Dialogue games for enforcement of argument acceptance and rejection via attack removal

Jérémie Dauphin[1]* and Ken Satoh[2]

[1] CSC, University of Luxembourg, Esch-sur-Alzette - Luxembourg
[2] National Institute of Informatics, Tokyo - Japan

Abstract. Argumentation is dynamic in nature and most commonly exists in dialogical form between different agents trying to convince each other. While abstract argumentation framework are mostly static, many studies have focused on dynamical aspects and changes to these static frameworks. An important problem is the one of *argument enforcement*, modifying an argumentation framework in order to ensure that a certain argument is accepted. In this paper, we use *dialogue games* to provide an exhaustive list of minimal sets of attacks such that when removed, a given argument is credulously accepted with respect to preferred semantics. We then extend the method to enforce other acceptability statuses and cope with sets of arguments.

Keywords: abstract argumentation · dialogue games · argument enforcement · attack removal · dynamical argumentation

1 Introduction

Argumentation is a dynamic process where one party attempts to convince another, or itself, of the validity of some statement. This process involves both parties putting arguments in favor or against the validity of said statement in turns, until one party has been convinced by running out of counter-arguments.

Abstract argumentation [6] provides a static approach where all arguments and their interactions are given from the beginning in what is called an argumentation framework, and the acceptability of the arguments is determined based on the entire state of the framework. While the dialogical nature might appear to be missing at first, dialogue games can be extracted from such a framework where one focuses on the acceptability status of a single argument with a process very similar to the one described in the earlier paragraph.

When focusing on adding a dynamic aspect to abstract argumentation, one common problem is the problem of argument enforcement: given a particular argument in a framework, what changes are required in order to change the status of the argument from rejected to accepted or vice-versa? This problem is tightly linked to persuasion [8], as the goal of an agent there is for the other agent

* The work of Jérémie Dauphin was supported by the H2020 Marie Skłodowska-Curie grant number 690974 for the project MIREL.

to accept or perhaps reject a given argument of interest, and hence the question of what the best approach is for this goal. There is also an element of strategy, as the other party might have an objective of their own, perhaps wanting to conserve their beliefs about certain arguments more than about others.

There has been much work already on the problem of argument enforcement. In his paper, Baumann [2] studies the minimal necessary additions, in terms of arguments and attacks, one needs to perform in order to enforce the acceptability of a given argument. Coste-Marquis et al. [5] examine the enforcement issue from the point of view of belief revision by allowing only for change in the attacks.

On another side, there also exists studies on the effects of argumentation framework manipulation. In their paper, Cayrol et al. [4] study the possible repercussions that the addition of an argument might have on the framework. Another study by Boella et al. [3] focuses on the effects that removing arguments and attacks in an argumentation framework have on the grounded extension. Liao et al. [9] propose a partitioning method to efficiently compute the repercussions of changes in an abstract argumentation framework.

In this paper, we provide an algorithm to compute an exhaustive list of sets of attacks, which, when removed, toggle the acceptance status of an argument, as well as a formal framework to ensure the well behavior of the algorithm. In practice, the operation of removing an attack could for example be performed by arguing for the preference of the attacked argument over its attacker, or by reformulating the argument so that it still reaches the same conclusion but from different, less vulnerable premises. Since operations of this kind come with a cost, we wish to minimize these costs, but since these costs might differ between the different attacks, we also wish to provide an exhaustive list of solutions.

In Section 2, we provide preliminary definitions of abstract argumentation upon which we build our results. In Section 3, we delve into the problem of enforcing the credulous acceptance of an argument with respect to the preferred semantics. We finish in Section 4 with a conclusion and discussion of potential future work.

2 Preliminaries

In this section we provide definitions of existing notions of abstract argumentation and dialogue games which will be used later on.

An argumentation framework [6] (AF) is a pair $\langle A, R\rangle$ where A is a finite set of atomic entities called *arguments*, and $R \subseteq A \times A$ is a relation of *attack*.

