Verifiability of Argumentation Semantics

Ringo BAUMANN [a], Thomas LINSBICHLER [b], and Stefan WOLTRAN [b]
[a] *Computer Science Institute, University of Leipzig, Germany*
[b] *Institute of Information Systems, TU Wien, Austria*

Abstract. Dung's abstract argumentation theory is a widely used formalism to model conflicting information and to draw conclusions in such situations. Hereby, the knowledge is represented by so-called argumentation frameworks (AFs) and the reasoning is done via semantics extracting acceptable sets. All reasonable semantics are based on the notion of conflict-freeness which means that arguments are only jointly acceptable when they are not linked within the AF. In this paper, we study the question which information on top of conflict-free sets is needed to compute extensions of a semantics at hand. We introduce a hierarchy of so-called verification classes specifying the required amount of information. We show that well-known standard semantics are exactly verifiable through a certain such class. Our framework also gives a means to study semantics lying inbetween known semantics, thus contributing to a more abstract understanding of the different features argumentation semantics offer.

Keywords. abstract argumentation, argumentation semantics, verifiability, strong equivalence, intermediate semantics

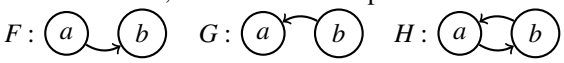
1. Introduction

In the late 1980s the idea of using *argumentation* to model nonmonotonic reasoning emerged (see [1,2] as well as the survey [3]). Nowadays argumentation theory is a vibrant subfield of Artificial Intelligence, covering aspects of knowledge representation, multi-agent systems, and also philosophical questions. Among other approaches which have been proposed for capturing representative patterns of inference in argumentation theory [4], Dung's abstract argumentation frameworks (AFs) [5] play an important role within this research area. At the heart of Dung's approach lie the so-called *argumentation semantics* (cf. [6] for an excellent overview). Given an AF F, which is set-theoretically just a directed graph encoding arguments and attacks between them, a certain argumentation semantics σ returns acceptable sets of arguments $\sigma(F)$, so-called σ-*extensions*. Each of these sets represents a reasonable position w.r.t. F and σ.

Over the last 20 years a series of abstract argumentation semantics were introduced. The motivations of these semantics range from the desired treatment of specific examples to fulfilling a number of abstract principles. The comparison via abstract criteria of the different semantics available is a topic which emerged quite recently in the community ([7] can be seen as the first paper in this line). Our work takes a further step towards a comprehensive understanding of argumentation semantics. In particular, we study the following question: Do we really need the entire AF F to compute a certain argumentation semantics σ? In other words, is it possible to unambiguously determine acceptable

sets w.r.t. σ, given only partial information of the underlying framework F. In order to solve this problem let us start with the following reflections:

1. As a matter of fact, one basic requirement of almost all existing semantics (exemptions are [8,9,10]) is that of conflict-freeness, i.e. arguments within a reasonable position are not allowed to attack each other. Consequently, knowledge about conflict-free sets is an essential part for computing semantics.
2. The second step is to ask the following: Which information on top on conflict-free sets has to be added? Imagine the set of conflict-free sets given by $\{\emptyset, \{a\}, \{b\}\}$. Consequently, there has to be at least one attack between a and b. Unfortunately, this information is not sufficient to compute any standard semantics (except naive extensions, which are defined as $\subseteq$-maximal conflict-free sets) since we know nothing precise about the neighborhood of a and b. The following three AFs possess exactly the mentioned conflict-free sets, but differ with respect to other semantics.



3. The final step is to try to minimize the added information. In other words, which kind of knowledge about the neighborhood is somehow dispensable in the light of computation? Clearly, this will depend on the considered semantics. For instance, in case of stage semantics [11], which requests conflict-free sets of maximal range, we do not need any information about incoming attacks. This information can not be omitted in case of admissible-based semantics since incoming attacks require counterattacks.

The above considerations motivate the introduction of *verification classes* specifying a certain amount of information. In a first step, we study the relation of these classes to each other. We therefore introduce the notion of being *more informative*, capturing the intuition that a certain class can reproduce the information of another. We present a hierarchy w.r.t. this ordering, containing 15 different verification classes only. This is because many syntactically different classes collapse to the same amount of information.

