A Logical Theory about Dynamics in Abstract Argumentation

Anonymous

No Institute Given

Abstract. We address dynamics in abstract argumentation using a logical theory where an agent's belief state consists of an argumentation framework (AF, for short) and a constraint that encodes the outcome the agent believes the AF *should* have. Dynamics enters in two ways: (1) the constraint is strengthened upon learning that the AF should have a certain outcome and (2) the AF is expanded upon learning about new arguments/attacks. A problem faced in this setting is that a constraint may be inconsistent with the AF's outcome. We discuss two ways to address this problem: First, it is still possible to form consistent *fallback beliefs*, i.e., beliefs that are most plausible given the agent's AF and constraint. Second, we show that it is always possible to find AF expansions to restore consistency. Our work combines various individual approaches in the literature on argumentation dynamics in a general setting.

Keywords: Argumentation, Dynamics, Knowledge Representation

1 Introduction

In Dung-style argumentation [1] the argumentation framework (AF for short) is usually assumed to be static. However, there are many scenarios where argumentation is a dynamic process: Agents may learn that an AF must have a certain outcome and may learn about new arguments/attacks. These are two basic issues that a theory about argumentation dynamics should address.

Some of these aspects have received attention in recent years. For example, the so called *enforcing problem* [2] is concerned with the question of whether and how an AF can be modified to make a certain set of arguments accepted. Other work studies the impact on the outcome of an AF when a new argument comes into play [3] or studies the issue of reasoning with incomplete AFs [4].

We address the problem by answering the following research questions: *How can we model an agent's belief about the outcome of an AF?* and *How can we characterize the effects of an agent learning that the AF should have a certain outcome, or learning about new arguments/attacks?*

The basis of our approach is a logical *labeling language*, interpreted by labelings that assign to each argument a label indicating that it is *accepted, rejected* or *undecided* [5]. Formulas in this language are statements about the acceptance of the arguments of an AF. This allows us to reason about the outcome of an AF in terms of beliefs, rather than extensions or labelings.

We take an agent's belief state to consist of an AF and a formula encoding a constraint on the outcome of the AF. The constraint is strengthened upon learning that the AF should have a certain outcome. Furthermore, the agent's AF is expanded upon learning about new arguments and attacks. These two operations are modeled by a *constraint expansion* and *AF expansion* operator.

A problem faced in this setting is that the constraint on the AF's outcome may be inconsistent with its actual outcome, preventing the agent from forming consistent beliefs. We call such a state *incoherent.* We appeal to the intuition that an AF provides the agent with the ability to argue for the plausibility of the beliefs that it induces. Incoherence thus means that the agent is unable to argue for the plausibility of her beliefs using the AF.

We show that there are two ways to deal with this. First, we show that, given an incoherent belief state, it is always possible to come up with an expansion of the AF that restores coherence. Such AF expansions can be thought of as providing the missing arguments necessary to argue for her beliefs. Second, we show that it is always possible to form consistent *fallback beliefs*, which represent the "most plausible" outcome of the agent's AF, given the constraint.

For the novel notion of fallback belief, we present an encoding of the associated decision problem (i.e., determining whether or not some formula is a fallback belief in a particular belief state) as an answer-set program.

Our theory about argumentation dynamics combines several individual approaches in the literature in a general setting. For example, the issue of restoring coherence is related to the enforcing problem [2]; other ways to characterize the effect of an AF expansion have been studied in [3] and our notion of fallback belief is related to principles developed in [4].

A brief outline of this paper: In section 2 we introduce our labeling logic, together with the necessary basics of argumentation theory. Next, we present our belief state model and associated expansion operators in section 3. We then discuss in sections 4 and 5 how to deal with incoherent belief states, i.e., by restoring coherence via AF expansion and by using fallback belief. In section 6 we present an ASP encoding of the decision problem for fallback belief. Having focused in these sections on the complete semantics, we turn in section 7 to a discussion of a number of additional semantics. In section 8 we discuss related work and we conclude and discuss future work in section 9.

2 Preliminaries

We start out with some preliminaries concerning Dung-style abstract argumentation theory [1]. According to this theory, argumentation can be modeled using an *argumentation framework*, which captures two basic notions, namely arguments and attacks among arguments. We limit ourselves to the abstract setting, meaning that we do not specify the content of arguments in a formal way. Nevertheless, arguments should be understood to consist of a *claim* and a *reason*, i.e., some consideration that counts in favor of believing the claim to be true, while attacks among arguments stem from conflicts between different claims and reasons. We assume in this paper that argumentation frameworks are finite.

Definition 1. *An* argumentation framework *(AF for short) is a pair* $(A, \Rightarrow)$ *where* A *is a finite set of* arguments *and* $\Rightarrow \subseteq A \times A$ *is an* attack *relation.*

Given an AF $(A, \Rightarrow)$ we say that x *attacks* y, or that x is an *attacker* of y, whenever $(x, y) \in \Rightarrow$. The outcome of an AF consists of possible points of view on the acceptability of its arguments. In the literature, these points of view are represented either by sets of acceptable arguments, called *extensions* or by *argument labelings*, which are functions assigning to each argument a label *in*, *out* or *undecided*, indicating that the argument is respectively accepted, rejected or neither [5]. The two representations are essentially reformulations of the same idea, because in most cases, the two can be mapped 1-to-1, such that extensions correspond to sets of in-labeled arguments [5]. For the current purpose we choose to adopt the labeling-based approach.

Definition 2. *A* labeling *of an AF* $F = (A, \Rightarrow)$ *is a function* $L : A \to \{I, O, U\}$*. We denote by* $I(L), O(L)$ *and* $U(L)$ *the set of all arguments* $x \in A$ *such that* $L(x) = I$*,* $L(x) = O$ *or* $L(x) = U$*, respectively, and by* $\mathcal{M}_F$ *the set of all labelings of* F*.*

Various conditions have been proposed to single out labelings that represent actual “rational” points of view. One of the most elementary ones, and the one we focus on in this paper, gives rise to what is called the *complete* semantics.

