanation semantics satisfying properties X, explaining the traditional semantics.*

3. Principles for explanation semantics

We start with three elementary properties. The first property is called U for Uniqueness and says that each accepted argument has a unique explanation.

Property 1 (Uniqueness) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R, a^S \in E$, we have $R = S$.*

The second property is called A for *Acceptance* and says that an explanation consists of a set of accepted arguments.

Property 2 (Acceptance) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R \in E$, the explanation R consists of arguments that are part of the extension $\{x \mid x^S \in E\}$, i.e., $R \subseteq \{x \mid x^S \in E\}$.*

The third property says that the explanation defends the accepted argument, possibly recursively. It uses the following characteristic function returning all arguments in F recursively defended by the arguments in S, which we write as c.

Definition 7 (Characteristic function) $c_0(S, F) = \{a \in F \mid S \textit{ defends } a\}$. $c_{i+1}(S, F) = c_0(S \cup c_i(S, F), F)$. $c_\infty(S, F) = \cup_{i=0}^{\infty} c_i(S, F)$.

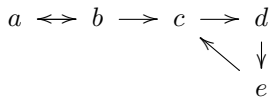
The third property is called I for *Indirect Defense* and says that for all $a^R \in E$, if we iteratively apply the characteristic function to explanation R, then we get a set of arguments containing a.[1]

Property 3 (Indirect Defense) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R \in E$, we have $a \in c_\infty(R, F)$.*

The fourth property strengthens indirect defense to direct defense. Obviously property D implies property I, in the sense that if an explanation semantics satisfies property D, it also satisfies property I.

Property 4 (Direct defense) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R \in E$, we have $a \in c_0(R, F)$.*

Example 1 (Two-three cycle) *Consider the following widely discussed two-three cycle framework:*



There are two preferred extensions $\{a\}$ and $\{b, d\}$ under Dung's argumentation semantics. The unique explanation extensions satisfying Properties UAID and explaining these preferred extensions are $\{a^a\}$ and $\{b^b, d^b\}$.

Proposition 1 (Explainable semantics, Prop. UAID) *All traditional Dung semantics are explainable with respect to Properties UAID.*

[1] Alternatively, we could require that for all $a^R \in E$, if we iteratively apply the characteristic function to label R, then we get a set of conflict-free arguments containing a. However, since all semantics we consder are conflict free, in the sense that the accepted arguments do not attack each other, we do not consider this variant of Property 3.

Proof For all $\mathcal{E} \in \sigma(\mathcal{F})$, $\forall a \in \mathcal{E}$, since $\mathcal{E}$ defends a, there exist a set of sets $R_1, \ldots, R_n \subseteq \mathcal{E}$ where $n \geq 1$, such that $a \in c_0(R_i, F)$ where $i = 1, \ldots, n$. Let $R_a \in \{R_1, \ldots, R_n\}$ be a minimal set with respect to set inclusion. Let $E = \{a^{R_a} \mid a \in \mathcal{E}\}$. It holds that E satisfies Properties UAID. □

The fifth property says that explanations are minimal in the sense that they do not contain superfluous arguments.

Property 5 (Minimality) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R \in E$, for all $S \subset R$ we have $a \notin c_\infty(S, F)$.*

Example 2 (Four-cycle) *Consider the following four-cycle framework:*

$$\begin{array}{ccc} a & \longrightarrow & b \\ \uparrow & & \downarrow \\ d & \longleftarrow & c \end{array}$$

There are two preferred extensions $\{a, c\}$ and $\{b, d\}$. For $\{a, c\}$, there are four different choices for the explanation extensions satisfying Properties UAIM, $\{a^a, c^a\}$, $\{a^c, c^c\}$, $\{a^a, c^c\}$, $\{a^c, c^a\}$, but only the latter also satisfies Property D.

Note that concerning Properties UAIDM, in contrast to Proposition 1, not all traditional Dung semantics are explainable. Consider the following counterexample.

