Labellings and Games for Extended Argumentation Frameworks

Sanjay Modgil
Department of Computer Science
Imperial College, London
sanjaymodgil@yahoo.co.uk

Abstract

Dung's abstract theory of argumentation has become established as a general framework for various species of non-monotonic reasoning, and reasoning in the presence of conflict. A Dung framework consists of arguments related by attacks, and the extensions of a framework, and so the status of arguments, are defined under different semantics. Developments of Dung's work have also defined argument labellings as an alternative way of characterising extensions, and dialectical argument game proof theories for establishing the status of individual arguments. Recently, Extended Argumentation Frameworks extend Dung's theory so that arguments not only attack arguments, but attacks themselves. In this way, the extended theory provides an abstract framework for principled integration of meta-level argumentation about defeasible preferences applied to resolve conflicts between object level arguments. In this paper we formalise labellings and argument games for a selection of Dung's semantics defined for the extended frameworks.

1 Introduction

Argumentation theory has a wide range of application in Artificial Intelligence [Bench-Capon and Dunne, 2007], including formalisation of non-monotonic reasoning, decision making over action, and negotiation and persuasion dialogues. Much of this work builds on Dung's seminal theory of argumentation [Dung, 1995]. A Dung *argumentation framework* (*DF*) is a directed graph consisting of a set of arguments $\mathcal{A}$ related by a binary conflict based *attack* relation $\mathcal{R}$. The extensions, and so the justified status of their contained arguments, are then defined under different semantics. Extensions are defined based on the acceptability of arguments w.r.t sets of arguments; argument x is acceptable w.r.t $S \subseteq \mathcal{A}$ if any y that attacks x is itself attacked by some $z \in S$. Thus, the core *admissible* semantics defines an admissible extension as a subset S of $\mathcal{A}$, all of whose contained arguments are acceptable w.r.t. S, and an extension under the *preferred* semantics is a set inclusion maximal admissible extension.

The widespread influence of Dung's work can be attributed to its abstract nature. The underlying logic, and definition of the logic's constructed arguments $\mathcal{A}$ and relation $\mathcal{R}$ is left unspecified, thus enabling instantiation of a framework by various logical formalisms. A theory's inferences are then defined as the claims of the justified arguments constructed from the theory (an argument essentially being a proof of a candidate inference - the argument's claim - in the underlying logic). Dung's theory thus provides a general framework for non-monotonic reasoning, and indeed, many logic programming formalisms and non-monotonic logics (e.g. default, auto-epistemic, and defeasible logics) have been shown to conform to Dung's semantics (e.g., in [Dung, 1995]).

Dung's extensional semantics may yield multiple extensions, raising the problem of how to choose between conflicting arguments in different extensions. The problem has been addressed by applying preferences to determine the success of attacks. For example, if x and y attack each other, then each is contained in a distinct preferred extension. However, given a preference for x over y, then y's attack on x does not succeed and we are left with x asymmetrically attacking y, and so $\{x\}$ is the unique preferred extension. Thus, Dung's framework has been augmented with a preference ordering on arguments [Amgoud and Cayrol, 2002], and in *value based argumentation* [Bench-Capon, 2003], y's attack on x does not succeed if the value promoted by x is ranked higher than y's value, according to some given value ordering. However, one often needs to reason, and indeed argue *about*, as well as *with*, defeasible and possibly conflicting preference information. Hence, [Modgil, 2009] has recently extended Dung's theory to integrate 'metalevel' argumentation about preferences between arguments. The extended theory preserves the abstract nature of Dung's approach; no assumptions are made about the structure of arguments expressing preferences, and application of preferences is abstractly characterised, by defining a new attack relation that originates from a preference argument, and that *attacks an attack* between the arguments that are the subject of the preference claim. A new notion of acceptability is defined for the extended theory, and the extensions of an *Extended Argumentation Framework* (*EAF*) are then defined in the same way as for Dung frameworks.

The extensions of a *DF* can equivalently be defined in terms of labellings assigned to arguments [Caminada, 2007; Verheij, 2007], and this approach had led to development

of algorithms for computing the extensions of a *DF* [Caminada, 2007]. The inherently dialectical nature of argumentation has also led to formulation of argument game proof theories, in which a proponent attempts to show that an argument is justified by countering attacking arguments moved by an opponent (e.g., [Cayrol *et al.*, 2003; Modgil and Caminada, 2009]). This work has also underpinned algorithm development [Vreeswijk, 2006], and development of general frameworks for conflict resolution and persuasion dialogues [Prakken, 2005].

This paper formalises labellings and argument games for *EAF*s, and thus establishes foundations for development of algorithms for *EAF*s. In Section 2 we review Dung's argumentation theory and the extended theory. Sections 3, 4 and 5 then describe the three main contributions of this paper: 1) Section 3 defines the admissible, preferred and stable labellings of an *EAF*, and states soundness and completeness results with the acceptability based definitions. Our approach builds on the work of [Caminada, 2007] by additionally assigning labels to attacks; 2) Section 4 defines a dialectical framework for argument games for the extended theory. The framework generalises existing frameworks for argument games (e.g. [Cayrol *et al.*, 2003]) to additionally allow players to move arguments that attack attacks; 3) Section 5 defines a specific game for deciding membership of admissible and preferred extensions of an *EAF*. Finally, Section 6 concludes and discusses future work.