Definition 1. *Let $F = \langle A, R\rangle$ be an AF and $S \subseteq A$ a set of arguments. We say that S is:* conflict-free *iff there exist no $a, b \in S$ such that a attacks b;* admissible *iff it is conflict-free and for all $a \in S$, $b \in A$ such that b attacks a, there exists $c \in S$ such that c attacks b; and a* preferred extension *iff it is a $\subseteq$-maximal admissible set.*

The preferred semantics is the function which returns all preferred extensions of a given AF. Since it may return more than one extension, we also define different degrees of acceptance.

Definition 2. *Let $F = \langle A, R \rangle$ be an AF, $a \in A$ an argument. We say that a is:* credulously accepted *iff for some preferred extensions E, $a \in E$;* rejected *iff for all preferred extensions E, $a \notin E$.*

Dialogue games [7, 10, 12] provide a proof theory to test the acceptability of a given argument in a fixed AF. The games involve two players, the *proponent* and the *opponent*, where moves are of the form a_p with a an argument and $p \in \{\mathsf{pro}, \mathsf{opp}\}$ a player. The game starts with a_{pro}, where a is the argument to be tested. The opponent then moves forward an argument to attack it, attempting to undermine its acceptability. A *legal move function* is a function from sequences of moves to sets of moves which, given a sequence of moves, dictates which moves could possibly be put forward next by the opposing player. We also define a *dispute* on an argument a as a finite sequence of moves $a_{\mathsf{pro}} \leftarrow b_{\mathsf{opp}} \leftarrow ...$ starting with a_{pro}, such that every move is a legal move with respect to the earlier part of the dispute according to a set legal move function. If a dispute has no more legal moves, we say that the dispute is *final* and that the player who put forward the last move is the *winner* of the dispute.

Definition 3. *Let $F = \langle A, R \rangle$ be an argumentation framework and d a dispute in F with last move a_p. The legal move function f_{pref} for the preferred semantics is defined as follows:*

1. *if $p = \mathsf{pro}$, then $f_{pref}(d) = \{b_{\mathsf{opp}} \mid (b, a) \in R, \nexists b_{\mathsf{opp}} \in d\}$;*
2. *otherwise, $f_{pref}(d) = \{b_{\mathsf{pro}} \mid (b, a) \in R\}$.*

The acceptability of an argument is then determined by whether or not the proponent has a *strategy* to win the dialogue game. This is a conditional plan which is formally represented by a set of disputes including a dispute for each possible move of the opponent. The definition makes use of the notion of *sub-dispute*, which is a sub-sequence of a dispute with the same initial argument.

Definition 4. *Let $F = \langle A, R \rangle$ be an argumentation framework, $a \in A$ an argument and f a legal move function. A* defending strategy *for a in F with respect to f is a non-empty set of disputes T such that:*

1. *each dispute in T has initial argument a and is won by pro;*
2. *for each $d \in T$ and for each sub-dispute d' of d, if the last move in d' is an argument b moved by pro, then for any $c \in f(d')$, there exists a $d'' \in T$ such that $d \leftarrow c$ is a sub-dispute of d'';*
3. *there is no $d, d' \in T, b \in A$ such that $b_{\mathsf{pro}} \in d$ and $b_{\mathsf{opp}} \in d'$.*

The last item represents the requirement that the strategy must be conflict-free, since the goal is to construct an admissible set. Note that these are usually called winning strategies in the literature, however since we will later define strategies for the victory of the opponent, we prefer using the term *defending.*

The existence of a defending strategy for a given argument is equivalent to its credulous acceptability [10].

Theorem 1. *Let $F = \langle A, R \rangle$ be an argumentation framework and $a \in A$ an argument. a is credulously accepted in F with respect to the preferred semantics iff there is a defending strategy for a in F with respect to f_{pref}.*

3 Enforcing credulous acceptance for preferred semantics

In this section, we first provide a few new definitions and associated results for dialogue games, and then use these to provide a procedure for identifying minimal sets of attacks to be removed in order to enforce credulous acceptability of a given argument with respect to the preferred semantics. We start by defining a counterpart to the defending strategies, i.e. a notion of strategy for the opponent to win the dispute.