We then formally define the essential property of a semantics σ being *verifiable* w.r.t. a certain verification class. We present a general theorem stating that any *rational* semantics is exactly verifiable w.r.t. one of the 15 different verification classes. Roughly speaking, a semantics is rational if attacks inbetween two self-loops can be omitted without affecting the set of extensions. An important aside hereby is that even the most informative class contains indeed less information than the entire framework by itself.

In this paper we consider a representative set of standard semantics. All of them satisfy rationality and thus, are exactly verifiable w.r.t. a certain class. Since the theorem does not provide an answer to which verification class perfectly matches a certain rational semantics we study this problem one by one for any considered semantics. As a result, only 6 different classes are essential to classify the considered standard semantics.

In the last part of the paper we study an application of the concept of verifiability. More precisely, we address the question of strong equivalence for semantics lying inbetween known semantics, so-called intermediate semantics. Strong equivalence is the natural counterpart to ordinary equivalence in monotonic theories (see [12,13] for abstract argumentation and [14,15,16,17] for other nonmonotonic theories). We provide characterization theorems relying on the notion of verifiability and thus, contributing to a more abstract understanding of the different features argumentation semantics offer. Besides these main results, we also give new characterizations for strong equivalence with respect to naive extensions and strong admissible sets [7,18].

Due to limited space we have to refer to an extended version [19] for full proofs.

2. Preliminaries

An *argumentation framework* (AF) $F = (A,R)$ is a directed graph whose nodes $A \subseteq \mathscr{U}$ (with $\mathscr{U}$ being an infinite set of arguments, so-called *universe*) are interpreted as *arguments* and whose edges $R \subseteq A \times A$ represent *conflicts* between them. We assume that all AFs possess finitely many arguments only and denote the collection of all AFs by $\mathscr{A}$. If $(a,b) \in R$ we say that *a attacks b*. Alternatively, we write $a \rightarrowtail b$ as well as, for some $S \subseteq A$, $a \rightarrowtail S$ or $S \rightarrowtail b$ if there is some $c \in S$ attacked by a or attacking b, respectively. An argument $a \in A$ is *defended* by a set $S \subseteq A$ if for each $b \in A$ with $b \rightarrowtail a$, $S \rightarrowtail b$. We define the *range* of S (in F) as $S_F^+ = S \cup \{a \mid S \rightarrowtail a\}$. Similarly, we use S_F^- to denote the *anti-range* of S (in F) as $S \cup \{a \mid a \rightarrowtail S\}$. Furthermore, we say that a set S is *conflict-free* (in F) if there is no argument $a \in S$ s.t. $S \rightarrowtail a$. The set of all conflict-free sets of an AF F is denoted by $cf(F)$. For an AF $F = (B,S)$ we use $A(F)$ and $R(F)$ to refer to B and S, respectively. Furthermore, we use $L(F) = \{a \mid (a,a) \in R(F)\}$ for the set of all self-defeating arguments. Finally, we introduce the union of AFs F and G as $F \cup G = (A(F) \cup A(G), R(F) \cup R(G))$.

A *semantics* σ assigns to each $F = (A,R)$ a set $\sigma(F) \subseteq 2^A$ where the elements are called σ*-extensions*. Numerous semantics are available. Each of them captures different intuitions about how to reason about conflicting knowledge. We consider $\sigma \in \{ad, na, stb, pr, co, gr, ss, stg, id, eg\}$ for admissible, naive, stable, preferred, complete, grounded, semi-stable, stage, ideal, and eager semantics [5,11,20,21,22].

Definition 1. Given an AF $F = (A,R)$ and let $S \subseteq A$.