Definition 3. *Let* $F = (A, \Rightarrow)$ *be an AF and* $L \in \mathcal{M}_F$ *a labeling. We say that* L *is* complete *if and only if for each* $x \in A$ *it holds that:*

- $L(x) = I$ *iff* $\forall y \in A$ *s.t.* $(y, x) \in \Rightarrow$*,* $L(y) = O$*,*
- $L(x) = O$ *iff* $\exists y \in A$ *s.t.* $(y, x) \in \Rightarrow$ *and* $L(y) = I$*,*

Thus, under the complete semantics, the outcome of an AF consists of labelings in which arguments are in if and only if all attackers are out and are out if and only if there is an attacker that is in. The complete semantics is an elementary semantics because many of the other semantics proposed in the literature, such as the *grounded*, *preferred* and *stable* semantics [1] are all based on selecting particular subsets of the set of complete labelings:

Definition 4. *Let* $F = (A, \Rightarrow)$ *be an AF and* L *a complete labeling of* F*.* L *is called:*

- grounded *iff there is no complete labeling* L' *of* F *s.t.* $I(L') \subset I(L)$*,*
- preferred *iff there is no complete labeling* L' *of* F *s.t.* $I(L) \subset I(L')$*,*
- stable *iff* $U(L) = \emptyset$*.*

While we focus in the following sections on the complete semantics, we will briefly discuss the other ones in section 7.

Example 1. Consider the AF shown in figure 1, which has three complete labelings, namely `IOOI`, `OIOI` and `UUUU`. (We denote labelings by strings of the form `ABC...` where `A`, `B`, `C`, ... are the labels of the arguments $a, b, c, \ldots$)

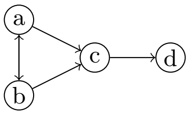

Fig. 1. An argumentation framework

A flexible way to reason about the outcome of an AF is by using a logical *labeling language*. Formulas in this language assign a label to an argument or are boolean combinations of such assignments. The language, given an AF $F = (A, \Rightarrow)$, is denoted by $\mathcal{L}_F$ and is generated by the following BNF, where $x \in A$:

$$\phi := \mathtt{in}_x \mid \mathtt{out}_x \mid \mathtt{u}_x \mid \neg\phi \mid \phi \vee \phi \mid \top \mid \bot.$$

For a set $X \subseteq A$ we use $\mathtt{in}_X$ as shorthand for the conjunction $\wedge_{x \in X} \mathtt{in}_x$ and similarly for $\mathtt{out}_X$ and $\mathtt{u}_X$. We also use the connectives $\wedge, \rightarrow, \leftrightarrow$, defined as usual in terms of $\neg$ and $\vee$. Next, we define a *satisfaction* relation $\models_F$ between labelings of F and formulas in $\mathcal{L}_F$:

Definition 5. *Let F be an AF. The satisfaction relation $\models_F \subseteq \mathcal{M}_F \times \mathcal{L}_F$ is defined by:*
- *$L \models_F \mathtt{in}_x$ iff $L(x) = I$,*
- *$L \models_F \mathtt{out}_x$ iff $L(x) = O$,*
- *$L \models_F \mathtt{u}_x$ iff $L(x) = U$,*
- *$L \models_F \phi \vee \psi$ iff $L \models_F \phi$ or $L \models_F \psi$,*
- *$L \models_F \neg\phi$ iff $L \not\models_F \phi$,*
- *$L \models_F \top$and $L \not\models_F \bot$.*

A model *of a formula ϕ is a labeling $L \in \mathcal{M}_F$ such that $L \models_F \phi$. We denote by $[\phi]_F$ the set of labelings satisfying ϕ, defined by $[\phi]_F = \{L \in \mathcal{M}_F \mid L \models_F \phi\}$. We write $\phi \models_F \psi$ iff $[\phi]_F \subseteq [\psi]_F$ and $\phi \equiv_F \psi$ iff $[\phi]_F = [\psi]_F$.*

Whenever the AF we talk about is clear from the context, we drop the subscript F from $\models_F$, $[\dots]_F$ and $\equiv_F$.

Using this labeling language, we can reason about the outcome of an AF by talking about *beliefs* induced by the AF. These beliefs can be represented by a formula ϕ such that $[\phi]$ is exactly the set of complete labelings of F. It is worth noting that ϕ can be formulated in a straightforward way:

Proposition 1. *Let $F = (A, \Rightarrow)$ be an AF. It holds that a labeling L is a complete labeling of F iff L is a model of the formula*

$$\wedge_{x \in A}\big((\mathtt{in}_x \leftrightarrow (\wedge_{(y,x) \in \Rightarrow} \mathtt{out}_y)) \wedge (\mathtt{out}_x \leftrightarrow (\vee_{(y,x) \in \Rightarrow} \mathtt{in}_y))\big).$$

Example 2. Among the beliefs induced by AF in figure 1 are $\neg\mathtt{out}_d$ and $(\mathtt{in}_a \vee \mathtt{in}_b) \leftrightarrow \mathtt{in}_d$ and $\neg(\mathtt{in}_a \wedge \mathtt{in}_b)$.

Finally, *conflict-freeness* is considered to be a necessary (but not sufficient) condition for any labeling to be considered rational. We will make use of the following definition:

Definition 6. *Let $F = (A, \Rightarrow)$ be an AF. A labeling L of F is said to be* conflict-free *if and only if L is a model of the formula*

$$\wedge_{x \in A}(\mathtt{in}_x \to ((\wedge_{(y,x) \in \Rightarrow} \mathtt{out}_y) \wedge (\wedge_{(y,x) \in \Rightarrow} \mathtt{out}_y))).$$

We denote this formula by Cf_F. A formula ϕ is said to be conflict-free iff $Cf_F \not\models \neg\phi$.