Example 3 (Direct defense vs Minimality) *For the argumentation framework below, there is only one complete extension $\{a, c, e\}$, and only $E = \{a^{\{\}}, c^{\{a\}}, e^{\{c\}}\}$ satisfies Properties UAID, but E does not satisfy Property M, since $E' = \{a^{\{\}}, c^{\{\}}, e^{\{\}}\}$ is also a complete explanation extension.*

$$a \longrightarrow b \longrightarrow c \longrightarrow d \longrightarrow e$$

We therefore consider only UAIM in the following two propositions.

Proposition 2 *For all explanation semantics satisfying Properties UAIM, the label of each element of the grounded explanation extension is an empty set.*

Proof Assume that there is an element a^R such that R is not an empty set. Since $a^\emptyset$ satisfies Properties AIM, according to Properties U and M, it turns out that a^R is not an element of the explanation extension. Contradiction. □

Proposition 3 (Explainable semantics, Prop. UAIM) *All traditional Dung semantics are explainable with respect to Properties UAIM.*

Proof For all $\mathcal{E} \in \sigma(\mathcal{F})$, $\forall a \in \mathcal{E}$, since $\mathcal{E}$ defends a, there exist a set of sets $R_1, \ldots, R_n \subseteq \mathcal{E}$ where $n \geq 1$, such that $a \in c_\infty(R_i, F)$ where $i = 1, \ldots, n$. Let $R_a \in \{R_1, \ldots, R_n\}$ be a minimal set with respect to set inclusion. Let $E = \{a^{R_a} \mid a \in \mathcal{E}\}$. It holds that E satisfies Properties UAIM. □

The sixth property relates explanations by a kind of transitivity.

Property 6 (Transitivity) *For all argumentation frameworks F, all explanation extensions E of F, and all explained arguments $a^R, b^S \in E$, if $b \in R$, then $S \subseteq R$.*

Example 4 (Continue Example 2) *Among $\{a^a, c^c\}$, $\{a^a, c^a\}$, $\{a^c, c^c\}$ and $\{a^c, c^a\}$, only $\{a^c, c^a\}$ does not satisfy Property T, while others do.*

Transitivity together with the properties UAIM has as a surprising consequence that explanation arguments are themselves self-explanatory.

Proposition 4 (Self-explanation) *For all explanation semantics satisfying Properties UAIMT, if $a^R \in E$ and $b \in R$, then there exists $b^S \in E$; when S is a singleton, $b^b \in E$, i.e. b is self-explanatory.*

Proof According to Property A, b is in the corresponding extension $\{x \mid x^T \in E\}$ under Dung's argumentation semantics. So, there exists $S \subseteq \{x \mid x^T \in E\}$ such that $b^S \in E$. Then, according to Property T, $S \subseteq R$. Assume that b is not in S. Then, we may remove b from R to obtain $R' = R \setminus \{b\}$ and $a^{R'} \in E$. So, R is not minimal, contradicting Property M. Therefore, $b \in S$. When S is a singleton, $S = \{b\}$. So, $b^b \in E$. □

Note that in Proposition 4, when S is a not singleton, it might not hold that $b^b \in E$.

Example 5 *Consider the argumentation framework below. We have an explanation extension $E = \{b^{\{b,d\}}, d^d\}$. It holds that $b^{\{b,d\}} \in E$ and $b \in \{b, d\}$, but $b^b \notin E$.*

$$a \leftrightarrow b \leftarrow c \leftrightarrow d$$

Proposition 5 (Explainable semantics, Prop. UAIMT) *All traditional Dung semantics are explainable with respect to Properties UAIMT.*

Proof We need only to verify that Property T holds for each explanation extension, on the condition that Properties UAIM hold. According to the proof of Proposition 3, for all minimal sets $a^{R_a}, b^{R_b} \in E$, when $b \in R_a$, let $R'_a = (R_a \setminus \{b\}) \cup R_b$. Since we have $b \in c_\infty(R_b, F)$ and R_b is minimal, after replacing b with R_b, R'_a is minimal. So, there exists $E' = (E \setminus \{a^{R_a}\}) \cup \{a^{R'_a}\}$ such that $R_b \subseteq R'_a$. □

The following example further illustrates the idea in the above proof.

