How to Argue for Anything: Enforcing Arbitrary Sets of Labellings Using AFs

Sjur K. Dyrkolbotn
Durham Law School, Durham University, UK
s.k.dyrkolbotn@durham.ac.uk

Abstract

We contribute to the investigation of possible outcomes of argumentation under semantics formulated using argumentation frameworks (AFs). In particular, we study this question for the labelling-based formulation of such semantics, generalizing previous work which has focused on extensions. In this paper, we restrict attention to the preferred and semi-stable semantics, showing that as long as we have a sufficient number of fresh arguments available, we can in fact argue for anything. That is, for any set of finite labellings there is an AF that enforces exactly this set as the outcome of argumentation.

1 Introduction

Formal argumentation in the style of (Dung 1995) is becoming an increasingly important formalism in artificial intelligence.[1] Dung observed that by representing argumentation scenarios by directed graphs, referred to as argumentation frameworks (AFs) in this context, various denotational semantics for non-monotonic reasoning could be formulated using graph-theoretic terms. Such semantics work by prescribing to an AF a set of *extensions*, sets of arguments that are regarded as successful when held together. In (Caminada 2006) it was shown that extensions could be defined in terms of three-valued labellings to the argument set, with every argument obtaining the status of being either accepted, defeated or undetermined. Hence an AF can be viewed as a theory in three-valued logic, and it has since been observed that Łukasiewicz logic is particularly well-suited for reasoning about argumentation (Dyrkolbotn and Walicki 2013; Dyrkolbotn 2013).

In recent work, the question of the *signature* of an argumentation semantics has been studied, asking about the structure of the different possible sets of extensions that can arise from an AF under a given semantics (Dunne et al. 2013). This work is interesting in that it promises to provide a bird's-eye view on the behavior of argumentation semantics, allowing us to develop a better understanding of their expressive power and meta-logical properties.

[1] We point to (Rahwan and Simari 2009) for a recent volume devoted to the use of abstract argumentation theory in AI.

However, developing this theory with respect to sets of extensions is unsatisfactory since the behavior of argumentation semantics is fundamentally three-valued. In particular, it seems to us that the question should also be studied when outcomes of argumentation are represented more realistically as sets of labellings. In this paper we make a contribution in this regard, by characterizing the possible outcomes of finite argumentation under the preferred and semi-stable semantics.[2] In particular, we show that *every* set of labellings can be enforced, as long as we have a sufficient number of additional arguments available. The proof is constructive and also demonstrates that the time needed to enforce a given set of labellings is linear in the number of possible labellings that are not in the outcome.

The structure of the paper is as follows. In Section 2 we present necessary background on AFs and semantics formulated using labellings. Then in Section 3 we prove the main result by constructing canonical AFs that can be used to enforce sets of labellings. In Section 4 we offer a conclusion.

2 Background

We give a terse background on AFs and the labelling-based approach to argumentation semantics, for more details we point to (Rahwan and Simari 2009, Chapter 2). First let us fix a countably infinite set Π of arguments. Then we can define an AF simply as a set of directed edges $E \subseteq \Pi \times \Pi$. If $(p, q) \in E$ we think of it as encoding the fact that p represents an argument that attacks the argument represented by q. We use the notation $E^+(x) = \{y \in \Pi \mid (x, y) \in E\}, E^-(x) = \{y \in \Pi \mid (y, x) \in E\}$, and extend it to sets such that $E^*(A) = \bigcup_{x \in A} E^*(x)$ for $* \in \{+, -\}$. We also define $\Pi(E) = \{x \mid E^+(x) \cup E^-(x) \neq \emptyset\}$, the set of arguments from Π which appear in some attack from E. We will often present AFs by simply depicting the attacks they contain, as in Figure 1.

Given an AF E, an argumentation semantics is used to identify sets of arguments that can be held successfully together. Typically, this involves various formalizations of the

[2] We remark that (Dunne et al. 2013) also notes the possibility of studying signatures with respect to labellings rather than extensions. They have little to say about it, however, apart from listing it as an important direction for future work.