1. $S \in ad(F)$ iff $S \in cf(F)$ and each $a \in S$ is defended by S,
2. $S \in na(F)$ iff $S \in cf(F)$ and there is no $S' \in cf(F)$ s.t. $S \subsetneq S'$,
3. $S \in stb(F)$ iff $S \in cf(F)$ and $S_F^+ = A$,
4. $S \in pr(F)$ iff $S \in ad(F)$ and there is no $S' \in ad(F)$ s.t. $S \subsetneq S'$,
5. $S \in co(F)$ iff $S \in ad(F)$ and for any $a \in A$ defended by S, $a \in S$,
6. $S \in gr(F)$ iff $S \in co(F)$ and there is no $S' \in co(F)$ s.t. $S' \subsetneq S$,
7. $S \in ss(F)$ iff $S \in ad(F)$ and there is no $S' \in ad(F)$ s.t. $S_F^+ \subsetneq S_F'^+$,
8. $S \in stg(F)$ iff $S \in cf(F)$ and there is no $S' \in cf(F)$ s.t. $S_F^+ \subsetneq S_F'^+$,
9. $S \in id(F)$ iff $S \in ad(F)$, $S \subseteq \bigcap pr(F)$ and $\nexists S' \in ad(F)$ s.t. $S' \subseteq \bigcap pr(F) \wedge S \subsetneq S'$,
10. $S \in eg(F)$ iff $S \in ad(F)$, $S \subseteq \bigcap ss(F)$ and $\nexists S' \in ad(F)$ s.t. $S' \subseteq \bigcap ss(F) \wedge S \subsetneq S'$.

For two semantics σ, τ we use $\sigma \subseteq \tau$ to indicate that $\sigma(F) \subseteq \tau(F)$ for each AF $F \in \mathscr{A}$. If we have $\rho \subseteq \sigma$ and $\sigma \subseteq \tau$ for semantics ρ, σ, τ, we say that σ is ρ-τ-*intermediate*. Well-known relations between semantics are $stb \subseteq ss \subseteq pr \subseteq co \subseteq ad$, meaning, for instance, that *ss* is *stb*-*pr*-intermediate.

Definition 2. We call a semantics σ *rational* if self-loop-chains are irrelevant. That is, for every AF F it holds that $\sigma(F) = \sigma(F^l)$, where $F^l = (A(F), R(F) \setminus \{(a,b) \in R(F) \mid (a,a),(b,b) \in R(F), a \neq b\})$.

Indeed, all semantics introduced in Definition 1 are rational. A prominent semantics that is based on conflict-free sets, but is not rational is the *cf2*-semantics [23], since here chains of self-loops can have an influence on the SCCs of an AF (see also [24]).

The main notions of equivalence available for non-monotonic formalisms are *ordinary* (or *standard*) *equivalence* and *strong* (or *expansion*) *equivalence*. A detailed overview of equivalence notion including their relations can be found in [25,26].

Definition 3. Given a semantics σ. Two AFs F and G are *standard equivalent* w.r.t. σ ($F \equiv^{\sigma} G$) iff $\sigma(F) = \sigma(G)$, and *expansion equivalent* w.r.t. σ ($F \equiv_E^{\sigma} G$) iff for each AF H: $F \cup H \equiv^{\sigma} G \cup H$.

Expansion equivalence can be decided syntactically via so-called *kernels* [12]. A kernel is a function $k : \mathscr{A} \mapsto \mathscr{A}$ mapping each AF F to another AF $k(F)$ (which we may also denote as F^k). Consider the following definitions.

Definition 4. Given an AF $F = (A,R)$ and a semantics σ. We define σ-kernels $F^{k(\sigma)} = (A, R^{k(\sigma)})$ whereby

- $R^{k(stb)} = R \setminus \{(a,b) \mid a \neq b, (a,a) \in R\}$,
- $R^{k(ad)} = R \setminus \{(a,b) \mid a \neq b, (a,a) \in R, \{(b,a),(b,b)\} \cap R \neq \emptyset\}$,
- $R^{k(gr)} = R \setminus \{(a,b) \mid a \neq b, (b,b) \in R, \{(a,a),(b,a)\} \cap R \neq \emptyset\}$,
- $R^{k(co)} = R \setminus \{(a,b) \mid a \neq b, (a,a),(b,b) \in R\}$.