2 Extended Argumentation Frameworks

A *Dung argumentation framework* (*DF*) [Dung, 1995] is a tuple $(\mathcal{A}, \mathcal{R})$, where $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation on the arguments in $\mathcal{A}$. An argument $x \in \mathcal{A}$ is then said to be acceptable w.r.t. some $S \subseteq \mathcal{A}$ iff $\forall y$ s.t. $(y, x) \in \mathcal{R}$ implies $\exists z \in S$ s.t. $(z, y) \in \mathcal{R}$ (i.e., z *reinstates* x). We now recall Dung's definition of extensions under the admissible, preferred and stable semantics[1], in which we refer to a set of arguments as *conflict free* iff $\forall x, y \in S$, (x, y), $(y, x) \notin \mathcal{R}$:

Definition 1 Let $(\mathcal{A}, \mathcal{R})$ be a *DF*, and S a conflict free subset of $\mathcal{A}$. Then:

- S is an *admissible* extension iff every argument in S is acceptable w.r.t. S
- S is a preferred extension iff it is a set inclusion maximal admissible extension
- S is a stable extension iff $\forall y \notin S$, $\exists x \in S$ such that $(x, y) \in \mathcal{R}$

An argument is *sceptically preferred (stable)* justified if it belongs to all preferred (stable) extensions, and only *credulously preferred (stable)* justified if it belongs to at least one, but not all, preferred (stable) extensions.

We now recall the extended argumentation theory [Modgil, 2009]. By way of motivation, consider individuals **P** and **O** exchanging arguments a, $b \ldots$ about the weather forecast:

P : "Today will be dry since the BBC forecast sunshine" = a
O : "Today will be wet since CNN forecast rain" = b
P : "But the BBC are more trustworthy than CNN" = c

[1]Grounded semantics for Dung's and the extended theory will be discussed in Section 6

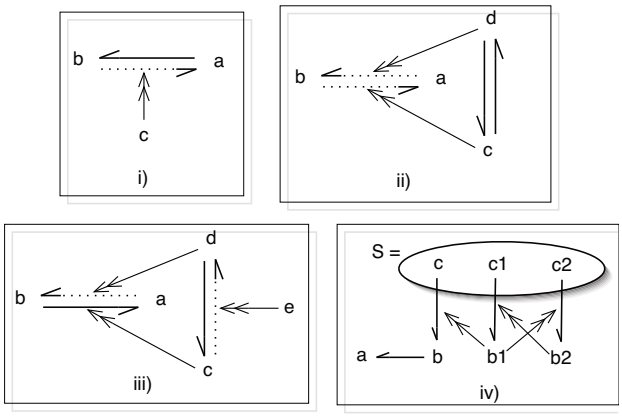

Figure 1: Motivating *EAF*s

O : "However, statistics show that CNN are more accurate than the BBC" = d
O : "And a statistical comparison is more rational than a comparison based on instincts about relative trustworthiness" = e

Arguments a and b symmetrically attack ((a, b),$(b, a) \in \mathcal{R}$), yielding the admissible and preferred extensions $\{a\}$ and $\{b\}$. To choose amongst the two credulously justified arguments, so that one is sceptically justified at the expense of the other, one can incorporate 'metalevel' arguments expressing preferences over other arguments. Thus, c is an *argument* claiming that a is preferred to b. Intuitively, c is an argument for a's repulsion of b's attack on a, i.e., c **attacks** *b's attack on a*[2] so that b's attack on a does not succeed and we are left only with a successfully attacking b (see Figure 1i) in which we introduce the notation $y \rightharpoonup x$ for an attack, and $z \twoheadrightarrow (y \rightharpoonup x)$ for an attack on an attack). Now $\{c, a\}$ is the only preferred extension and so a is sceptically justified. d claims b is preferred to a and so attacks a's attack on b. Now $\{c, a\}$ and $\{d, b\}$ are preferred since the choice between a and b is unresolved given that c and d claim contradictory preferences and so c and d attack each other (Figure 1ii)). However e then attacks the attack from c to d (Figure 1iii)), and so d successfully attacks c, b successfully attacks a, and the discussion concludes in favour of b ($\{e, d, b\}$ is the single preferred extension). *Extended Argumentation Framework*s (*EAF*s) thus extend Dung frameworks with a second attack relation $\mathcal{D}$ from arguments to attacks. If $(z, (x, y)) \in \mathcal{D}$ then z is an argument for preferring y to x, and if any two such *preference arguments* express contradictory preferences, then they attack each other.