Definition 5. *Let $F = \langle A, R \rangle$ be an argumentation framework, $a \in A$ an argument and f a legal move function. An* opposing strategy *for a in F with respect to f is a non-empty set of disputes T such that:*

1. *each dispute in T has initial argument a and either is won by opp, or contains a move b_{pro} such that there exists $d' \in T$ with $b_{\mathsf{opp}} \in d'$;*
2. *for each $d \in T$ and for each sub-dispute d' of d, if the last move in d' is an argument b moved by opp, then for any c that pro can legally move against b, there exists a $d'' \in T$ such that $d \leftarrow c_{\mathsf{opp}}$ is a sub-dispute of d''.*

The opponent's goal is to prevent the proponent from successfully defending the argument in focus and thus construct an admissible set containing it. The opponent does this by providing a set of argument attacks from which the proponent cannot fully defend, and thus shows no admissible set containing the argument in focus can be constructed.

Theorem 2. *Let $F = \langle A, R \rangle$ be an argumentation framework and $a \in A$ an argument. There exists a defending strategy for a in F with respect to f_{pref} iff there does not exist an opposing strategy for a in F with respect to f_{pref}.*

Proof sketch: This follows from Zermelo's Theorem [13]. □

Corollary 1. *Let $F = \langle A, R \rangle$ be an argumentation framework and $a \in A$ an argument. a is rejected in F with respect to the preferred semantics iff there is an opposing strategy for a in F with respect to f_{pref}.*

We have now laid down the foundations for dialogue games, in which we can identify whether an argument is accepted or rejected by providing defending strategies, respectively opposing strategies for said argument. In order to alter that argument's acceptance status, we can use this information in order to pinpoint minimal sets of attacks which, when removed, guarantee that the argument's status is changed. For this, we first have to be able to retrieve the attacks which correspond to each player's moves in a particular strategy.

Definition 6. *Let $F = \langle A, R \rangle$ be an AF and d a dispute. We define the* attacking set *of d to be $\mathcal{A}(d) = \{(a, b) \mid$ for some dispute d', $d' \leftarrow b_{\mathsf{pro}} \leftarrow a_{\mathsf{opp}}$ is a sub-dispute of $d\}$. We define the* defending set *of d to be $\mathcal{D}(d) = \{(a, b) \mid$ for some dispute $d', d' \leftarrow b_{\mathsf{opp}} \leftarrow a_{\mathsf{pro}}$ is a sub-dispute of $d\}$.*

For a set of disputes D, we write $\mathcal{A}(D)$ for $\bigcup_{d \in D} \mathcal{A}(d)$, and similarly, $\mathcal{D}(D)$ for $\bigcup_{d \in D} \mathcal{D}(d)$.

We wish to disrupt all winning strategies of a given player, which will hence give his counter-part a strategy to win. We do this by removing one of the player's possible moves from each winning strategy in the form of attack removal. To identify these attacks, since we want to work with minimal changes to the framework, we use *minimal hitting sets*.

Definition 7. *Let E be a set and S a set of subsets of E. We define the set of* hitting sets *of S to be $\mathcal{HS}(S) = \{s \subseteq E \mid \forall s' \in S, s \cap s' \neq \emptyset\}$. We also define the set of* minimal hitting sets *of S to be $\mathcal{MHS}(S) = \{s \in \mathcal{HS}(S) \mid \nexists s' \in \mathcal{HS}(S)$ such that $s' \subset s\}$.*

The enumeration of all minimal hitting sets can be efficiently done using for example the algorithm described by Satoh et al. [11].