We say that a relation $\equiv \subseteq \mathscr{A} \times \mathscr{A}$ is *characterizable through kernels* if there is a kernel k, s.t. $F \equiv G$ iff $F^k = G^k$. Moreover, we say that a semantics σ is *compatible with a kernel k* if $F \equiv_E^{\sigma} G$ iff $F^k = G^k$. All semantics (except naive semantics) considered in this paper are compatible with one of the four kernels introduced above.

Theorem 1. *[12,27] For any AFs F and G,*

1. $F \equiv_E^{\sigma} G \Leftrightarrow F^{k(\sigma)} = G^{k(\sigma)}$ *with* $\sigma \in \{stb, ad, co, gr\}$,
2. $F \equiv_E^{\tau} G \Leftrightarrow F^{k(ad)} = G^{k(ad)}$ *with* $\tau \in \{pr, id, ss, eg\}$,
3. $F \equiv_E^{stg} G \Leftrightarrow F^{k(stb)} = G^{k(stb)}$.

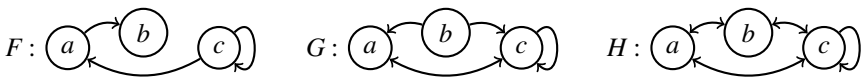
In Section 3 we will complete these results taking naive semantics and strongly admissible sets into account.

3. Complementing Previous Results

In order to provide an exhaustive analysis of intermediate semantics (cf. Section 5) we provide missing kernels for naive semantics as well as strongly admissible sets. We start with the *naive kernel* characterizing expansion equivalence w.r.t. naive semantics. Note that the following kernel is the first one which adds attacks to the former attack relation.

Definition 5. Given an AF $F = (A,R)$. We define the *naive kernel* $F^{k(na)} = (A, R^{k(na)})$ whereby $R^{k(na)} = R \cup \{(a,b) \mid a \neq b, \{(a,a),(b,a),(b,b)\} \cap R \neq \emptyset\}$.

Example 1. Consider the AFs F and G. Note that $na(F) = na(G) = \{\{a\},\{b\}\}$. Consequently, $F \equiv^{na} G$. In accordance with Definition 5 we observe that both AFs possess the same naive kernel $H = F^{k(na)} = G^{k(na)}$.



The following theorem shows that possessing the same kernels is necessary as well as sufficient for being strongly equivalent, i.e. $F \equiv_E^{na} G$.

Theorem 2. *For all AFs F and G, it holds that* $F \equiv_E^{na} G \Leftrightarrow F^{k(na)} = G^{k(na)}$.

We turn now to *strongly admissible sets* [7]. We will see that, besides grounded [12] and resolution-based grounded semantics [28,29], strongly admissible sets are characterizable through the grounded kernel. Consider the following definition from [18].

Definition 6. Given an AF $F = (A,R)$. A set $S \subseteq A$ is *strongly admissible*, i.e. $S \in sad(F)$ iff any $a \in S$ is defended by a strongly admissible set $S' \subseteq S \setminus \{a\}$.

The following properties are needed to prove the characterization theorem. (1) and (2) are already shown in [7], (3) is an immediate consequence of the former.

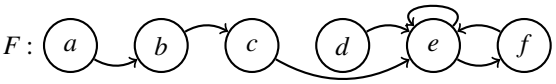
Proposition 3. *Given two AFs F and G, it holds that*

1. $gr(F) \subseteq sad(F) \subseteq ad(F)$,
2. *if* $S \in gr(F)$ *we have:* $S' \subseteq S$ *for all* $S' \in sad(F)$, *and*
3. $sad(F) = sad(G)$ *implies* $gr(F) = gr(G)$.

We now provide an alternative criterion for being a strongly admissible set. In contrast to the former it allows one to construct strongly admissible sets step by step. A proof that Definitions 6 and 7 are equivalent can be found in [19].

Definition 7. Given an AF $F = (A,R)$. A set $S \subseteq A$ is *strongly admissible*, i.e. $S \in sad(F)$ iff there are finitely many and pairwise disjoint sets $A_1,...,A_n$, s.t. $S = \bigcup_{1 \leq i \leq n} A_i$ and $A_1 \subseteq \Gamma_F(\emptyset)$[1] and furthermore, $\bigcup_{1 \leq i \leq j} A_i$ defends A_{j+1} for $1 \leq j \leq n-1$.