Definition 2 An *Extended Argumentation Framework* is a tuple $(\mathcal{A}, \mathcal{R}, \mathcal{D})$, where $\mathcal{A}$ is a set of arguments, $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$, and:

- $\mathcal{D} \subseteq \mathcal{A} \times \mathcal{R}$
- If $(z, (x, y))$, $(z', (y, x)) \in \mathcal{D}$ then (z, z'), $(z', z) \in \mathcal{R}$

The notion of a successful attack, from hereon referred to as a *defeat*, is then parameterised w.r.t. preferences specified by some given set S of arguments:

[2][Modgil, 2009] discusses why it is not appropriate to consider c as directly attacking b.

Definition 3 y defeats$_S$ x, denoted $y \rightarrow^S x$, iff $(y, x) \in \mathcal{R}$ and $\neg\exists z \in S$ s.t. $(z,(y, x)) \in S$.

In the weather example, a defeats$_\emptyset$ b but does not defeat$_{\{d\}}$ b.

A conflict free set of arguments is then defined to account for the case where y *asymmetrically* attacks x, but given a preference for x over y, both may appear in a conflict free set and hence an extension (as in citeBC03). Notice that a conflict free set does not admit arguments that symmetrically attack, irrespective of the preference arguments contained.

Definition 4 S is conflict free iff $\forall x, y \in S$: if $(y, x) \in \mathcal{R}$ then $(x, y) \notin \mathcal{R}$, and $\exists z \in S$ s.t. $(z,(y, x)) \in \mathcal{D}$.

The acceptability of an argument x w.r.t. a set S is now defined for an *EAF*. The basic idea is that for any attacker y of x, a reinstating attack $z \rightharpoonup y$ from $z \in S$, must be reinstated against preference argument attacks on $z \rightharpoonup y$. The definition is motivated in more detail in [Modgil, 2009] and relates to an intuitive requirement (captured by Dung's fundamental lemma in [Dung, 1995]) on what it means for an argument to be acceptable w.r.t. an admissible set S of arguments: *if x is acceptable with respect to S, then $S \cup \{x\}$ is admissible*. To ensure satisfaction of this requirement, acceptability for *EAF*s requires the notion of a *reinstatement set* for a defeat.

Definition 5 Let $S \subseteq \mathcal{A}$ in ($\mathcal{A}$, $\mathcal{R}$, $\mathcal{D}$). Let $R_S = \{x_1 \rightarrow^S y_1, \ldots, x_n \rightarrow^S y_n\}$ where for $i = 1 \ldots n$, $x_i \in S$. Then R_S is a reinstatement set for $a \rightarrow^S b$, iff $a \rightarrow^S b \in R_S$, and
• $\forall x \rightarrow^S y \in R_S$, $\forall y'$ s.t. $(y',(x, y)) \in \mathcal{D}$, $\exists x' \rightarrow^S y' \in R_S$

Definition 6 x is acceptable w.r.t. $S \subseteq \mathcal{A}$ iff $\forall y$ s.t. $y \rightarrow^S x$, $\exists z \in S$ s.t. $z \rightarrow^S y$ and there is a *reinstatement set* for $z \rightarrow^S y$.

In Figure 1iv), a is acceptable w.r.t. S. We have $b \rightarrow^S a$, $c \rightarrow^S b$, and there is a reinstatement set $\{c \rightarrow^S b, c1 \rightarrow^S b1, c2 \rightarrow^S b2\}$ for $c \rightarrow^S b$. Note that if we had $b3 \twoheadrightarrow (c2 \rightharpoonup b2)$, and no argument in S defeating $b3$, there would be no reinstatement set, and a would not be acceptable w.r.t. S.

Given the definitions of conflict free and acceptability for *EAF*s, admissible, preferred and stable semantics for *EAF*s are now defined as for *DF*s in Definition 1 (except that x defeats$_S$ y replaces $(x, y) \in \mathcal{R}$). [Modgil, 2009] shows that *EAF*s inherit many of the fundamental results that hold for *DF*s; in particular, Dung's fundamental lemma holds, and for each admissible S there exists a preferred extension S' such that $S \subseteq S'$.

Extended argumentation has been proposed as a general framework for non-monotonic logics that accommodate defeasible reasoning about priorities on rules. For example, [Modgil, 2009] shows that the inferences from logic programming theories with defeasible priorities [Prakken and Sartor, 1997] correspond to the grounded extension of the EAFs they instantiate. Furthermore, unlike [Prakken and Sartor, 1997], one can provide a well founded definition of the admissible and preferred extensions of such theories. The extended theory is also proposed as a unifying framework for formalising and extending works augmenting Dung frameworks with preferences and values [Modgil and Bench-Capon, 2008; Modgil, 2009], as a semantics for adaptive agent defeasible reasoning and conflict resolution [Modgil, 2007], and for conflict resolution in normative systems [Modgil and Luck, 2008]. Figure 2i) shows an *EAF* for argumentation over a course of medical action (logical formalisms for constructing these arguments are described in [Modgil, 2006]). $a1$ and $a2$ are arguments for prescribing drugs aspirin and chlopidogrel respectively, given that both realise a treatment goal to reduce blood clotting. $b1$ and $b2$ are arguments (based on clinical trials 1 and 2 respectively) expressing the contradictory conclusions that chlopidogrel is more efficacious than aspirin, and aspirin is more efficacious than chlopidogrel. $c1$ claims that trial 1 is more statistically robust than trial 2. $a3$ claims that chlopidogrel is costly, and so attacks $a2$, and $b3$ expresses that the value of improving patient health (promoted by $a2$) is greater than $a3$'s value of cost. $b4$'s contradictory value preference for cost over health mutually attacks $b3$, and finally, $c3$ is a utilitarian argument preferring $b4$ to $b3$ on the grounds that the cost of using chlopidogrel will compromise treatment of other patients. The *EAF* has a single preferred extension $\{c1, b1, c2, b4, a3, a1\}$; aspirin is the preferred choice.