Thus, a labeling is conflict-free if and only if all 'neighbors' of every in-labeled argument are labeled out. It can be checked that every complete labeling is also a conflict-free labeling but not vice versa.

Example 3. Some examples of conflict-free labelings of the AF in figure 1 are `IOOO`, `UUOI` and `OOOO`. Some examples of labelings that are not conflict-free are `IIOO`, `UIOO` and `UUIO`.

3 Belief states

One the one hand, AFs interpreted under the complete semantics induce beliefs about the status of arguments (and, consequently, about argument's claims and reasons) that are rational in the sense that the arguments and attacks in the AF can be used to argue for the plausibility of these beliefs. For example, given the AF $(\{b, a\}, \{(b, a)\})$, the belief $\mathtt{out}_a$ can, informally speaking, be argued for by pointing out that a is attacked by b which, in turn, is not attacked and should thus be accepted. Furthermore, these beliefs are defeasible, because learning about new arguments and attacks may cause old beliefs to be retracted.

On the other hand, an agent may learn, observe or come to desire some claim to be true or false, without being aware of arguments to argue for the plausibility of it. This bears on the outcome that the AF *should* have, according to the agent. To model scenarios like these, we define an agent's belief state to consist not only of an AF, but also a constraint that the agent puts on its outcome.

Definition 7. *A belief state is a pair $S = (F, K)$, where $F = (A, \Rightarrow)$ is an AF and $K \in \mathcal{L}_F$ the agent's* constraint. *We define $K(S)$ by $K(S) = K$ and $Bel(S)$ by $Bel(S) = \phi$ where ϕ is a formula such that $[\phi] = \{L \in [K] \mid L \text{ is a complete labeling of } F\}$. We say that the agent* believes *ψ if and only if $Bel(S) \models \psi$ and that S is* coherent *if and only if $Bel(S) \not\models \bot$.*

Thus, the belief $Bel(S)$ of an agent is formed by the outcome of the AF in conjunction with the constraint. Intuitively, the plausibility of the agent's belief can be argued for only if it is consistent, i.e., only if the belief state is coherent. An incoherent state is thus a state in which the agent is prevented from forming beliefs that can be shown to be plausible via the AF.

We turn again to incoherence in the following section. We first define two expansion operators: one that strengthens the agent's constraint and one that expands the AF. The *constraint expansion operator* takes as input a belief state and a formula ϕ representing a constraint that is to be incorporated into the new belief state. It is defined as follows.

Definition 8. *Let F be an AF, $S = (F, K)$ a belief state and $\phi \in \mathcal{L}_F$. The* constraint expansion *of S by ϕ, denoted $S \oplus \phi$ is defined by $S \oplus \phi = (F, K \wedge \phi)$.*

Example 4. Let $S_1 = (F, \top)$ where F is the AF shown in figure 1. We do not have $Bel(S_1) \models \texttt{in}_d$. That is, the agent does not believe that d is in. Consider the constraint expansion $S_2 = S_1 \oplus (\texttt{in}_a \vee \texttt{in}_b)$. Now we have $Bel(S_2) \models \texttt{in}_d$. That is, after learning that either a or b is in, the agent believes that d is in.

As to expanding the AF, we make two assumptions: First, we assume that arguments and attacks are not "forgotten". This means that elements can be added to an AF but not removed. Second, we assume that attacks between arguments are determined once the arguments are known. This means that no new attacks can be added between arguments already present in the agent's AF. A set of new arguments and attacks for an AF is called an *AF update*:

Definition 9. *Let $F = (A, \Rightarrow)$ be an AF. An* AF update *for F is a pair $F^* = (A^*, \Rightarrow^*)$ such that $A \cap A^* = \emptyset$ and $\Rightarrow^* \subseteq ((A \cup A^*) \times (A \cup A^*)) \setminus (A \times A)$.*

The *AF expansion operator* is defined as follows:

Definition 10. *Let $F = (A, \Rightarrow)$ be an AF, $S = (F, K)$ a belief state and $F^* = (A^*, \Rightarrow^*)$ an AF update for F. The* AF expansion *of S by F^*, denoted by $S \otimes F^*$ is defined by $S \otimes F^* = ((A \cup A^*, \Rightarrow \cup \Rightarrow^*), K)$.*

Example 5. Consider the belief state $S_1 = (F, \texttt{out}_a \vee \texttt{out}_b)$ where F is the AF shown in figure 1. Note that we do not have, e.g., $Bel(S_1) \models \texttt{in}_b$. Now consider the AF expansion $S_2 = (S_1 \otimes (\{e\}, \{(e, a)\})$. Now we do have $Bel(S_2) \models \texttt{in}_b$.

The two operators just defined allow us to study our belief state model in a dynamic setting, where an agent's belief state changes after new constraints on the AF's outcome are acquired or after adding new arguments and attacks.

4 Restoring coherence through AF expansion

In the previous section we presented a belief state model which includes, besides the agent's AF, a constraint on its outcome. We also explained that incoherence (i.e., the belief induced by the AF being inconsistent with the constraint) prevents the agent from forming beliefs that can be shown to be plausible via the agent's AF. The question is then: can the AF be expanded in such a way that the beliefs induced by it *are* consistent with the agent's constraints? In other words: can we restore coherence by expanding the AF in some way? Consider the following example.