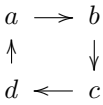
Example 6 *Continue Example 2, for $E = \{a^c, c^a\}$, let $R_a = \{c\}$ and $R_c = \{a\}$. Since $c \in R_a$, let $R'_a = (R_a \setminus \{c\}) \cup R_c = \{a\}$. Let $E' = (E \setminus \{a^{R_a}\}) \cup \{a^{R'_a}\} = \{a^a, c^a\}$. So, E' is an explanation extension satisfying Properties UAIMT.*

Let the defense set of $a^R \in E$ be the set $\{x^S \in E \mid \exists y : x \to y \to a\}$.

Property 7 (Explanation Inheritance) *For all $a^R \in E$ and $b \in R$, there is a c^S in the defense set of a^R such that $b \in S$.*

The following example illustrates property E.

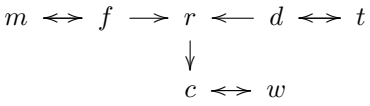
Example 7 *Consider again the four cycle framework:*



For $\{a, c\}$, the explanation extensions satisfying property UAIMTE are $\{a^a, c^a\}$, and $\{a^c, c^c\}$.

The following example is from Rienstra *et al.* [17].

Example 8 (Eating out) *The argumentation framework shown below represents the decision making of an agent planning to eat out.*



He will eat meat or fish (m or f) and take a taxi or drive himself (t or d). He drinks red wine (r) but not with fish or when driving (f and d attack r). Finally, he drinks either cola or water (c or w), but no cola if he drinks red wine (r attacks c).

The direction of the attacks implies that the agent first chooses independently between m and f and between t and d. Then he determines the status of r, which depends on f and d. Finally he chooses between c and w, which depends on r. Note that we can, of course, imagine different scenarios, but this would involve different directions of attack. E.g., if the decision about r came before *the decision between t and d, then the attack of d on r would be reversed.*

Now consider the preferred extenson $\{m, t, r, w\}$. The possible explanation extensions satisfying UAIMT are $\{m^m, t^t, r^m, w^m\}$, $\{m^m, t^t, r^t, w^t\}$, $\{m^m, t^t, r^m, w^t\}$, $\{m^m, t^t, r^t, w^m\}$, $\{m^m, t^t, r^m, w^r\}$, $\{m^m, t^t, r^t, w^r\}$, $\{m^m, t^t, r^m, w^w\}$, $\{m^m, t^t, r^t, w^w\}$. In other words, the explanation of r is either m or t, and the explanation of w is either m, t, r or w. Only the latter four satisfy property D.

Explanation $\{m^m, t^t, r^m, w^t\}$ and $\{m^m, t^t, r^t, w^m\}$ do not satisfy property E.

Proposition 6 *All traditional Dung semantics are explainable with respect to Properties UAIMTE.*

Proof We need only to verify that Property E holds on the conditions that Properties UAIMT hold. For an explanation extension E satisfying Properties UAIMT, and for all $a^R \in E$, since $a \in \{x \mid x^T \in E\}$, there exists $c, y \in A$ such that $c \to y \to a$ and $c \in \{x \mid x^T \in E\}$. So, there exists S' such that $c^{S'} \in E$. According to Proposition 5, given that $b \in R$, if $c = b \in R$, then $S' \subseteq R$. Then, according to the proof of Proposition 4, $b \in S'$. In this case, let $S = S'$, we have $b \in S$. Otherwise, $b \neq c$. Let $S = (S' \setminus c_\infty(\{b\}, F)) \cup \{b\}$. It holds that c^S satisfies Properties UAIMT. In this case, it holds that $b \in S$. □