3 Labellings for *EAF*s

This section builds on the work of [Caminada, 2007], and formalises labellings that characterise the admissible, preferred and stable extensions of an *EAF*. A labelling assigns exactly one label to each argument; either IN, OUT or UNDEC. The arguments labelled IN constitute an extension E under some given semantics, and the rules for deciding that an argument is legally IN intuitively correspond to deciding the acceptability of these arguments as defined in Section 2. OUT arguments are defeated$_E$ by arguments in E, and an argument is UNDEC if it is neither in the extension or defeated$_E$ by an argument in the extension. For *EAF*s, attacks on attacks and reinstatement of attacks may decide the acceptability of arguments. Hence, labels must also assigned to attacks in $\mathcal{R}$, so that if $(x, y) \in \mathcal{R}$ is IN, respectively OUT, then this denotes that the attack (x, y) is successful, respectively unsuccessful. Finally, attacks can also be assigned UNDEC.

Definition 7 A labelling for an *EAF*$(\mathcal{A}, \mathcal{R}, \mathcal{D})$ is a pair of total functions $(\mathcal{L}_\mathcal{A}, \mathcal{L}_\mathcal{R})$ such that:

1. $\mathcal{L}_\mathcal{A} : \mathcal{A} \mapsto \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$
2. $\mathcal{L}_\mathcal{R} : \mathcal{R} \mapsto \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$

For $\texttt{S} \in \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$: $\texttt{s}(\mathcal{L}_\mathcal{A}) = \{x | \mathcal{L}_\mathcal{A}(x) = \texttt{S}\}$; $\texttt{s}(\mathcal{L}_\mathcal{R}) = \{(x, y) | \mathcal{L}_\mathcal{R}((x, y)) = \texttt{S}\}$

We now define the notion of a *legal* labelling:

Definition 8 Let $\mathcal{L} = (\mathcal{L}_\mathcal{A}, \mathcal{L}_\mathcal{R})$ be a labelling for $(\mathcal{A}, \mathcal{R}, \mathcal{D})$. $\forall x \in \mathcal{A}$:

1. $x \in \texttt{out}(\mathcal{L}_\mathcal{A})$ is *legally* OUT iff $\exists (y, x) \in \mathcal{R}$ s.t. $\mathcal{L}_\mathcal{A}(y)$ = IN and $\mathcal{L}_\mathcal{R}((y, x))$ = IN.
2. $x \in \texttt{in}(\mathcal{L}_\mathcal{A})$ is *legally* IN iff $\forall (y, x) \in \mathcal{R}$, either $\mathcal{L}_\mathcal{A}(y)$ = OUT or $\mathcal{L}_\mathcal{R}((y, x))$ = OUT.
3. $x \in \texttt{undec}(\mathcal{L}_\mathcal{A})$ is *legally* UNDEC iff:
   (a) $\neg\exists (y, x) \in \mathcal{R}$ such that $\mathcal{L}_\mathcal{A}(y)$ = IN and $\mathcal{L}_\mathcal{R}((y, x))$ = IN, and;
   (b) it is not the case that: $\forall y \in \mathcal{A}$, $(y, x) \in \mathcal{R}$ implies $\mathcal{L}_\mathcal{A}(y)$ = OUT or $\mathcal{L}_\mathcal{R}((y, x))$ = OUT

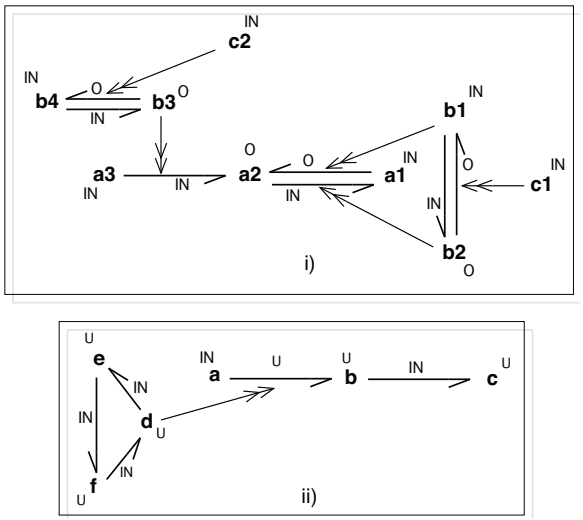

Figure 2: *EAF*s and their labellings (O stands for OUT and U stands for UNDEC)