Example 1. Let $S = \big\{\{(a,b),(c,d)\},\{(c,d),(e,f)\}\big\}$. Then, the minimal hitting sets of S are $\mathcal{MHS}(S) = \big\{\{(a,b),(e,f)\},\{(c,d)\}\big\}$.

If an argument is rejected, we can enumerate all the opposing strategies for it. We then identify candidate sets of attacks to be removed by computing the minimal hitting sets of the attacking sets in the opposing strategies.

Definition 8. *Let $F = \langle A, R\rangle$ and S a set of strategies in F. We define the* set of critical attack sets *of S to be $\mathcal{CA}(S) = \mathcal{MHS}(\{\mathcal{A}(T) \mid T \in S\})$. We also define the* set of critical defense sets *of S to be $\mathcal{CD}(S) = \mathcal{MHS}(\{\mathcal{D}(T) \mid T \in S\})$.*

Our first result with regard to argument status enforcement is that entirely removing at least one of these sets of attacks is required in order to enforce the acceptance of the argument of interest.

Lemma 1. *Let $F = \langle A, R\rangle$ be an AF where some argument $a \in A$ is rejected with respect to preferred semantics and S the set of opposing strategies for a with respect to f_{pref}. For all $R' \subseteq R$, if there is no $s \in \mathcal{CA}(S)$ such that $s \subseteq (R \setminus R')$, then a is rejected in $F' = \langle A, R'\rangle$.*

Proof sketch: There exists at least one opposing strategy which is still viable in F', otherwise there would be a critical attack set which has been fully removed. Hence, a is still rejected in F'. □

This lemma shows that fully removing some of the critical attacks of S is a necessary condition to enforce the acceptability of the argument of interest a. Now the question is whether this is a sufficient condition to ensure it. An issue could arise when removing a *controversial* attack with respect to a, i.e. an attack which appears in $\mathcal{A}(d)$ and $\mathcal{D}(d')$ for two disputes d and d' with initial argument a. Deleting such an attack might also hinder some of the proponent's potential new defending strategies and give rise to new opposing strategies. Hence, when removing such an attack, one might need to iterate the argument enforcement procedure.

Algorithm 1 Enumeration of solutions for enforcement of credulous acceptance with respect to preferred semantics

Input: $\langle A, R\rangle$ is an AF with $a \in A$, $\mathit{Except} \subseteq R$ is a set of attacks.
Output: *sols* is the exhaustive set of minimal solutions to the enforcement problem.

```
 1: procedure EnumAtksAcc($\langle A, R\rangle, a, \mathit{Except}$)        ▷ Enumerate all solutions
 2:     $S \leftarrow$ OppStrats($F, a$)
 3:     if $S = \emptyset$ then return $\{\emptyset\}$        ▷ If already accepted, solution is no change
 4:     end if
 5:     $\mathit{sols} := \emptyset$
 6:     $C \leftarrow \mathrm{CA}(S)$
 7:     for every $s \in C$ such that $\nexists e \in \mathit{Except}$ with $e \subseteq s$ do
 8:         if $\exists r \in s$ such that $r$ is controversial then
 9:             $E := \mathit{Except} \cup (C \setminus \{s\})$
10:             $\mathit{sols} := \mathit{sols} \cup \{s \cup s' \mid s' \in$ EnumAtksAcc($\langle A, R \setminus s\rangle, a, E$)$\}$
11:         else
12:             $\mathit{sols} := \mathit{sols} \cup \{s\}$
13:         end if
14:     end for
15:     return $\mathit{sols}$
16: end procedure
```

Algorithm 1 provides the details of how to compute the exhaustive list of minimal sets of attack to be removed in order to enforce the acceptance of an argument a in a framework F. The algorithm relies on a function OppStrats which computes and returns all opposing strategies for a given argument in a given framework, and a function CA which computes the set of critical attack sets of a given set of strategies. Note that the procedure EnumAtksAcc should initially be called with an empty set of exceptions, however if some sets of attacks in the framework are already determined to be jointly crucial and impossible to remove, one can add them to the initial set of exceptions. This set of exceptions is mainly used in order to prevent the iterated procedure called in line 10 to consider removing attacks which could already have been selected at an earlier stage and hence ensure the minimality of the output. If the set of exceptions is initially too restrictive, it is possible that no solution is returned.