Example 9 *Consider again the four-cycle framework: For $\{a, c\}$, $\{a^a, c^c\}$ satisfies Properties UAIM but not Property E, since $a \notin \{c\}$. Given that $c_\infty(\{a\}, F) = \{a, c\}$, let $S = (\{c\} \setminus \{a, c\}) \cup \{a\} = \{a\}$. As a result, we have $\{a^a, c^a\}$ as an explanation extension satisfying Property E.*

4. Examples of explanation semantics

Based on the principles introduced in the previous section, we may define various explanation semantics.

Definition 8 *Let $F = (A, \rightarrow)$ be an argumentation framework, and $X_F = \{a^R \mid a \in A, R \subseteq A\}$. For all $E \subseteq X_F$,*

- *E is conflict-free if and only if $\{a \mid a^R \in E\}$ is conflict-free.*
- *E is direct if and only if it is conflict-free and satisfies Properties UAID.*
- *E is a minimal explanation extension if and only if it is conflict-free and satisfies Properties UAIM.*
- *E is transitive if and only if it is a minimal explanation extension and satisfies Property* T.
- *E is explanation inherited if and only if it is transitive and satisfies Property* E.

Meanwhile, orthogonally, we say that E is complete (respectively, preferred, stable, and grounded), if and only if $\{x \mid x^R \in E\}$ is complete (respectively, preferred, stable, and grounded) under Dung's argumentation semantics.

The set of explanation extensions is represented as $\Sigma_\sigma(F)$, where $\Sigma \in \{\mathbf{D}, \mathbf{M}, \mathbf{T}, \mathbf{E}\}$, indicating direct, minimal, transitive and explanation inherited semantics, respectively, and σ is a Dung's semantics.

Example 10 *Consider again the four-cycle framework:*

$$\begin{array}{ccc} a & \longrightarrow & b \\ \uparrow & & \downarrow \\ d & \longleftarrow & c \end{array}$$

- $\mathbf{M}_{pr}(F) = \{E_1, \ldots, E_8\}$, *where* $E_1 = \{a^a, c^a\}$, $E_2 = \{a^c, c^c\}$, $E_3 = \{a^a, c^c\}$, $E_4 = \{a^c, c^a\}$, $E_5 = \{b^b, d^b\}$, $E_6 = \{b^d, d^d\}$, $E_7 = \{b^b, d^d\}$, *and* $E_8 = \{b^d, d^b\}$.
- $\mathbf{D}_{pr}(F) = \{E_4, E_8\}$.
- $\mathbf{T}_{pr}(F) = \{E_1, E_2, E_3, E_5, E_6, E_7\}$.
- $\mathbf{E}_{pr}(F) = \{E_1, E_2, E_5, E_6\}$.

According to Definition 8, it is obvious that for all F, $\mathbf{E}_\sigma(F) \subseteq \mathbf{T}_\sigma(F) \subseteq \mathbf{M}_\sigma(F)$.

Meanwhile, according to Example 10, it seems that for all F, $\mathbf{D}_\sigma(F) \subseteq \mathbf{M}_\sigma(F)$. Unfortunately, this is not the case in general: remember that in Example 3, $\mathbf{D}_\sigma(F) = \{E\}$, $\mathbf{M}_\sigma(F) = \{E'\}$, $E \neq E'$, and therefore $\mathbf{D}_\sigma(F) \not\subseteq \mathbf{M}_\sigma(F)$.

5. Explanation based on weak defense graphs

In this section, we formulate two examples of explanation semantics based on a kind of meta-argumentation framework, of which the nodes are no longer arguments, but pairs of arguments, reflecting a weak notion of defense.