$\forall (y, x) \in \mathcal{R}$:

1. $(y, x) \in \texttt{out}(\mathcal{L}_\mathcal{R})$ is *legally* OUT iff $\exists$ $(z,(y, x)) \in \mathcal{D}$ s.t. $\mathcal{L}_\mathcal{A}(z)$ = IN
2. $(y, x) \in \texttt{in}(\mathcal{L}_\mathcal{R})$ is *legally* IN iff $\forall$ $(z,(y, x)) \in \mathcal{D}$, $\mathcal{L}_\mathcal{A}(z)$ = OUT
3. $(y, x) \in \texttt{undec}(\mathcal{L}_\mathcal{R})$ is *legally* UNDEC iff
   (a) $\neg\exists$ $(z,(y, x)) \in \mathcal{D}$ s.t. $\mathcal{L}_\mathcal{A}(z)$ = IN
   (b) it is not the case that: $\forall z \in \mathcal{A}$, $(z,(y, x)) \in \mathcal{D}$ implies $\mathcal{L}_\mathcal{A}(z)$ = OUT

For $\texttt{S} \in \{\texttt{IN}, \texttt{OUT}, \texttt{UNDEC}\}$:

- An argument x is said to be illegally S iff $x \in \texttt{s}(\mathcal{L}_\mathcal{A})$, and it is not legally S.
- An attack (y, x) is said to be illegally S iff $(y, x) \in \texttt{s}(\mathcal{L}_\mathcal{R})$, and it is not legally S.

Note, it is straightforward to show that $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{R}$ assign exactly one label to each argument, respectively attack. Also observe, that by definition, an argument or attack that is not attacked cannot be legally UNDEC. Also, Definition 8 implies that an attack (y, x) is legally UNDEC iff there exists at least one UNDEC labelled argument z that attacks (y, x), and no z' attacking (y, x) is labelled IN. Similarly, an argument x is legally UNDEC iff there exists at least one $(y, x) \in \mathcal{R}$ such that y or (y, x) are UNDEC, and there is no (y, x) such that y and (y, x) are IN.

We now define admissible, preferred and stable *EAF* labellings and state a correspondence with the extensions as defined in Section 2.

Definition 9 Let $\mathcal{L} = (\mathcal{L}_\mathcal{A}, \mathcal{L}_\mathcal{R})$ be a labelling for $(\mathcal{A}, \mathcal{R}, \mathcal{D})$.

- $\mathcal{L}$ is *admissible* iff :
  1. no $x \in \mathcal{A}$ is illegally IN or illegally OUT
  2. no $(y, x) \in \mathcal{R}$ is illegally IN or illegally OUT
  3. $\forall x, y \in \texttt{in}(\mathcal{L}_\mathcal{A})$, it is not the case that $(y, x) \in \mathcal{R}$ and $(x, y) \in \mathcal{R}$
- $\mathcal{L}$ is *prefered* iff $\mathcal{L}$ is *admissible* and there does not exist an admissible $\mathcal{L}'$ such that $\texttt{in}(\mathcal{L}'_\mathcal{A}) \supset \texttt{in}(\mathcal{L}_\mathcal{A})$
- $\mathcal{L}$ is *stable* iff $\mathcal{L}$ is *admissible*, and $\texttt{undec}(\mathcal{L}_\mathcal{A}) = \emptyset$, $\texttt{undec}(\mathcal{L}_\mathcal{R}) = \emptyset$.

Theorem 1 Let $\Delta = (\mathcal{A}, \mathcal{R}, \mathcal{D})$ be an *EAF*, and $E \subseteq \mathcal{A}$. For $m \in \{$admissible, preferred, stable$\}$:

E is an m extension of Δ iff there exists an m labelling $(\mathcal{L}_\mathcal{A}, \mathcal{L}_\mathcal{R})$ with $\texttt{in}(\mathcal{L}_\mathcal{A}) = E$

Consider the preferred labelling identifying the preferred extension for the medical example *EAF* in Figure 2i). Consider also the labelled *EAF* in Figure 2ii). It is easy to verify that none of the arguments in the odd loop $f \rightharpoonup d \rightharpoonup e \rightharpoonup f$ can be legally assigned IN or OUT. The undecided status of these arguments 'contaminates' the attack $a \rightharpoonup b$, so that only $\{a\}$ is admissible and preferred, and there does not exist a stable extension. Notice the requirement that admissible labellings require that attacks are legally labelled IN (OUT). Suppose this were not the case, so that $a \rightharpoonup b$ was illegally labelled IN. Then b would be legally OUT and c legally IN. However, $\{a, c\}$ is not admissible since c is not acceptable w.r.t. $\{a, c\}$ (there is no reinstatement set for $a \rightarrow_{\{a,c\}} b$).