Example 2. Consider the argumentation framework $F_2 = \langle A_2, R_2\rangle$ and final disputes for the argument a depicted in Figure 1. In this framework, there are two opposing strategies: $T_1 = \{d_1\}$ and $T_2 = \{d_4\}$. We get that $\mathcal{CA}(T_1 \cup T_2) = \{S_1, S_2, S_3, S_4\}$, where $S_1 = \{(b, a), (e, a)\}$, $S_2 = \{(j, h), (k, i)\}$, $S_3 = \{(e, a), (j, h)\}$ and $S_4 = \{(b, a), (k, i)\}$. Since S_1 contains no controversial attacks, it is directly listed as a solution. On the other hand, both (j, h) and (k, i) are controversial since (j, h) also appears in $\mathcal{D}(d_2)$ and (k, i) also appears in $\mathcal{D}(d_5)$. However, it turns out that a is accepted in $\langle A_2, R_2 \setminus S_2\rangle$ and hence $\{(j, h), (k, i)\}$ is a solution. Next, we consider S_3. This one contains a controversial attack as well, so we have to compute the solutions for enforcing a in $\langle A_2, R_2 \setminus S_3\rangle$, which gives us three solutions: $\{(c, a)\}$, $\{(h, f)\}$ and $\{(k, i)\}$. Since $\{(k, i), (j, h)\}$ is already a solution, we only take $\{(c, a)\}$ and $\{(h, f)\}$, giving us

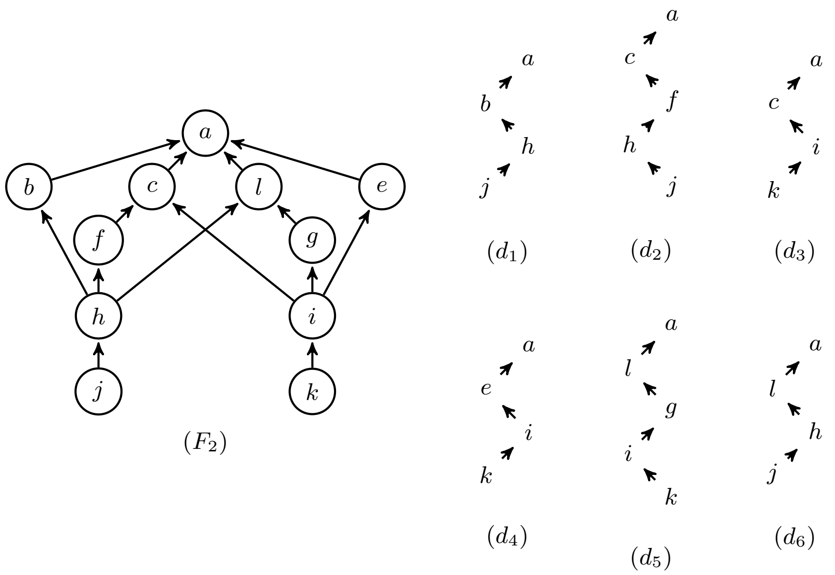

Fig. 1. Example argumentation framework F_2 and all final disputes for a in F_2.

two new solutions in F: $\{(j,h),(e,a),(c,a)\}$ and $\{(j,h),(e,a),(h,f)\}$. Lastly, S_4 also contains a controversial attack, so we once again iterate the procedure in $\langle A_2, R_2 \setminus S_4\rangle$, giving us three solutions: $\{(l,a)\}$, $\{(i,g)\}$ and $\{(j,h)\}$, of which we ignore $\{(j,h)\}$. In the end, we have 6 solutions:

1. $\{(e,a),(b,a)\}$
2. $\{(j,h),(k,i)\}$
3. $\{(e,a),(j,h),(c,a)\}$
4. $\{(e,a),(j,h),(h,f)\}$
5. $\{(b,a),(k,i),(l,a)\}$
6. $\{(b,a),(k,i),(i,g)\}$

Theorem 3. *Let $F = \langle A, R\rangle$ where $a \in A$ is rejected with respect to preferred semantics, and sols the set returned by calling* EnumAtksAcc$(F, a, \emptyset)$*. For all $s \in$ sols, a is credulously accepted in $\langle A, R \setminus s\rangle$ with respect to the preferred semantics, and for all R' such that $(R \setminus s) \subset R' \subseteq R$, a is rejected with respect to preferred semantics in $\langle A, R'\rangle$.*

Proof sketch: The minimality result follows from Lemma 1. Similarly, the correctness of the solutions follows from Theorem 2. In the cases of controversial attack removals, Lemma 1 might need to be applied multiple times while Theorem 2 applies only on the final framework. Note that since the initial set of attacks is finite, the algorithm is guaranteed to terminate. □

Algorithm 1 can be adapted to work for the enforcement of argument rejection by retrieving defending strategies in line 2 and computing critical defense sets in line 6 instead. This modified algorithm will then disrupt defending strategies in a similar way and thus give rise to at least one opposing strategy, ensuring the rejection of the argument in question.

4 Conclusion and future work

In this paper, we have described results in dialogue games for abstract argumentation frameworks which have allowed us to provide an algorithm for the computation of minimal sets of attacks which, when removed from the framework, enforce the credulous acceptability of a given argument with respect to preferred semantics, which can be easily modified to enforce rejection instead.

Future work could include similar procedures for other semantics, such as stable, semi-stable and ideal, of which the dialogue games have been briefly discussed by Modgil et al. [10], but also for semantics such as stage2 or cf2 [1]. One could also use the results in this paper to focus on the removal of arguments instead of attacks, or a mixture of both.

References

1. Pietro Baroni, Massimiliano Giacomin, and Giovanni Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 2005.
2. Ringo Baumann. What does it take to enforce an argument? minimal change in abstract argumentation. In *ECAI*, volume 12, pages 127–132, 2012.
3. Guido Boella, Souhila Kaci, and Leendert Van Der Torre. Dynamics in argumentation with single extensions: Abstraction principles and the grounded extension. In *ECSQARU*, pages 107–118. Springer, 2009.
4. Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *Journal of Artificial Intelligence Research*, 38:49–84, 2010.
5. Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. *KR*, 14:52–61, 2014.
6. Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.
7. Paul E Dunne and Trevor JM Bench-Capon. Two party immediate response disputes: Properties and efficiency. *Artificial Intelligence*, 149(2):221–250, 2003.
8. Simone Gabbriellini and Paolo Torroni. A new framework for ABMs based on argumentative reasoning. In *Advances in Social Simulation*, pages 25–36. Springer, 2014.
9. Beishui Liao, Li Jin, and Robert C Koons. Dynamics of argumentation systems: A division-based method. *Artificial Intelligence*, 175(11):1790–1814, 2011.
10. Sanjay Modgil and Martin Caminada. Proof theories and algorithms for abstract argumentation frameworks. In *Argumentation in artificial intelligence*, pages 105–129. Springer, 2009.
11. Ken Satoh and Takeaki Uno. Enumerating maximal frequent sets using irredundant dualization. In *International Conference on Discovery Science*, pages 256–268. Springer, 2003.
12. Gerard AW Vreeswik and Henry Prakken. Credulous and sceptical argument games for preferred semantics. In *European Workshop on Logics in Artificial Intelligence*, pages 239–253. Springer, 2000.
13. Ernst Zermelo. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In *Proceedings of the fifth international congress of mathematicians*, volume 2, pages 501–504. II, Cambridge UP, Cambridge, 1913.