Example 6. Let $S_1 = (F, \top)$ where F is the AF shown in figure 1. Suppose the agent learns that both a and b are out. The resulting state $S_2 = S_1 \oplus (\texttt{out}_a \wedge \texttt{out}_b)$ is incoherent, i.e., we have $Bel(S_2) \models \bot$. Now suppose the agent learns about arguments e and f, attacking a and b. The corresponding AF update is shown as

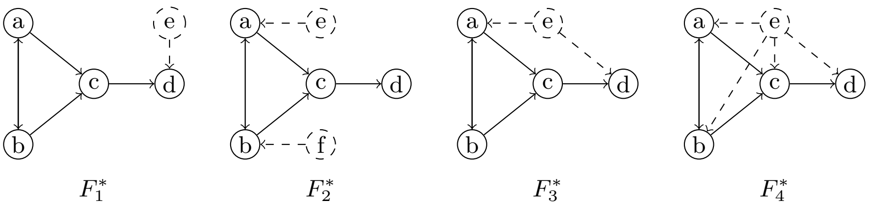

Fig. 2. Four argumentation framework updates

F_2^* in figure 2. The resulting state is $S_3 = S_2 \otimes (\{e, f\}, \{(e, a), (f, b)\})$. Coherence is now restored: $Bel(S_3) \not\models \bot$. In S_3 the agent believes, e.g., that c is in and d is out: $Bel(S_3) \models \texttt{in}_c \wedge \texttt{out}_d$. Notice that F_4^*, too, restores coherence in state S_2, whereas F_1^* and F_3^* do not.

This example shows that it is indeed possible to expand an AF such that coherence is restored. Note, also, that the AF updates F_2^* and F_4^* can be understood to provide the "missing explanation" for the agent's constraint ($\texttt{out}_a \wedge \texttt{out}_b$). That is, a and b are out *because* there are arguments attacking (among possibly other arguments) a and b. We can show that, as long as the agent's constraint is conflict-free, there always exists some AF expansion that restores coherence. That the agent's constraint is required to be conflict-free follows from the fact that attacks between existing arguments cannot be removed.

Theorem 1. *Let (F, K) be an incoherent belief state where K is conflict-free. There exists an AF update F^* for F such that $(F, K) \otimes F^*$ is coherent.*

Proof. Let (F, K) (with $F = (A, \Rightarrow)$) be an incoherent belief state with conflict-free K and let $L \in [K]$. Define $F^* = (A^*, \Rightarrow^*)$ by $A^* = \{a_I, a_U\}$ and $\Rightarrow^* = \{(a_U, a_{U)}\} \cup \{(a_I, x) \mid L(x) = O\} \cup \{(a_U, x) \mid L(x) = U\})$. It can be checked that F^* restores coherence.

This result essentially says that incoherence of a belief state can be understood to mean that the agent's AF is incomplete and needs to be expanded with additional arguments and attacks. A related result, called the *conservative strong enforcing* result, was presented by Baumann [2]. However, this result deals only with the possibility of making some set of arguments accepted. By contrast, we deal with arbitrary formulas expressible in the logical labeling language.

5 Fallback belief

Consider again the situation sketched in example 6. The agent learns that a and b are out, resulting in the belief state becoming incoherent. Consequently, the agent's belief in this state is inconsistent. It is, however, still possible to form reasonable, consistent *fallback* beliefs. To see what we mean, it is enough to just look at the framework in figure 1 and see that, once a and b are out, c becomes in and d becomes out.

However, there is no complete labeling in which a, b and d are out and c is in. So we need another mechanism to form fallback beliefs.

The starting point is to assume that the agent is equipped with a *rationality order* over the set of conflict-free labelings of AF, used to determine their "relative rationality". We assume this order to be a total pre-order (i.e., a complete, transitive and reflexive binary relation). When the agent is in an incoherent state (i.e., when all "fully rational" labelings are ruled out) she can fall back on the "most rational" labelings, permitting her to still form consistent fallback beliefs that represent the "most plausible" beliefs given the agent's constraint.

We can characterize the type of belief we end up with using an appropriate adaptation of the *KM postulates* introduced by Katsuno and Mendelzon [6]. In the following, we assume that the agent's rationality order is a function of F and denote it by $\preceq_F$. That is, the agent's AF (and *only* the agent's AF) guides her in comparing different, conflict-free labelings. Given a set $M \subseteq [Cf_F]$ we define $max_{\preceq_F}(M)$ by $max_{\preceq_F}(M) = \{L \in M \mid \forall L' \in M, L' \preceq_F L\}$. Letting $Bel^*(S)$ denote the fallback belief of an agent in the state S, we have the following result:

Proposition 2. *Let $S = (F, K)$ be a belief state. The following are equivalent:*

1. *Bel^* is defined by $Bel^*(S) = \phi$ where ϕ is a formula such that $[\phi] = max_{\preceq_F}([K] \cap [Cf_F])$ and $\preceq_F$ is a total pre-order over $[Cf_F]$.*
2. *Bel^* satisfies the following set of postulates:*
   P1: $Bel^*(S \oplus \phi) \models \phi$
   P2: If $Bel^*(S) \not\models \neg\phi$ then $Bel^*(S \oplus \phi) \equiv Bel^*(S) \wedge \phi$
   P3: If $K(S \oplus \phi)$ is conflict-free then $Bel^*(S \oplus \phi)$ is satisfiable
   P4: If $F_1 = F_2$, $K_1 \equiv K_2$ and $\phi_1 \equiv \phi_2$ then
   $Bel^*((F_1, K_1) \oplus \phi_1) \equiv Bel^*((F_2, K_2) \oplus \phi_2)$.
   P5: $Bel^*(S \oplus \phi) \wedge \psi \models Bel^*(S \oplus (\phi \wedge \psi))$.
   P6: If $Bel^*(S \oplus \phi) \not\models \neg\psi$ then $Bel^*(S \oplus (\phi \wedge \psi)) \models Bel^*(S \oplus \phi) \wedge \psi$.