4 A dialectical framework for *EAF* games

Argument game proof theories establish the justified status of an argument to be tested, and provide a basis for algorithm development. In this section we define a dialectical framework for *EAF* game proof theories played as dialogues between two players — P (for "proponent") and O (for "opponent") — each of which are referred to as the other's 'counterpart'. We will from hereon assume finite *EAF*s that contain a finite number of arguments. A game begins with P moving an initial argument x to be tested. O and P then take turns in moving arguments that attack their counterpart's last move, where unlike games defined for Dung frameworks, a player can attack either an argument *or an attack* moved by its counterpart.

Definition 10 Let $(\mathcal{A}, \mathcal{R}, \mathcal{D})$ be an *EAF*. A dialogue d is a possibly infinite sequence of moves $m_0, m_1, \ldots$ such that:

- $d_\emptyset$ denotes the empty sequence, m_0 the 'initial move'
- each m_i is of the form $x_{\texttt{Pl}}$ where $x \in \mathcal{A}$ is the argument moved in m_i, denoted by $arg(m_i)$, and $\texttt{Pl} \in \{\texttt{P}, \texttt{O}\}$ is the player of m_i, denoted $pl(m_i)$.
- $pl(m_0)$ = P, and $pl(m_i) \neq pl(m_{i+1})$.
- m_1 *attack replies* m_0, and for $\forall i > 1$, either
  – m_i *attack replies* m_{i-1} (denoted $m_{i-1} \leftharpoonup m_i$), or
  – m_i *pref attack replies* m_{i-1} (denoted $m_{i-1} \twoheadleftarrow m_i$), and m_i does not both *attack and pref attack reply* m_{i-1}, where:
  • m_i *attack replies* m_{i-1} iff $(arg(m_i), arg(m_{i-1})) \in \mathcal{R}$
  • m_i *pref attack replies* m_{i-1} iff $(arg(m_i), (arg(m_{i-1}), arg(m_{i-2})) \in \mathcal{D}$

A finite dialogue $d = m_o - m_1 - \ldots - m_n$ is said to be won by Pl if $pl(m_n)$ = Pl (note that from hereon, if we write $m_{i-1} - m_i$ then '$-$' denotes either $\leftharpoonup$ or $\twoheadleftarrow$).

The rules of the game encode restrictions on the legality of a player's attack on its counterpart's previously moved argument or attack. Different sets of rules capture the different semantics under which justification of the argument moved by P in the initial move is to be shown, by effectively establishing when O or P run out of legal moves, and thus which player wins the dialogue. In what follows, we will refer to a generic legal move function ϕ that places restrictions on players' moves, and dialogues played according to ϕ as ϕ-dialogues.

In general, a player can backtrack to a counterpart's previous move and initiate a new dialogue. Consider the dialogue $a_{\mathtt{P}} \leftharpoondown b_{\mathtt{O}} \leftharpoondown c_{\mathtt{P}} \nleftarrow d_{\mathtt{O}} \leftharpoondown e_{\mathtt{P}} \leftharpoondown f_{\mathtt{O}}$ won by O ($x_{\mathtt{Pl}}$ denotes argument x moved by player Pl). P must then try and backtrack to move an argument against either O's move of b, or the attack $b \rightharpoonup a$, or d, and so try and establish an alternative P winning dialogue (i.e., 'line of defense') for a. Suppose such a dialogue $a_{\mathtt{P}} \leftharpoondown b_{\mathtt{O}} \leftharpoondown g_{\mathtt{P}}$. Then O can backtrack and try an alternative O winning dialogue (i.e., 'line of attack') moving h against a, so that P must now try and win the newly initiated dialogue $a_{\mathtt{P}} \leftharpoondown h_{\mathtt{O}}$.

Thus, a ϕ game that establishes whether x is justified, is a tree of ϕ-dialogues whose root is P's initial move of x, and such that O fully fulfills its burden of attack by moving all ϕ legally allowed replies to each argument and attack moved by P, and P fully fulfills its burden of defense by moving at least one ϕ legally allowed reply to each argument or the associated attack moved by O. If every dialogue in such a game is won by P, then x is shown to be justified. Such a game is defined below as a *winning strategy*, in which we refer to the notion of a *sub-dialogue* d' of a dialogue d, which is any sub-sequence of d that starts with the same initial move as d.

Definition 11 Let $(\mathcal{A}, \mathcal{R}, \mathcal{D})$ be an EAF and T a non-empty finite set of finite ϕ-dialogues with initial move $x_{\mathtt{P}}$. Then T is a ϕ winning strategy for x iff:

1. Each dialogue in T is won by P.
2. $\forall d \in T$, $\forall d'$ such that $d' = m_0 - \ldots - m$ is some sub-dialogue of d and $pl(m) = \mathtt{P}$, then if O can ϕ legally reply to m with m', there is a $d'' \in T$ such that $d' - m'$ is a sub-dialogue of d''.