Example 2. Consider the following AF F.



We have $\Gamma_F(\emptyset) = \{a,d\}$. Hence, for all $S \subseteq \{a,d\}$, $S \in sad(F)$. Furthermore, $\Gamma_F(\{a\}) = \{a,c\}$, $\Gamma_F(\{d\}) = \{d,f\}$ and $\Gamma_F(\{a,d\}) = \{a,d,c,f\}$. This means, additionally $\{a,c\},\{d,f\},\{a,d,c\},\{a,d,f\},\{a,d,c,f\} \in sad(F)$. Finally, $\Gamma_F(\{a,c\}) = \{a,c,f\}$ justifying the last missing set $\{a,c,f\} \in sad(F)$.

The grounded kernel is insensitive w.r.t. strongly admissible sets, which then allows us to state the main result for strongly admissible sets.

Lemma 4. *For any AF F,* $sad(F) = sad(F^{k(gr)})$.

Theorem 5. *For any two AFs F and G, we have* $F \equiv_E^{sad} G \Leftrightarrow F^{k(gr)} = G^{k(gr)}$.

Proof. ($\Rightarrow$) We show the contrapositive, i.e. $F^{k(gr)} \neq G^{k(gr)} \Rightarrow F \not\equiv_E^{sad} G$. Assuming $F^{k(gr)} \neq G^{k(gr)}$ implies $F \not\equiv_E^{gr} G$ (cf. Theorem 1). This means, there is an AF H, s.t. $gr(F \cup H) \neq gr(G \cup H)$. Due to statement 3 of Proposition 3, we deduce $sad(F \cup H) \neq sad(G \cup H)$ proving $F \not\equiv_E^{sad} G$. ($\Leftarrow$) Given $F^{k(gr)} = G^{k(gr)}$. Since expansion equivalence is a congruence w.r.t. $\cup$ we obtain $(F \cup H)^{k(gr)} = (G \cup H)^{k(gr)}$ for any AF H. Consequently, $sad((F \cup H)^{k(gr)}) = sad((G \cup H)^{k(gr)})$. Due to Lemma 4 we deduce $sad(F \cup H) = sad(G \cup H)$, concluding the proof. □

[1] Hereby, Γ is the so-called *characteristic function* [5] with $\Gamma_F(S) = \{a \in A \mid a \text{ is defended by } S \text{ in } F\}$. The term $\Gamma_F(\emptyset)$ can be equivalently replaced by $\{a \in A \mid a \text{ is unattacked}\}$.

4. Verifiability

In this section we study the question whether we really need the entire AF F to compute the extensions of a given semantics. Consider naive semantics. Obviously, in order to determine naive extensions it suffices to know all conflict-free sets. Conversely, knowing $cf(F)$ only does not allow to reconstruct F unambiguously. This means, knowledge about $cf(F)$ is indeed less information than the entire AF by itself. In fact, most of the existing semantics do not need information about the entire AF. We will categorize the amount of information by taking the conflict-free sets as a basis and distinguish between different amounts of knowledge about the neighborhood (range and anti-range) of these sets.

Definition 8. We call a function $\mathfrak{r}^x : 2^{\mathscr{U}} \times 2^{\mathscr{U}} \to (2^{\mathscr{U}})^n$ $(n > 0)$, which is expressible via basic set operations only, *neighborhood function*. A neighborhood function $\mathfrak{r}^x$ induces the *verification class* mapping each AF F to $\widetilde{F}^x = \{(S, \mathfrak{r}^x(S_F^+, S_F^-)) \mid S \in cf(F)\}$.