This result means that fallback belief behaves like revised belief as studied in belief revision. A detailed discussion of the postulates can be found in the paper where they were introduced [6] which, in turn, built on the well-known AGM approach to belief revision [7]. We suffice by pointing out the differences with the original KM postulates. Namely, P3 requires that ϕ is not just satisfiable but that the agent's constraint is conflict-free after learning ϕ, and P4 requires both that the AFs (and thus orderings) in the two belief states are the same and that the constraint is equivalent.

The question we need to answer now is: how do we define the rationality order $\preceq_F$ or, in other words, how do we determine which conflict-free labelings of F are "more rational" than others? A natural way to do this is by looking at the arguments that are *illegally* labeled [8]. This is defined as follows:

Definition 11. *Let $F = (A, \Rightarrow)$ be an AF and $L \in \mathcal{M}_F$ a labeling of F. An argument $x \in A$ is said to be:*

– Illegally in *if and only if $L(x) = I$ and $\exists y \in A, (y, x) \in \Rightarrow$ and $L(y) \neq O$,*

- Illegally out *if and only if* $L(x) = O$ *and* $\nexists y \in A, (y, x) \in \Rightarrow$ *such that* $L(y) = I$,
- Illegally undecided *if and only if* $L(x) = U$ *and* $\exists y \in A, (y, x) \in \Rightarrow$ *and* $L(y) = I$ *or* $\nexists y \in A, (y, x) \in \Rightarrow$ *such that* $L(y) = U$.

We denote by $Z^I_F(L), Z^O_F(L)$ *and* $Z^U_F(L)$ *the sets of arguments that are, respectively, illegally in, out and undecided in* L.

Intuitively, an illegally labeled argument indicates a local violation of the condition imposed on the argument's label according to the complete semantics. It can be checked, for example, that a labeling L is a complete labeling if and only if it has no arguments illegally labeled. It can also be checked that, in a conflict-free labeling, arguments are never illegally in. Thus in judging the relative rationality of a conflict-free labeling L, we only have to look at the sets $Z^O_F(L)$ and $Z^U_F(L)$.

What, exactly, do the sets $Z^O_F(L)$ and $Z^U_F(L)$ tell us about the relative rationality of a conflict-free labeling L? To answer this we have to look at what it takes to turn L into a complete labeling. We say that an AF update that turns L into (part of) a complete labeling of the (expanded) AF is an AF update that *completes* L. Formally:

Definition 12. *Let* $F^* = (A^*, \Rightarrow^*)$ *be an AF update for* $F = (A, \Rightarrow)$ *and* L *a conflict-free labeling of* F. *We say that* F^* completes L *if and only if there is a complete labeling* L' *of the AF* $(A \cup A^*, \Rightarrow \cup \Rightarrow^*)$ *such that* $(L' \downarrow A) = L$, *where* $(L \downarrow A)$ *is a function defined by* $(L \downarrow A)(x) = L(x)$, *for all* $x \in A$.

To determine the "impact" of an AF update we can look at what we call its *attack degree*, which is the number of arguments in the existing AF that are attacked by the AF update.

Definition 13. *Let* $F^* = (A^*, \Rightarrow^*)$ *be an AF update for* $F = (A, \Rightarrow)$. *We denote by* $\delta_F(F^*)$ *the* attack degree *of* F^*, *defined by* $\delta_F(F^*) = |\{x \in A \mid \exists y \in A^*, (y, x) \in \Rightarrow^*\}|$.

We now show how, given a conflict-free labeling L, the sets $Z^O_F(L)$ and $Z^U_F(L)$ inform us about the minimal impact it would take to complete L, or to turn L into a fully rational point of view. We use this as the criterion to define the rationality order $\preceq_F$, making the assumption that the agent believes that conflict-free labelings that require minimal impact to be turned into a complete labeling are most rational. The following result links up the cardinality of the sets $Z^O_F(L)$ and $Z^U_F(L)$ associated with a labeling L and the minimal impact required to complete L:

Proposition 3. *Let* L *be a conflict-free labeling of an AF* F. *If* F^* *completes* L *then* $\delta_F(F^*) \geq |Z^O_F(L) \cup Z^U_F(L)|$.

Proof. Let L be a conflict-free labeling of an AF $F = (A, \Rightarrow)$ and let $F^* = (A^*, \Rightarrow^*)$ be an AF update that completes L. Let L' be a labeling of $(A \cup F^*, \Rightarrow \cup \Rightarrow^*)$ such that $L = (L' \downarrow A)$. We have that L' is a complete labeling only if, for all $x \in Z^O_F(L)$, x is legally out in L' and for all $x \in Z^U_F(L)$, x is legally undecided in L'. It can be checked that for all $x \in Z^O_F(L) \cup Z^U_F(L)$, x must therefore be attacked by some argument in A^*. Hence we have that $\delta_F(F^*) \geq |Z^O_F(L) \cup Z^U_F(L)|$.

We thus have that the cardinality of the sets $Z^O_F(L)$ and $Z^U_F(L)$ combined is a lower bound on the attack degrees of the set of AF updates that complete L. We now define the rationality order $\preceq_F$ as follows. Let $L, L' \in [Cf_F]$,

$$L \preceq_F L' \text{ iff } |Z^O_F(L) \cup Z^U_F(L)| \geq |Z^O_F(L') \cup Z^U_F(L')|$$

By using this rationality order, the outcome of the AF according to the agent's fallback belief is the outcome that would hold if some minimal impact, coherence restoring AF update would be performed.

Example 7. The table below represents $\preceq_F$ for the AF F shown in figure 1.