We now define notation that will be of use when specifying legal move functions:

Notation 1 Let d be a dialogue. For $\mathtt{Pl} \in \{\mathtt{O}, \mathtt{P}\}$:

- $\mathtt{Pl}_{\mathcal{A}}(d) = \{x | arg(m) = x, pl(m) = \mathtt{Pl}, m$ is a move in $d\}$ is the set of arguments moved by Pl in d.
- $\mathtt{Pl}_{\mathcal{R}}(d) = \{(x, y) | arg(m) = x, arg(m') = y, pl(m) = \mathtt{Pl}, m$ attack replies m' in $d\}$ is the set of attacks moved by Pl in d.
- $\mathtt{Pl}_{\mathcal{D}}(d) = \{(z, (x, y)) \mid arg(m) = z, arg(m') = x, pl(m) = \mathtt{Pl}, m$ pref attack replies m' in $d\}$ is the set of pref attacks moved by Pl in d.

5 An argument game for the credulous preferred semantics

In this section we define a legal move function ϕ_{PC} for the preferred credulous game. Since every admissible extension of an *EAF* is a subset of a preferred extension, it suffices to show membership of an admissible extension in order to show membership of a preferred extension.

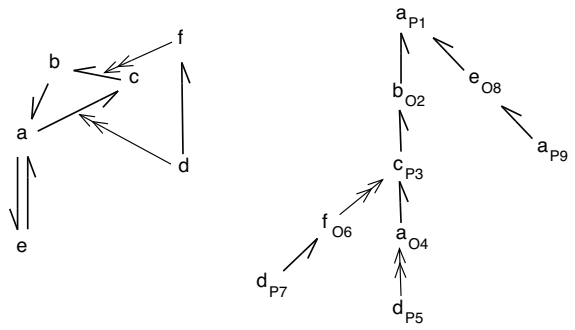

Figure 3: An *EAF*s and ϕ_{PC} winning strategy for a

The function ϕ_{PC} prevents O from moving arguments *and* attacks that have already been attacked by P in a dialogue, since P will have already fulfilled its burden of defense with respect to these arguments / attacks. One need only consider the framework $x \rightleftharpoons y$. Without the restriction on O, P's attempt to show that x is in the admissible $\{x\}$ will result in an infinite dialogue $x_{\mathtt{P}} \leftharpoondown y_{\mathtt{O}} \leftharpoondown x_{\mathtt{P}} \leftharpoondown y_{\mathtt{O}} \ldots$. Since every admissible extension is conflict free, ϕ_{PC} also prevents P from introducing a conflict into the arguments it has already moved in a dialogue. That is to say, P can only move an argument x in d if: 1) x does not attack itself, and; 2) no argument y, *and* attack (y, x) or (x, y) has been moved by P, and; 3) x does not symmetrically attack some y moved by P.

Definition 12 Given $(\mathcal{A}, \mathcal{R}, \mathcal{D})$, and a dialogue d, then $conflict(d) =$
$\{x | (x, x) \in \mathcal{R}\} \cup \{x | \exists y \in \mathtt{P}_{\mathcal{A}}(d), (x, y), (y, x) \in \mathcal{R}\} \cup$
$\{x | \exists y \in \mathtt{P}_{\mathcal{A}}(d), (x, y) \in \mathtt{P}_{\mathcal{R}}(d)$ or $(y, x) \in \mathtt{P}_{\mathcal{R}}(d)\}$

Definition 13 Given $(\mathcal{A}, \mathcal{R}, \mathcal{D})$, and a dialogue d, then ϕ_{PC} is a legal move function such that:

- $\phi_{PC}(d_{\emptyset}) = \mathcal{A} - \{x | (x, x) \in \mathcal{R}\}$ (the initial move by P)
- If d is of odd length (next move is by O), then:
  if $d = x_{\mathtt{P}}$, then $\phi_{PC}(d) = \{y \mid (y, x) \in \mathcal{R}\}$,
  else if $d = d' - z_{\mathtt{O}} - x_{\mathtt{P}}$ then $\phi_{PC}(d) =$
  $\{y \mid (y, x) \in \mathcal{R}, (x', y) \notin \mathtt{P}_{\mathcal{R}}(d), (x', (y, x)) \notin \mathtt{P}_{\mathcal{D}}(d)\}$
  $\cup$
  $\{y \mid (y, (x, z)) \in \mathcal{D}, (x', y) \notin \mathtt{P}_{\mathcal{R}}(d)\}$
- If $d = d' - z_{\mathtt{P}} - x_{\mathtt{O}}$ is of even length (next move is by P), then $\phi_{PC}(d) =$
  $\{y \mid (y, x) \in \mathcal{R}$ or $(y,(x, z)) \in \mathcal{D}$, and $y \notin conflict(d)\}$

One can then show that the following holds:

Theorem 2 Given an *EAF* $\Delta = (\mathcal{A}, \mathcal{R}, \mathcal{D})$, $x \in \mathcal{A}$ is in an admissible extension of Δ iff there exists a ϕ_{PC} winning strategy T for x such that $\bigcup_{d \in T}(\mathtt{P}_{\mathcal{A}}(d))$ is conflict free [3].