Figure 1: An AF E such that $\Pi(E) = \{p, q, q', p'\}$

intuition that the set should be internally consistent and able to defend itself against attacks from other arguments. Different semantics differ about the details, but they have the same signature: they are defined as an operator $\mathcal{S}$ which takes an AF E and returns a set of sets of $\mathcal{S}(E) \subseteq 2^{\Pi}$, the set of acceptable sets, called extensions. Moreover, to the best of our knowledge, all semantics share the property that arguments in an extension should be internally consistent, free of internal conflict. Formally, for all such semantics $\mathcal{S}$, all AFs E and all $A \in \mathcal{S}(E)$, we have $E^-(A) \subseteq \Pi \setminus A$, such that no two arguments in A attack each other.

For a given argument, it is accepted or it is not, but a boolean perspective fails to do justice to the nature of the structure $(E, \mathcal{S})$ in two important ways. First, it is not clear whether we should say that p is accepted on E under $\mathcal{S}$ when there *exists* some $A \in \mathcal{S}(E)$ such that $p \in A$, or whether we should require $p \in A$ for *all* such A. Both notions of acceptance have been used, and the former is typically dubbed *credulous* while the latter is referred to as *skeptical*.[3]

The second sense in which acceptance is not a binary notion has to do with the structure of E. In particular, given any $A \in \mathcal{S}(E)$ the status of p with respect to A can be any of the following:

$$\begin{array}{ll} 1 : p \in A & 2 : p \in E^+(A) \\ 3 : p \in \Pi \setminus (A \cup E^+(A)) & \end{array} \quad (1)$$

Notice that by conflict-freeness of A, it follows that if $p \in E^+(A)$ then $p \notin A$. Hence when the focus is on the status of individual arguments, we might as well view $\mathcal{S}(E)$ as a set of partitions of Π into three disjoint sets or, equivalently, as a collection of *labellings* (Caminada 2006), functions $\mathsf{c} : \Pi \to \{1, \frac{1}{2}, 0\}$ such that for all $x \in \Pi$:

$$\mathsf{c}(x) = 0 \iff \exists y \in E^-(x) : \mathsf{c}(y) = 1 \quad (2)$$

For any AF E we let $\mathsf{cf}(E)$ be the set of all labellings for E, and we define $\mathsf{c}^1 = \{x \in \Pi \mid \mathsf{c}(x) = 1\}, \mathsf{c}^0 = \{x \mid \mathsf{c}(x) = 0\}$ and $\mathsf{c}^{\frac{1}{2}} = \{x \in \Pi \mid \mathsf{c}(x) = \frac{1}{2}\}$. This, in particular, defines a semantics for argumentation such that for all E, we regard $A \subseteq \Pi$ as acceptable if there is some $\mathsf{c} \in \mathsf{cf}(E)$ such that $\mathsf{c}^1 = A$.[4] In applications of argumentation theory, this is usually considered too permissive, and a range of various restrictions has been considered, each giving rise to a new semantics, the most well-known of which are defined in Figure 2.

Admissible:	$a(E) = \{\mathsf{c} \in \mathsf{cf}(E) \mid E^-(\mathsf{c}^1) \subseteq \mathsf{c}^0\}$
Complete:	$c(E) = \{\mathsf{c} \in \mathsf{cf}(E) \mid \mathsf{c}^1 = \{x \in \Pi \mid E^-(x) \subseteq \mathsf{c}^0\}\}$
Grounded:	$g(E) = \{\bigcap c(E)\}$
Preferred:	$p(E) = \{\mathsf{c}_1 \in a(E) \mid \forall \mathsf{c}_2 \in a(E) : \mathsf{c}_1^1 \not\subset \mathsf{c}_2^1\}$
Semi-stable:	$ss(E) = \{\mathsf{c}_1 \in a(E) \mid \forall \mathsf{c}_2 \in a(E) : \mathsf{c}_1^{\frac{1}{2}} \not\supset \mathsf{c}_2^{\frac{1}{2}}\}$
Stable:	$s(E) = \{\mathsf{c} \in a(E) \mid \mathsf{c}^{\frac{1}{2}} = \emptyset\}$