0	1	2	3	4
$\mathtt{OIOI}$	$\mathtt{OIO\underline{O}}$ $\mathtt{UUU\underline{O}}$	$\mathtt{\underline{OO}IO}$ $\mathtt{UU\underline{OO}}$	$\mathtt{\underline{OOO}I}$ $\mathtt{\underline{OU}U\underline{O}}$	$\mathtt{\underline{OOOO}}$ $\mathtt{\underline{OUOU}}$ $\mathtt{\underline{OUOO}}$
$\mathtt{UUUU}$	$\mathtt{OIO\underline{U}}$ $\mathtt{IOO\underline{O}}$	$\mathtt{\underline{OU}UU}$ $\mathtt{UU\underline{OU}}$	$\mathtt{\underline{OOU}U}$ $\mathtt{\underline{UOO}I}$	$\mathtt{\underline{OOOU}}$ $\mathtt{\underline{UOOO}}$
$\mathtt{IOOI}$	$\mathtt{UU\underline{O}I}$ $\mathtt{IOO\underline{U}}$	$\mathtt{\underline{UO}UU}$	$\mathtt{\underline{OUO}I}$ $\mathtt{\underline{UO}U\underline{O}}$	$\mathtt{\underline{OOUO}}$ $\mathtt{\underline{UOOU}}$

The header of each column shows the number of arguments illegally labeled. This determines the ordering $\preceq_F$ in a straightforward way, i.e., $L \prec_F L'$ iff L is another column to the right of L'. Note that we underline arguments that are illegally labeled. The following fallback beliefs can be read off this table:

- $Bel^*(F, \mathtt{out}_a \wedge \mathtt{out}_b) \models \mathtt{in}_c$ (if a and b are out then c is in).
- $Bel^*(F, \mathtt{in}_c) \models \mathtt{out}_a \wedge \mathtt{out}_b$ (if c is in then a and b must be out).
- $Bel^*(F, \mathtt{out}_d) \models \neg(\mathtt{in}_a \wedge \mathtt{in}_b)$ (even if d is out, a and b cannot both be in).
- $Bel^*(F, \mathtt{out}_d) \models \mathtt{u}_a \rightarrow \mathtt{u}_c$ (even if d is out, if a undecided then so is c).

Note that none of these inferences can be made just by looking at the complete labelings of F alone.

As expected, regular and fallback belief coincides in coherent states, i.e.:

Proposition 4. *Let S be a belief state. We have that $Bel(S) \models Bel^*(S) \models Cf_F$ and if S is coherent then $Bel^*(S) \models Bel(S)$.*

As the following theorem states more formally, and as we pointed out before, fallback belief is formed by assuming the most rational outcome of an AF in an incoherent state to be the outcome that would hold after a coherence restoring AF update with minimal impact. That is, if coherence is restored using an AF update with a minimal attack degree, then the agent's regular belief in the updated state includes the agent's fallback belief in the old state.

Theorem 2. *Let S be an incoherent belief state and F^*_1 a minimal coherence restoring update (i.e., $S \otimes F^*_1$ is coherent and there is no F^*_2 such that $S \otimes F^*_2$ is coherent and $\delta_F(F^*_2) < \delta_F(F^*_1)$). It holds that $Bel(S \otimes F^*_1) \models Bel^*(S)$.*

Proof. Assume $S = (F, K)$ (with $F = (A, \Rightarrow)$) is incoherent and F^* a minimal coherence restoring update. We show that $L \in [K(S \otimes F_1^*)]$ implies $(L \downarrow A) \in max_{\preceq_F}([K] \cap [Cf_F])$. Let $L \in [K(S \otimes F_1^*)]$ and suppose the contrary: $(L \downarrow A) \notin max_{\preceq_F}([K] \cap [Cf_F])$. We have that $(L \downarrow A) \in [K] \cap [Cf_F]$ thus there is some $L' \in [K] \cap [Cf_F]$ such that $(L \downarrow A) \prec L'$. Then there is also an F_2^* that completes L' such that $\delta_F(F_2^*) < \delta_F(F^*)$ (proposition 3). But then F^* is not minimal because F_2^* also restores coherence. Contradiction! Thus we have that $L \in [K(S \otimes F_1^*)]$ implies $(L \downarrow A) \in max_{\preceq_F}([K] \cap [Cf_F])$. From this it follows that $Bel(S \otimes F_1^*) \models Bel^*(S)$.

Example 8. Let $S = (F, \mathtt{out}_c)$ be a belief state with $F = (\{a, b, c, d, e\}, \{(a, b), (b, c), (d, e), (e, c)\})$. We have $Bel^*(S) \equiv \phi$ where ϕ is a formula s.t. $[\phi] = \{\mathtt{IO\underline{O}IO}, \mathtt{\underline{O}IOIO}, \mathtt{IOO\underline{O}I}\}$. Three minimal coherence restoring AF updates are: $F_1^* = (\{f\}, \{(f, c)\})$, $F_2^* = (\{f\}, \{(f, a)\})$ and $F_3^* = (\{f\}, \{(f, d)\})$. We have that $Bel(S \otimes F_n^*) \equiv \psi$, where ψ is a formula s.t. $[\psi] = \{\mathtt{IOOIOI}\}$ if $n = 1$; $[\psi] = \{\mathtt{OIOIO}\}$, if $n = 2$ and $[\psi] = \{\mathtt{IOOOII}\}$, if $n = 3$. It can be checked that, for all $n \in \{1, 2, 3\}$, $Bel(S \otimes F_n^*) \models \phi$ and thus $Bel(S \otimes F_n^*) \models Bel^*(S)$.

6 The decision problem for fallback belief in ASP

Answer-set programming has proven to be a useful mechanism to compute extensions of AFs under various semantics [9, 10]. The idea is to encode both the AF and a so called *encoding* of the semantics in a single program of which the stable models correspond to the extensions of the AF.