Before we formally introduce this meta-argumentation theory, we introduce this weak notion of defense, which we call defense graphs. We start by making two obser-

vations concerning the role of defense in Dung's theory. The first observation is that the notion of defense by itself is too weak to capture all relevant properties of an argumentation framework. For example, an argumentation framework with three arguments, each attacking the next one in the sequence $a \longrightarrow b \longrightarrow c$ has a defense graph where a defends c, but nothing is said about b. If we represent the defense relation by a double arrow, then the defense graph may be visualized by $a \Longrightarrow c \qquad b$

We cannot take this defense graph as the basis for formal argumentation, because it is no longer clear whether argument b can be accepted or not. Thus, a defense graph represents some information about argumentation frameworks, but not everything.

The second observation concerning the notion of defense in formal argumentation is that it is not a binary relation over arguments, like the attack relation is a binary relation over arguments, but it is a relation between a set of arguments and an argument. In this sense, the defense relation is different from the so-called support relation, which is often studied in abstract argumentation.

The following definition of defense graph deals with these two issues in the following way. First, a defense graph is defined relatively to an argumentation framework. Thus, it is not meant to replace the attack relation, but it is used in addition to it. Also, we consider arguments defended by the empty set, i.e. arguments which are not attacked (called initial arguments in graph theory). Second, whereas a set of arguments S defends argument b when it attacks *all* attackers of b, we say that a defends b when a attacks *some* attacker of b. In defense graphs, we are thus slightly abusing the word "defense" for a similar but distinct notion. We could distinguish the two notions by writing $\forall$defends and $\exists$defends, but since the difference is always clear from context, we prefer to overload the concept of defense.

Definition 9 (Weak defense) *Let $F = (A, \rightarrow)$ be an argumentation framework. For $a, b \in A$,*

- *$\langle a, b\rangle$ is a weak defense if and only if $\exists c \in A$ such that $a \rightarrow c$ and $c \rightarrow b$.*
- *$\langle \emptyset, b\rangle$ is a weak defense iff b is initial.*

The set of weak defenses of F is denoted as DEF_F. Given a weak defense $\langle a, b\rangle$ or $\langle \emptyset, b\rangle \in \mathrm{DEF}_F$, we call a the *defender*, and b the *defendee*, of the defense. Given a set $D \subseteq \mathrm{DEF}_F$, we write $\mathrm{defendee}(D) = \{b \mid \langle a, b\rangle, \langle \emptyset, b\rangle \in D\}$ to denote the set of defendees in D, $\mathrm{defender}(D) = \{a \mid \langle a, b\rangle \in D\}$ to denote the set of defenders in D, and $\mathrm{arg}(D) = \mathrm{defendee}(D) \cup \mathrm{defender}(D)$ be the set of arguments who are defendees and defenders in D.

We now define the attacks of the meta-argumentation framework, which are attacks between weak defenses.

Definition 10 (Attacks between weak defenses) *For all $\langle x, a\rangle$, $\langle y, b\rangle \in \mathrm{DEF}_F$ where $x, y \in A \cup \{\emptyset\}$ and $a, b \in A$, we say that $\langle x, a\rangle$ attacks $\langle y, b\rangle$, denoted as $\langle x, a\rangle \rightarrow \langle y, b\rangle$ iff $x \rightarrow y$, or $x \rightarrow b$, or $a \rightarrow y$, or $a \rightarrow b$.*

The set of attacks between weak defenses and their defeaters is denoted as $\rightarrow_F$. We call $\mathrm{DG}_F = (\mathrm{DEF}_F, \rightarrow_F)$ a defense graph. Given an extension $\mathcal{E}$ of F under Dung's argumentation semantics, let $\mathrm{defense}(\mathcal{E}) = \{\langle x, y\rangle \in \mathrm{DEF}_F \mid x \in \mathcal{E} \cup \{\emptyset\}, y \in \mathcal{E}\}$.