Figure 2: Various semantics, defined for any $E \subseteq \Pi \times \Pi$

To illustrate the definition, consider the AF in Figure 1. We notice, in particular, that the set of admissible and complete extensions is $a(E) = c(E) = \{\emptyset, \{q\}, \{q'\}\}$, that the grounded extension is empty, that the preferred and semi-stable extensions are $p(E) = ss(E) = \{\{q\}, \{q'\}\}$ and that E does not have any stable extensions (the labellings corresponding to these sets can be easily recovered using Equation (1)).

We can now illustrate why the approach from (Dunne et al. 2013), which focuses on extensions rather than labellings, fails to do full justice to the question of characterizing the possible outcomes of argumentation. For instance, consider enforcing the outcome where one argument, call it w, is accepted, while all other arguments from Π are defeated. That is, there should be only one acceptable labelling and it should assign 1 to w and 0 to all other arguments. Looking at extensions alone there is no way to distinguish this from the case when w is accepted and all other arguments are undecided: in both cases it will hold that $\{w\}$ is the only extension. If we go on by attempting to enforce the outcome that *all* arguments in Π are rejected we come across a *validity* of argumentation that (Dunne et al. 2013) is unable to capture. It is easy to see, in particular, that while many AFs exist for which the empty set is the only admissible extension, it is impossible to construct an AF such that every argument is defeated, since then no arguments labelled 1 would be around to defeat them.

Due to validities of this type, it seems difficult in general to characterize the set of all infinite labellings that can be enforced without the addition of new arguments. In the following we therefore restrict attention to finite AFs, a restriction which is also made in (Dunne et al. 2013). The upshot is that we can always assume availability of additional arguments to help ensure that a given outcome for $V \subset \Pi$ is obtained.[5]

[3] It is natural to view skeptical and credulous acceptance as dual *modalities*, suggesting the study of the set of validities characterizing their interactions. This, in particular, is a different approach to modal reasoning about AFs than that explored in (Grossi 2010b; 2010a), where modalities are used to talk about AFs, to allow semantics to be defined in terms of modal formulas addressing the graph structure. It is closer to what is called an "object level" approach in (Caminada and Gabbay 2009), where arguments are regarded as atoms in a propositional language.

[4] Hence it is not hard to see that values assigned by labellings correspond to the three points of Equation 1 whenever we restrict attention to conflict-free sets of accepted arguments. Notice, in particular, that $p \in \mathsf{c}^0 \iff p \in E^+(\mathsf{c}^1)$ and $p \in \mathsf{c}^{\frac{1}{2}} \iff p \in \Pi \setminus (\mathsf{c}^1 \cup \mathsf{c}^0)$

[5] We remark that in the context of logic-based approaches to argumentation, such as those explored in (Dyrkolbotn and Walicki 2013; Dyrkolbotn 2013), this restriction is suitable as long as we do not consider languages with *infinitary* connectives. Moreover, we note that while the finiteness restriction is not in itself crucial and may be dropped, the presence of a sufficient amount of additional arguments is indeed needed for the results presented in the next section.

3 Enforcing an Arbitrary Set of Labellings

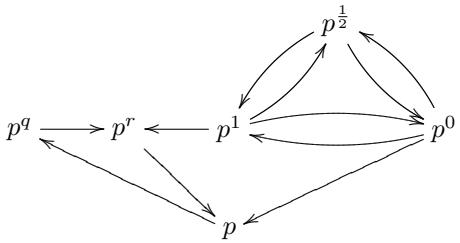
We now proceed to show how canonical AFs can be used to enforce arbitrary sets of labellings to finite $V \subset \Pi$ under preferred and semi-stable semantics. The same AFs will be canonical for both, so our result is relevant also to the investigation of translations between semantics (Dvorák and Woltran 2011; Dvorák and Spanring 2012). In particular, our constructions provide an alternative to the approach of modifying AFs so that their behavior under one semantics can be captured by another. For *any* argumentation semantics which prescribes labellings, we can capture it using preferred and semi-stable semantics, by using the canonical AF.