In this section we show that the problem of deciding whether a formula ϕ is a fallback belief in a state (F, K) can be solved, too, using an answer-set program. The encoding, shown in listing 6, turns out to be surprisingly simple, and works as follows. The AF is assumed to be encoded (line 1) using the predicates `arg/1` and `att/2`. For example, the AF of figure 1 is encoded by the facts `arg(a)`, `arg(b)`, `arg(c)`, `arg(d)`, `att(a,b)`, `att(a,c)`, `att(b,a)`, `att(b,c)` and `att(c,d)`. The choice rule on line 2 ensures that each argument $x \in A$ gets one of thee labels, expressed by the predicates `in`, `out` and `undec`. On lines 3 and 4 conflict-freeness is ensured. Given just these constraints, stable models correspond to conflict-free labelings of F. Lines 5-10 are used to establish whether an argument $x \in A$ is illegally labeled, expressed by the predicate `illegal(x)`. The cardinality of this predicate is minimized on line 12. Finally, the agent's constraint is assumed to be encoded (line 11) using statements restricting the possible labels assigned to arguments. For example, the constraint $\mathtt{out}_a \vee \mathtt{out}_b$ can be encoded by adding the choice rule `1 {out(a), out(b)} 2`, and the formula $\mathtt{out}_a \wedge \mathtt{out}_b$ by the two facts `out(a)` and `out(b)`. The (optimal) stable models now correspond to maximally rational conflict-free labelings that satisfy the constraint. (Note that, with no constraint the program would yield complete labelings of F.)

The program is compatible with the *Gringo* grounder (version 3.0.5) in combination with the *Clasp* answer set solver (version 2.1.2) [11]. The optimal stable

```
% <-- Framework encoding here -->
1 { in(X), out(X), undec(X) } 1 :- arg(X).
out(Y) :- att(X, Y), in(X).
out(X) :- att(X, Y), in(Y).
legally_out(X) :- out(X), att(Y, X), in(Y).
legally_undec(X) :- undec(X), att(Y, X), undec(Y).
illegally_out(X) :- out(X), not lllegally_out(X).
illegally_undec(X) :- undec(X), not legally_undec(X).
illegal(X) :- illegally_out(X).
illegal(X) :- illegally_undec(X).
% <-- Knowledge encoding here -->
#minimize { illegal(X) }.
```

Program 1: The ASP encoding for the decision problem for fallback belief

models can be obtained by running the program (assuming it is stored in a file called `program`) with the command `gringo program | clasp --opt-all`. The final step of the complete procedure amounts to checking whether the formula ϕ is true in every optimal stable model. Alternatively, the set of stable models of the program can be converted into a formula in disjunctive normal form that represents the agent's whole fallback belief.

7 Additional semantics

We have focused in this paper on the complete semantics. Some of the notions we introduced can be adapted to other semantics in a straightforward way. For example, we can define a family of types of *s-belief* for a semantics $s \in \{Co, St, Pr, Gr\}$ (for Complete, Stable, Preferred, Grounded) as follows:

Definition 14. *Let $F = (A, \Rightarrow)$ be an AF, $S = (F, K)$ be the agent's belief state and $s \in \{Co, St, Pr, Gr\}$. We define $Bel_s(S)$ by $Bel_s(S) = \phi$, where ϕ is a formula such that $[\phi] = \{L \in [K] \mid L \text{ is an } s\text{-labeling of } F\}$. We say that that the agent s-*believes *ϕ if and only if $Bel_s(S) \models \phi$.*

It can be checked that we have $Bel_{Gr}(S) \models Bel(S)_{Co}$ and $Bel_{St}(S) \models Bel_{Pr}(S) \models Bel(S)_{Co}$. This follows directly from the fact that grounded labelings are also complete, stable also preferred, and so on. Now consider e.g. the following notion of 's-coherence':

Definition 15. *Let S be a belief state and $s \in \{Co, St, Pr, Gr\}$. We say that S is s-coherent if and only if $Bel_s(S) \not\models \bot$.*

Given these definitions of s-belief and s-coherence we can state a more general version of theorem 1:

Theorem 3. *Let $s \in \{Co, St, Pr, Gr\}$ and let (F, K) be an s-incoherent belief state where K is conflict-free. There exists an AF update F^* for F such that $(F, K) \otimes F^*$ is s-coherent.*

Proof. Let (F, K) (with $F = (A, \Rightarrow)$) be an s-incoherent belief state with conflict-free K and let $L \in [K]$. Define $F^* = (A^*, \Rightarrow^*)$ by $A^* = \{a_I, a_U\}$ and $\Rightarrow^* = \{(a_U, a_{U)}\} \cup \{(a_I, x) \mid L(x) = O\} \cup \{(a_U, x) \mid L(x) = U\})$. It can be checked that F^* restores s-coherence.

Fallback belief, however, is less straightforward to adapt, as the corresponding rationality orderings would have to combine different criteria, i.e. minimizing/-maximizing in-labeled arguments w.r.t. set-inclusion and minimizing illegally labeled arguments. This could, depending on what properties one expect grounded and preferred fallback belief to satisfy, mean that the corresponding rationality orders, too, become partial pre-orders.

8 Related work

In this section we give a short overview of related work. We already mentioned the relation of our work with the *enforcing problem* [2]. The authors present a result stating that every conflict-free extension can be enforced (i.e., made accepted under a semantics) with an appropriate AF expansion. In our setting we consider more general types of enforcing, not limited only to acceptance of sets of arguments. Our theorem 3 thus strengthens their possibility result.

Next, different ways to characterize the impact of AF expansions have been studied. This includes minimality w.r.t. the number of added attacks, studied in the context of the enforcing problem [12]. Further criteria have been defined in the study of the impact on the outcome of an AF of adding an argument [3]. A limitation in that work is that it considers only additions of a single argument.