We coined the term neighborhood function because the induced verification classes apply these functions to the neighborhoods, i.e. range and anti-range of conflict-free sets. The notion of *expressible via basic set operations* simply means that (in case of $n = 1$) the expression $\mathfrak{r}^x(A,B)$ is in the language generated by the BNF $X ::= A \mid B \mid (X \cup X) \mid (X \cap X) \mid (X \setminus X)$. Consequently, in case of $n = 1$, we may distinguish eight set theoretically different neighborhood functions, namely

$$\mathfrak{r}^{\varepsilon}(S,S') = \emptyset \qquad \mathfrak{r}^{+}(S,S') = S \qquad \mathfrak{r}^{-}(S,S') = S' \qquad \mathfrak{r}^{\mp}(S,S') = S' \setminus S$$
$$\mathfrak{r}^{\pm}(S,S') = S \setminus S' \quad \mathfrak{r}^{\cap}(S,S') = S \cap S' \quad \mathfrak{r}^{\cup}(S,S') = S \cup S' \quad \mathfrak{r}^{\Delta}(S,S') = (S \cup S') \setminus (S \cap S')$$

The names of the neighborhood functions are inspired by their usage in the verification classes they induce (cf. Definition 8). A verification class encapsulates a certain amount of information about an AF, as the following example illustrates.

Example 3. Consider the AF $F = (\{a,b,c\}, \{(a,b),(b,a),(b,b),(c,b)\})$. Now take, for instance, the verification class induced by $\mathfrak{r}^+$, that is $\widetilde{F}^+ = \{(S, \mathfrak{r}^+(S_F^+, S_F^-)) \mid S \in cf(F)\} = \{(S, S_F^+) \mid S \in cf(F)\}$, storing information about conflict-free sets together with their associated ranges w.r.t. F. It contains the following tuples: $(\emptyset,\emptyset)$, $(\{a\},\{a,b\})$, $(\{c\},\{b,c\})$, and $(\{a,c\},\{a,b,c\})$. For he verification class induced by $\mathfrak{r}^{\pm}$, on the other hand, we have $\widetilde{F}^{\pm} = \{(\emptyset,\emptyset),(\{a\},\emptyset),(\{c\},\{b\}),(\{a,c\},\emptyset)\}$.

Intuitively, it should be clear that the set $\widetilde{F}^+$ suffices to compute stage extensions (i.e., range-maximal conflict-free sets) of F. This intuitive understanding of *verifiability* will be formally specified in Definition 10. Note that a neighborhood function $\mathfrak{r}^x$ may return n-tuples. Consequently, in consideration of the eight basic functions we obtain (modulo reordering, duplicates, empty set) $2^7 + 1$ syntactically different neighborhood functions and therefore the same number of verification classes. As usual, we denote the n-ary combination of basic functions $(\mathfrak{r}^{x_1}(S,S'), \ldots, \mathfrak{r}^{x_n}(S,S'))$ as $\mathfrak{r}^x(S,S')$ by $x = x_1 \ldots x_n$.

With the following definition we can put neighborhood functions into relation w.r.t. their information. This will help us to show that actually many of the induced classes collapse to the same amount of information.

Definition 9. Given neighborhood functions $\mathfrak{r}^x$ and $\mathfrak{r}^y$ returning n-tuples and m-tuples, respectively, we say that $\mathfrak{r}^x$ is *more informative* than $\mathfrak{r}^y$, for short $\mathfrak{r}^x \succeq \mathfrak{r}^y$, iff there is a

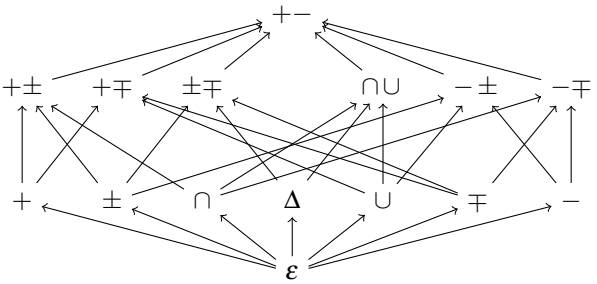


Figure 1. Representatives of neighborhood functions and their relation w.r.t. information; a node x stands for the neighborhood function $\mathfrak{r}^x$; an arrow from x to y means $\mathfrak{r}^x \prec \mathfrak{r}^y$.

function $\delta : (2^{\mathscr{U}})^n \to (2^{\mathscr{U}})^m$ such that for any two sets of arguments $S, S' \subseteq \mathscr{U}$, we have $\delta(\mathfrak{r}^x(S,S')) = \mathfrak{r}^y(S,S')$. We denote the strict part of $\succeq$ by $\succ$, i.e. $\mathfrak{r}^x \succ \mathfrak{r}^y$ iff $\mathfrak{r}^x \succeq \mathfrak{r}^y$ and $\mathfrak{r}^y \not\succeq \mathfrak{r}^x$. Finally, $\mathfrak{r}^x \approx \mathfrak{r}^y$ ($\mathfrak{r}^x$ *represents* $\mathfrak{r}^y$ and vice versa) in case $\mathfrak{r}^x \succeq \mathfrak{r}^y$ and $\mathfrak{r}^y \succeq \mathfrak{r}^x$.