We have the following proposition, corresponding to Theorems 1, 2 and Corollaries 1, 2 in [13] with slightly modified notations.

Proposition 7 *Given $F = (A, \rightarrow)$ and its defense graph $\mathrm{DG}_F = (\mathrm{DEF}_F, \rightarrow_F)$, it holds that $\forall D \in \sigma(\mathrm{DG}_F)$, $\arg(D) \in \sigma(F)$; and $\forall \mathcal{E} \in \sigma(F)$, $\mathrm{defense}(\mathcal{E}) \in \sigma(\mathrm{DG}_F)$.*

Definition 11 *Given $F = (A, \rightarrow)$ and its defense graph $\mathrm{DG}_F = (\mathrm{DEF}_F, \rightarrow_F)$, $\forall D \in \sigma(\mathrm{DG}_F)$, let $E = \{a^{R_a} \mid \langle x, a\rangle \in D\}$ where $R_a = \{b \mid \langle b, a\rangle \in D\} \setminus \{\emptyset\}$. We call E a Direct explanation extension. The set of Direct explanation extensions is denoted $\mathrm{Direct}(F)$.*

Proposition 8 *Direct explanation semantics satisfies Properties UAID.*

Proof According to Definition 11, Properties UAID hold by definition. □

In this paper, we view a defense as a transitive relation, i.e., if $\langle a, b\rangle$ and $\langle b, c\rangle$ then $\langle a, c\rangle$. Based on this notion, we have the following definition.

Definition 12 *Given $F = (A, \rightarrow)$ and its defense graph $\mathrm{DG}_F = (\mathrm{DEF}_F, \rightarrow_F)$, $\forall D \in \sigma(\mathrm{DG}_F)$, let D^* be the transitive closure of D. let $E = \{a^{R_a} \mid \langle x, a\rangle \in D\}$ where $R_a = \{a \mid \langle a, a\rangle \in D^*\} \cup \{b \mid \langle b, a\rangle \in D^*, \langle b, b\rangle \in D^*\}$. We call E a Root explanation extension. The set of Root explanation extensions is denoted $\mathrm{Root}(F)$.*

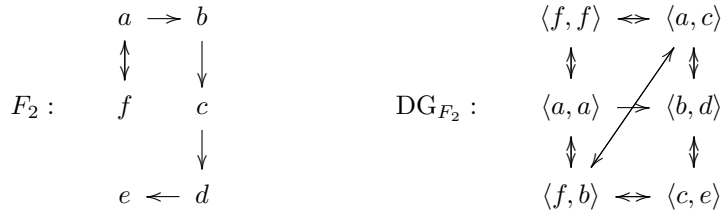
Proposition 9 *Root explanation semantics satisfies Properties UAITE.*

Proof First, since for each $a^{R_a} \in E$, R_a is unique, Property Uniqueness is satisfied. Second, according to Proposition 7, it holds that if $\langle a, b\rangle \in D$ then there exists $\langle c, a\rangle \in D$. So, in terms of Definition 12, $R_a \subseteq \arg(D)$ and $\{b \in E \mid b^{R_b}\} = \arg(D)$. So, $R_a \subseteq \{b \in E \mid b^{R_b}\}$, and Property Acceptance is satisfied. Third, according to Definition 12, $a \in c_\infty(R_a, F)$, and therefore Properties Indirect Defense hold. Fourth, for all $a^{R_a}, b^{R_b} \in E$, if $b \in R_a$, assume that $R_b \not\subseteq R_a$. Then, exists $c \in R_b$ such that $c \notin R_a$. So, $\langle c, a\rangle \notin D^*$. Since when $b \neq a \neq c$, $\langle b, a\rangle \in D^*$ and $\langle c, b\rangle \in D^*$. As a result, $\langle c, a\rangle \in D^*$. Contradiction. So, Property Transitivity holds. Fifth, if $a = b$, then let $c^S = a^{R_a}$. In this case, Property Explanation Inheritance holds. Otherwise, $a \neq b$. In this case, $\langle b, a\rangle \in D^*$ and $\langle b, b\rangle \in D^*$. Let $c^S = b^{R_b}$ where $b \in R_b$. Property Explanation Inheritance also holds. □

Note that Root explanation semantics does not satisfy Properties Direct defense and Minimality, as illustrated by the following examples.