The rationality ordering used in section 5 is related to a preferential model semantics for argumentation [13] and a study of nonmonotonic inference relations to reason with AFs [4]. Also related are *open labelings* [14], which have a purpose similar to ours, i.e., to identify arguments to attack in order to make a labeling consistent with an AF.We should also mention other work in which (parts of) argumentation theory are formalized using logics. They include models using modal logics [15, 16]; translations of the problem of computing extensions to problems in classical logic or ASP [17, 18]; and a study of a logical language consisting of attack and defense connectives [19].

Finally, our model is related to the concept of a *constrained AF*, which combines an AF with a formulas expressing a constraint on the status of the arguments [20]. A limitation is that the constraint is assumed to be admissible, meaning that constraints that are inconsistent with the outcome of the AF cannot, in general, be dealt with. Furthermore, this work does not explore the relation between constraints and AF expansions.

9 Conclusion and future work

We believe that theories about dynamics in abstract argumentation should address two issues: First, agents may observe or otherwise learn that an AF must have a certain outcome and second, agents may learn about new arguments/attacks. We presented such a theory, using a logical labeling language, and focusing on the problem that observations may make an agent's belief state incoherent. We discussed two days to deal with this: AF expansions and fallback belief.

We plan to extend our work in a number of directions. First, our model allows iterated updates only under the assumption that new observations never contradict old ones. In order to allow this we have to look at revising the agent's constraint in the light of conflicting observations. Second, a number of generalizations are possible. For example, we may drop requirement that observations are conflict-free and we can allow removal of arguments and attacks.

Finally, we plan to investigate connections between the areas of abstract argumentation and belief revision beyond those presented in this paper. We believe that the approach of using a logical labeling language to reason about the outcome of an AF is an essential step towards establishing such connections.

References

1. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2) (1995) 321–358
2. Baumann, R., Brewka, G.: Expanding argumentation frameworks: Enforcing and monotonicity results. In Baroni, P., Cerutti, F., Giacomin, M., Simari, G.R., eds.: COMMA. Volume 216 of Frontiers in Artificial Intelligence and Applications., IOS Press (2010) 75–86
3. Cayrol, C., de Saint-Cyr, F., Lagasquie-Schiex, M.: Change in abstract argumentation frameworks: Adding an argument. Journal of Artificial Intelligence Research **38**(1) (2010) 49–84
4. Booth, R., Kaci, S., Rienstra, T., van der Torre, L.: Monotonic and non-monotonic inference for abstract argumentation. In: FLAIRS. (2013)
5. Caminada, M.: On the issue of reinstatement in argumentation. In: JELIA. (2006) 111–123
6. Katsuno, H., Mendelzon, A.O.: Propositional knowledge base revision and minimal change. Artificial Intelligence **52**(3) (1991) 263–294
7. Alchourrón, C.E., Gärdenfors, P., Makinson, D.: On the logic of theory change: Partial meet contraction and revision functions. Journal of symbolic logic (1985) 510–530
8. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowledge Eng. Review **26**(4) (2011) 365–410
9. Toni, F., Sergot, M.: Argumentation and answer set programming. In: Logic programming, knowledge representation, and nonmonotonic reasoning. Springer (2011) 164–180
10. Egly, U., Gaggl, S.A., Woltran, S.: Aspartix: Implementing argumentation frameworks using answer-set programming. In: Logic Programming. Springer (2008) 734–738
11. Gebser, M., Kaufmann, B., Kaminski, R., Ostrowski, M., Schaub, T., Schneider, M.: Potassco: The potsdam answer set solving collection. AI Communications **24**(2) (2011) 107–124

12. Baumann, R.: What does it take to enforce an argument? Minimal change in abstract argumentation. In Raedt, L.D., Bessière, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F., Lucas, P.J.F., eds.: ECAI. Volume 242 of Frontiers in Artificial Intelligence and Applications., IOS Press (2012) 127–132
13. Roos, N.: Preferential model and argumentation semantics. In: Proceedings of the 13th International Workshop on Non-Monotonic Reasoning (NMR-2010). (2010)
14. Gratie, C., Florea, A.M.: Argumentation semantics for agents. In Cossentino, M., Kaisers, M., Tuyls, K., Weiss, G., eds.: EUMAS. Volume 7541 of Lecture Notes in Computer Science., Springer (2011) 129–144
15. Grossi, D.: On the logic of argumentation theory. In van der Hoek, W., Kaminka, G.A., Lespérance, Y., Luck, M., Sen, S., eds.: AAMAS, IFAAMAS (2010) 409–416
16. Schwarzentruber, F., Vesic, S., Rienstra, T.: Building an epistemic logic for argumentation. In del Cerro, L.F., Herzig, A., Mengin, J., eds.: JELIA. Volume 7519 of Lecture Notes in Computer Science., Springer (2012) 359–371
17. Besnard, P., Doutre, S.: Checking the acceptability of a set of arguments. In Delgrande, J.P., Schaub, T., eds.: NMR. (2004) 59–64
18. Egly, U., Gaggl, S.A., Woltran, S.: Answer-set programming encodings for argumentation frameworks. Argument and Computation **1**(2) (2010) 147–177
19. Boella, G., Hulstijn, J., van der Torre, L.W.N.: A logic of abstract argumentation. In Parsons, S., Maudet, N., Moraitis, P., Rahwan, I., eds.: ArgMAS. Volume 4049 of Lecture Notes in Computer Science., Springer (2005) 29–41
20. Coste-Marquis, S., Devred, C., Marquis, P.: Constrained argumentation frameworks. In Doherty, P., Mylopoulos, J., Welty, C.A., eds.: KR, AAAI Press (2006) 112–122