[3] [Modgil and Caminada, 2009] discuss why checks for conflict freeness of winning strategies are required

Figure 3 shows an *EAF*, and ϕ_{PC} winning strategy for a in which moves are individuated by numerical indicies indicating the order in which they are played. a is in the admissible and preferred extension $\{a, d, c\}$. Notice that P can move c at 3, since although c is attacked by P's previously moved a, the attack by a on c has not been moved by P. Indeed, $c_{\mathtt{P}3}$ forces O to move the attack of a on c, exposing this attack to P's pref attack with d at 5. Finally, the reader can easily verify that there is a ϕ_{PC} winning strategy for $a1$, and no ϕ_{PC} winning strategy for $a2$, for the EAF in Figure 2i).

6 Conclusions

In this paper we have built on the labelling approach of [Caminada, 2007] in order to formalise a labelling based characterisation of the admissible, preferred and stable extensions of *EAF*s, and thus provided foundations for future development of algorithms for computing extensions of *EAF*s. Future work will also formalise a labelling based characterisation of the grounded extension of an *EAF*. The grounded extension of a Dung framework (*DF*) is the least fixed point of a characteristic function F that takes as input a set S of arguments and returns the set S' of arguments acceptable w.r.t. S. The characteristic function G for *EAF*s is not monotonic, so that a least fixed point cannot be guaranteed. Hence the grounded extension of an *EAF* is defined constructively, by iteration of G, beginning with the empty set (analogous to construction of a *DF*'s grounded extension). This in turn means that the grounded extension of an *EAF* cannot be readily characterised by a labelling that adapts the grounded labellings defined by [Caminada, 2007] for *DF*s[4].

In this paper we have defined a dialectical framework for *EAF* argument game proof theories that determine the justified status of arguments. We have defined a specific game within the framework for the credulous preferred semantics, and future work will focus on specifying a game for the grounded semantics. An advantage of dialectical games is that they relate formal entailment to something most people are familiar with in everyday life: debates and discussions. These games will thus not only provide guidelines for the design of algorithms for computing the justified arguments of an *EAF*, but will also inform development of frameworks for argumentation based negotiation, deliberation and persuasion dialogues in which participants can debate preferences.

References

[Amgoud and Cayrol, 2002] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.

[Bench-Capon and Dunne, 2007] T. J. M. Bench-Capon and P. E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171:10–15, 2007.

[Bench-Capon, 2003] T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[Caminada, 2007] M. Caminada. An algorithm for computing semi-stable semantics. In *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, pages 222–234, 2007.

[Cayrol *et al.*, 2003] C. Cayrol, S. Doutre, and J. Mengin. On Decision Problems Related to the Preferred Semantics for Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):377–403, 2003.

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

[Modgil and Bench-Capon, 2008] S. Modgil and T. J. M. Bench-Capon. Integrating object and meta-level value based argumentation. In *Proc. 2nd Int. Conf. on Computational Models of Argument*, pages 240–251, 2008.

[Modgil and Caminada, 2009] S. Modgil and M. Caminada. Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G. Simari, editors, *Argumentation in AI (in press)*. Springer, 2009.

[Modgil and Luck, 2008] S. Modgil and M. Luck. Argumentation based resolution of conflicts between desires and normative goals. In *Proc. 5th Int. Workshop on Argumentation in Multi-Agent Systems (best paper NorMAS 2009)*, pages 252–263, 2008.

[Modgil, 2006] S. Modgil. Value based argumentation in hierarchical argumentation frameworks. In *Proc. 1st Int. Conf. on Computational Models of Argument*, pages 297–308, 2006.

[Modgil, 2007] S. Modgil. An argumentation based semantics for agent reasoning. In *Proc. Workshop on Languages, methodologies and development tools for multi-agent systems (LADS 07)*, pages 37–53, Durham, UK, 2007.

[Modgil, 2009] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, in press (doi:10.1016/j.artint.2009.02.001) 2009.

[Prakken and Sartor, 1997] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75, 1997.

[Prakken, 2005] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15:1009–1040, 2005.

[Verheij, 2007] H.B. Verheij. A labeling approach to the computation of credulous acceptance in argumentation. In *Proc. 20th Int. Joint Conf. on Artificial Intelligence (IJCAI 2007)*, pages 623–628, 2007.

[Vreeswijk, 2006] G. Vreeswijk. An algorithm to compute minimally grounded and admissible defence sets in argument systems. In *Proc. 1st Int. Conf. on Computational Models of Argument*, pages 109–120, UK, 2006.

[4] However G is monotonic for hierarchical *EAF*s that stratify object and metalevel argumentation. These EAFs are studied in [Modgil and Bench-Capon, 2008] in which a correspondence result is shown with rewrites of hierarchical *EAF*s as *DF*s