Example 11 *Given F_1 and DG_{F_1} below, under preferred semantics, there are two extensions of DG_{F_1}: $D_1 = \{\langle a, c\rangle, \langle c, e\rangle, \langle e, a\rangle\}$, $D_2 = \{\langle b, d\rangle, \langle d, f\rangle, \langle f, b\rangle\}$. So, $D_1^* = D_1 \cup \{\langle a, e\rangle, \langle a, a\rangle, \langle c, a\rangle, \langle c, c\rangle, \langle e, c\rangle, \langle e, e\rangle\}$ and $D_2^* = D_2 \cup \{\langle b, f\rangle, \langle b, b\rangle, \langle d, b\rangle, \langle d, d\rangle, \langle f, d\rangle, \langle f, f\rangle\}$. So, we have two Root explanation extensions: $E_1 = \{a^{\{a,c,e\}}, c^{\{a,c,e\}}, c^{\{a,c,e\}}\}$, and $E_2 = \{b^{\{b,f,d\}}, f^{\{b,f,d\}}, d^{\{b,f,d\}}\}$, which do not satisfy Property Minimality.*

Example 12 *Given* F_2 *and* DG_{F_2} *below, under preferred semantics, there are two extensions of* DG_{F_1}*:* $D_1 = \{\langle f, f\rangle, \langle f, b\rangle, \langle b, d\rangle\}$, $D_2 = \{\langle a, a\rangle, \langle a, c\rangle, \langle c, e\rangle\}$. *So,* $D_1^* = D_1 \cup \{\langle f, d\rangle\}$ *and* $D_2^* = D_2 \cup \{\langle a, e\rangle\}$. *So, we have two Root explanation extensions:* $E_1 = \{f^{\{f\}}, b^{\{f\}}, d^{\{f\}}\}$, *and* $E_2 = \{a^{\{a\}}, c^{\{a\}}, e^{\{a\}}\}$, *which do not satisfy Property Direct defense.*



6. Conclusions and future work

We study explanation semantics of argumentation by using a principle-based approach. More specifically, in this paper we introduce the explanation principles Uniqueness, Acceptance, Indirect defense, Direct defense, Minimality, Transitivity, Explanation Inheritance. Furthermore, we define various examples of explanations of traditional abstract argumentation semantics. In further work, the formal approach in this paper needs to be extended to informal argumentation as well [3,7].

The work in this paper can be further developed in many ways for both the principles and the explanation semantics. For example, instead of only explaining why an argument is accepted, we can also explain why it is rejected. Explanations can be restricted to core arguments or to representations [14]. Explanation semantics can be combined with, for example, labeling semantics and ranking semantics can be used to rank explanations as well. Moreover, support relations or numerical arguments or attacks can be used to define more sophisticated notions of explanation. The abstract theory of explanation can be further developed for structured argumentation. For example, explanation arguments can refer to evidence or to ethical or legal principles. We believe that such a study of explanation in structured argumentation can also inspire new theories of explanation in abstract argumentation.

More concepts from the general theory of explanation [15] can be studied in formal argumentation, and a general theory of explanation for abstract argumentation can be developed, combining explanation semantics with other notions of explanation in formal argumentation, for example in dialogue [4]. A striking similarity between both is that the notion of defense plays a central role, and such a unified theory of argumentation explanation may lead to a more formal argumentation in which attack and defense are at par. This may also bring the theory of formal argumentation closer to theories of attack and defense in other disciplines such as security [12,10] and in biology [16,2].