To construct it formally, assume we are given a finite $V \subseteq \Pi$ and a set of three-valued assignments $F \subseteq \left\{1, \frac{1}{2}, 0\right\}^V$. Then we will construct E_F and show that it enforces F on V in the following sense, for $\mathcal{S} \in \{p, ss\}$:

$$\{\mathsf{c}|_V \mid \mathsf{c} \in \mathcal{S}(E_F)\} = F \tag{3}$$

We rely on two basic constructions to show that there exists E_F which satisfies this equation, for all V, F. First we introduce, for all $p \in \Pi$, the AF Ł_p, which we refer to as a *circuit* for p. It is defined as depicted below.

(4)



The reader may easily verify that Ł_p satisfies the following property, for both preferred and semi-stable $\mathcal{S}$:

$$\{\mathsf{c}|_p \mid \mathsf{c} \in \mathcal{S}(\text{Ł}_p)\} = \{\{p \mapsto 0\}, \{p \mapsto 1\}, \{p \mapsto \frac{1}{2}\}\} \tag{5}$$

That is, in the AF Ł_p, the semantic status of p is completely open: for any of the three values p could have, the preferred and semi-stable semantics admit labellings for Ł_p that give p this value. This implies that for any finite set $V = \{p_1, \ldots, p_n\}$ there is an AF which characterizes the set $X = \left\{1, \frac{1}{2}, 0\right\}^V$, containing *all* functions from V to $\left\{1, \frac{1}{2}, 0\right\}$. In particular, we can take E_X to be the following AF (assuming, of course, that all arguments p_i^x used to construct a circuit for p_i are fresh, i.e. do not occur in V).

$$E_X = \bigcup_{1 \leq i \leq n} \text{Ł}_{p_i} \tag{6}$$

To show the general claim, for arbitrary $F \subseteq X$, we will start from E_X and then inductively define E_F in terms of $E_{F'}$ such that $F = F' \setminus \{f\}$. We will show, in particular, how to obtain E_F from $E_{F'}$ by ensuring that f becomes *forbidden* while all other assignments from F' are still permitted. To explain how to do this, let us first observe the following crucial property, allowing us to characterize any of the three values that p_i can have by an argument that is accepted if, and only if, p_i has this value. In particular, the reader can easily verify that the following holds, for preferred and semi-stable $\mathcal{S}$ and all $\mathsf{c} \in \mathcal{S}(\text{Ł}_{p_i})$:

$$\begin{cases} \mathsf{c}(p_i) = 1 \iff \mathsf{c}(p_i^1) = 1 \\ \mathsf{c}(p_i) = 0 \iff \mathsf{c}(p_i^0) = 1 \\ \mathsf{c}(p_i) = \frac{1}{2} \iff \mathsf{c}(p_i^{\frac{1}{2}}) = 1 \end{cases} \tag{7}$$

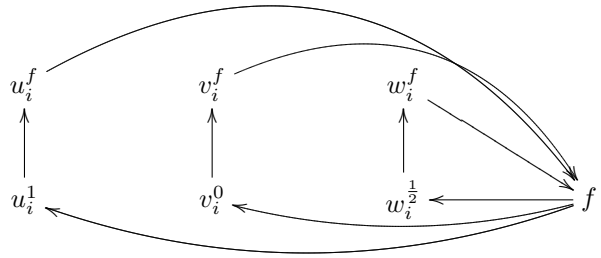
Notice also that $\mathsf{c}(p_i^y) \in \{1, 0\}$ for all y, i and $\mathsf{c} \in \mathcal{S}(\text{Ł}_{p_i})$. That is, the arguments witnessing to the value of p_i are always assigned one of the boolean values. This will be important to keep in mind below. Now, assume as induction hypothesis that we have been able to define $E_{F'}$ for $F' = F \cup \{f\}$ based on E_X such that Equation (3) holds for F', and such that Equation (7) holds for all $\mathsf{c} \in \mathcal{S}(E_{F'})$. To define E_F we first partition V according to the values assigned by f:

$$V = V_1 \cup V_0 \cup V_{\frac{1}{2}}$$

where $V_1 = \{x \mid f(x) = 1\} = \{u_1, \ldots, u_k\}, V_0 = \{x \mid f(x) = 0\} = \{v_1, \ldots, v_l\}, V_{\frac{1}{2}} = \{x \mid f(x) = \frac{1}{2}\} = \{w_1, \ldots, w_m\}$ (notice that this implies that all V_i's are disjoint). Next we let $\{f, x^f \mid x \in V\}$ be a collection of $|V|+1$ fresh arguments, and we define E_F as follows:

$$\begin{aligned} E_F = E_{F'} &\cup \bigcup_{1 \leq i \leq k} \{(u_i^1, u_i^f), (u_i^f, f), (f, u_i^1)\} \\ &\cup \bigcup_{1 \leq i \leq l} \{(v_i^0, v_i^f), (v_i^f, f), (f, v_i^0)\} \\ &\cup \bigcup_{1 \leq i \leq m} \{(w_i^{\frac{1}{2}}, w_i^f), (w_i^f, f), (f, w_i^{\frac{1}{2}}\} \end{aligned}$$

In the figure below we sketch this construction, depicting the pattern made up of the edges added in the induction step.



(8)

First let us note that Equation 7 holds for all $\mathsf{c} \in \mathcal{S}(E_F)$. Indeed, none of the edges that have been added interfere with the fact that for all $x_i \in V$, we have $\mathsf{c}(x_i^y) = 1$ if, and only if, $\mathsf{c}(x_i) = y$. It is then straightforward to establish that there is no $\mathsf{c} \in \mathcal{S}(E_F)$ such that $\mathsf{c}|_V = f$. In particular, if there was such a c it would follow that all arguments of the form $x_i^{f(x_i)}$ would be assigned 1, so all attackers of f would be labelled 0, which in turn would imply that the fresh argument f would be labelled 1. This would be contradiction, however, since f attacks all $x_i^{f(x_i)}$. On the other hand, it is easy to see that any other $f' \in F'$ is still permitted: as long as one or more arguments from V are labelled by something

different from their value under f, it follows from Equation (7) that the argument named f will be rejected, thus rendering all new attacks on existing arguments, which all come from f, irrelevant to the labelling of arguments from V.

Having demonstrated how to construct the desired AF E_F by induction, we conclude as follows.

Theorem 3.1. *For all finite $V \subseteq \Pi$, and all $F \subseteq \{1, \frac{1}{2}, 0\}^V$, there is a finite AF E_F such that for all $\mathcal{S} \in \{p, ss\}$ we have*

$$\{\mathsf{c}|_V \mid \mathsf{c} \in \mathcal{S}(E_F)\} = F$$

The proof of this result essentially consisted in an algorithm for computing E_F. To analyze the complexity of this construction, notice first that E_X has linear size in V. Moreover, remember that we needed $|V| + 1$ new arguments in each step of the procedure, and that the number of total steps required was equal to the number of forbidden labellings. In light of this, we obtain the following corollary.

Corollary 3.2. *For any finite $V \subseteq \Pi$ and any $F \subseteq \{1, \frac{1}{2}, 0\}^V$, we can construct E_F that enforces F under preferred and semi-stable semantics in time $\mathcal{O}(|V| \times |\{1, \frac{1}{2}, 0\}^V \setminus F|)$.*

Before we conclude, we comment briefly on the other semantics from Figure 2. It is easy to see, in particular, that Theorem 3.1 does not hold for any of the other semantics defined there. For the stable semantics this is trivial since it only permits boolean labellings.[6] For the other semantics, notice that all of them always prescribe a labelling such that all other permissible labellings extend it (the empty labelling for the admissible semantics, the grounded labelling for the others). This means, in particular, that no set of labellings can be enforced if it fails to include this labelling.