Acknowledgement

This material is based in part upon work supported by the "2030 Megaproject" - New Generation Artificial Intelligence of China under Grant No. 2018AAA0100904, the Nat-

ural Science Foundation of Zhejiang Province under Grant No. LY20F030014, and the National Social Science Foundation of China No.18ZDA290 and No.17ZDA026.

References

[1] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert van der Torre, editors. *Handbook of formal argumentation*, volume 1. College Publications, 2018.

[2] Howard Barringer, Dov M. Gabbay, and John Woods. Temporal dynamics of support and attack networks: From argumentation to zoology. In *Mechanizing Mathematical Reasoning, Essays in Honor of Jörg H. Siekmann on the Occasion of His 60th Birthday*, pages 59–98, 2005.

[3] Marcos Cramer and Mathieu Guillaume. Empirical study on human evaluation of complex argumentation frameworks. In *JELIA 2019*, pages 102–115, 2019.

[4] Kristijonas Cyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Syst. Appl.*, 127:141–156, 2019.

[5] Kristijonas Cyras, Ken Satoh, and Francesca Toni. Explanation for case-based reasoning via abstract argumentation. In *COMMA 2016*, pages 243–254, 2016.

[6] Jérémie Dauphin, Marcos Cramer, and Leendert W. N. van der Torre. Abstract and concrete decision graphs for choosing extensions of argumentation frameworks. In *COMMA 2018*, pages 437–444, 2018.

[7] Jérôme Delobelle and Serena Villata. Interpretability of gradual semantics in abstract argumentation. In *ECSQARU 2019*, pages 27–38, 2019.

[8] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[9] Xiuyi Fan and Francesca Toni. On explanations for non-acceptable arguments. In *TAFA 2015*, pages 112–127, 2015.

[10] Dov Gabbay, Ross Horne, Sjouke Mauw, and Leendert van der Torre. Argumentation-based semantics for attack-defense networks. In *Proceedings of The Seventh International Workshop on Graphical Models for Security (GRAMSEC2020)*, 2020.

[11] Alejandro Javier García, Nicolás D. Rotstein, and Guillermo Ricardo Simari. Dialectical explanations in defeasible argumentation. In *ECSQARU 2007*, pages 295–307, 2007.

[12] Barbara Kordy, Sjouke Mauw, Sasa Radomirovic, and Patrick Schweitzer. Foundations of attack-defense trees. In Pierpaolo Degano, Sandro Etalle, and Joshua D. Guttman, editors, *FAST 2010*, volume 6561 of *Lecture Notes in Computer Science*, pages 80–95. Springer, 2010.

[13] Beishui Liao and Leendert W. N. van der Torre. Defense semantics of argumentation: encoding reasons for accepting arguments. *CoRR*, abs/1705.00303, 2017.

[14] Beishui Liao and Leendert W. N. van der Torre. Representation equivalences among argumentation frameworks. In *COMMA 2018*, pages 21–28, 2018.

[15] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[16] Sergio M. Pellis and Vivien C. Pellis. Differential rates of attack, defense, and counterattack during the developmental decrease in play fighting by male and female rats. *Developmental Psychobiology*, 23, 1990.

[17] Tjitze Rienstra, Matthias Thimm, Beishui Liao, and Leendert W. N. van der Torre. Probabilistic abstract argumentation based on SCC decomposability. In *KR 2018*, pages 168–177, 2018.

[18] Dunja Seselja and Christian Straßer. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217, 2013.

[19] L. Richard Ye and Paul E. Johnson. The impact of explanation facilities in user acceptance of expert system advice. *MIS Quarterly*, 19(2):157–172, 1995.

[20] Zhiwei Zeng, Chunyan Miao, Cyril Leung, Zhiqi Shen, and Jing Jih Chin. Computing argumentative explanations in bipolar argumentation frameworks. In *AAAI 2019*, pages 10079–10080, 2019.