4 Conclusion

We have studied the finite signatures of the preferred and semi-stable semantics, in terms of the sets of finite labellings they may give rise to. As it turns out, any set of labellings is a possible outcome under these semantics, so in a sense we have indeed shown how to argue for anything. Moreover, while the construction we used takes exponential time to compute in general, it is linear in an interesting parameter: the number of labellings that are *not* supposed to be in the outcome.

This result provides theoretical insight concerning the preferred and semi-stable semantics and it also strengthens the connection between argumentation and three-valued Łukasiewicz logic. As observed in (Dyrkolbotn 2013), any semantics applied to an AF can be seen as a theory in this logic. Hence Theorem 3.1 serves to complete the picture by showing that any finite theory of Łukasiewicz logic can also be represented as an AF. That is, for any such theory there is a corresponding AF such that a three-valued assignment to atoms is a model of the theory if, and only if, it corresponds to a preferred/semi-stable labelling for the corresponding AF.

This is interesting in its own right, but also suggests the possibility of obtaining completeness results for modal logics that allow us to reason about skeptical and credulous acceptance under preferred and semi-stable semantics. It seems, in particular, that we are able to instantiate a canonical class of three-valued doxastic Kripke frames using AFs.[7] This will be explored further in future work.

[6] However, notice that the argument used to prove Theorem 3.1 can be easily adapted to show that the stable semantics is canonical for boolean theories.

[7] In particular, by taking labellings as worlds and letting belief sets be induced by an AF under an argumentation semantics.

References

Caminada, M. W. A., and Gabbay, D. M. 2009. A logical account of formal argumentation. *Studia Logica* 93(2-3):109–145.

Caminada, M. W. A. 2006. On the issue of reinstatement in argumentation. In Fisher, M.; van der Hoek, W.; Konev, B.; and Lisitsa, A., eds., *JELIA*, volume 4160 of *Lecture Notes in Computer Science*, 111–123. Springer.

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

Dunne, P. E.; Dvorák, W.; Linsbichler, T.; and Woltran, S. 2013. Characteristics of multiple viewpoints in abstract argumentation. In Beierle, C., and Kern-Isberner, G., eds., *4th Workshop on Dynamics of Knowledge and Belief (DKB-2013)*. FernUniversität in Hagen. 16–31.

Dvorák, W., and Spanring, C. 2012. Comparing the expressiveness of argumentation semantics. In Verheij, B.; Szeider, S.; and Woltran, S., eds., *COMMA*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, 261–272. IOS Press.

Dvorák, W., and Woltran, S. 2011. On the intertranslatability of argumentation semantics. *J. Artif. Intell. Res. (JAIR)* 41:445–475.

Dyrkolbotn, S. K., and Walicki, M. 2013. Propositional discourse logic. *Synthese* 1–37. Available online first at Springer.

Dyrkolbotn, S. K. 2013. The same, similar, or just completely different? Equivalence for argumentation in light of logic. In Libkin, L.; Kohlenbach, U.; and de Queiroz, R. J. G. B., eds., *WoLLIC*, volume 8071 of *Lecture Notes in Computer Science*, 96–110. Springer.

Grossi, D. 2010a. Argumentation in the view of modal logic. In McBurney, P.; Rahwan, I.; and Parsons, S., eds., *ArgMAS*, volume 6614 of *Lecture Notes in Computer Science*, 190–208. Springer.

Grossi, D. 2010b. On the logic of argumentation theory. In van der Hoek, W.; Kaminka, G. A.; Lespérance, Y.; Luck, M.; and Sen, S., eds., *AAMAS*, 409–416. IFAAMAS.

Rahwan, I., and Simari, G. R., eds. 2009. *Argumentation in artificial intelligence*. Springer.