It turns out that many neighborhood functions amount to the same amount of information. In particular, $\mathfrak{r}^{+-}$ represents all $\mathfrak{r}^{x_1,\ldots,x_n}$ with $n > 2$.

Lemma 6. *All neighborhood functions are represented by the ones depicted in Figure 1 and the $\prec$-relation represented by arcs in Figure 1 holds.*

If the information provided by a neighborhood function is sufficient to compute the extensions under a semantics, we say that the semantics is verifiable by the class induced by the neighborhood function.

Definition 10. A semantics σ is *verifiable* by the verification class induced by the neighborhood function $\mathfrak{r}^x$ returning n-tuples (or simply, *x-verifiable*) iff there is a function (also called *criterion*) $\gamma_\sigma : (2^{\mathscr{U}})^n \times 2^{\mathscr{U}} \to 2^{2^{\mathscr{U}}}$ s.t. for every AF $F \in \mathscr{A}$ we have: $\gamma_\sigma(\widetilde{F}^x, A(F)) = \sigma(F)$. Moreover, σ is *exactly x-verifiable* iff σ is x-verifiable and there is no verification class induced by $\mathfrak{r}^y$ with $\mathfrak{r}^y \prec \mathfrak{r}^x$ such that σ is y-verifiable.

We proceed with a list of criteria showing that any semantics mentioned in Definition 1 is verifiable by a verification class induced by a certain neighborhood function. In the following, we abbreviate the tuple $(\widetilde{F}^x, A(F))$ by $\widetilde{F}^x_A$.

$$\gamma_{na}(\widetilde{F}^{\varepsilon}_A) = \{S \mid S \in \widetilde{F}, S \textit{ is } \subseteq\textit{-maximal in } \widetilde{F}\};$$

$$\gamma_{stg}(\widetilde{F}^{+}_A) = \{S \mid (S,S^+) \in \widetilde{F}^+, S^+ \textit{ is } \subseteq\textit{-maximal in } \{C^+ \mid (C,C^+) \in \widetilde{F}^+\}\};$$

$$\gamma_{stb}(\widetilde{F}^{+}_A) = \{S \mid (S,S^+) \in \widetilde{F}^+, S^+ = A\};$$

$$\gamma_{ad}(\widetilde{F}^{\mp}_A) = \{S \mid (S,S^\mp) \in \widetilde{F}^\mp, S^\mp = \emptyset\};$$

$$\gamma_{pr}(\widetilde{F}^{\mp}_A) = \{S \mid S \in \gamma_{ad}(\widetilde{F}^{\mp}_A), S \textit{ is } \subseteq\textit{-maximal in } \gamma_{ad}(\widetilde{F}^{\mp}_A)\};$$

$$\gamma_{ss}(\widetilde{F}^{+\mp}_A) = \{S \mid S \in \gamma_{ad}(\widetilde{F}^{\mp}_A), S^+ \textit{ is } \subseteq\textit{-maximal in } \{C^+ \mid (C,C^+,C^\mp) \in \widetilde{F}^{+\mp}, C \in \gamma_{ad}(\widetilde{F}^{\mp}_A)\}\};$$

$$\gamma_{id}(\widetilde{F}^{\mp}_A) = \{S \mid S \textit{ is } \subseteq\textit{-maximal in } \{C \mid C \in \gamma_{ad}(\widetilde{F}^{\mp}_A), C \subseteq \bigcap \gamma_{pr}(\widetilde{F}^{\mp}_A)\}\};$$

$$\gamma_{eg}(\widetilde{F}^{+\mp}_A) = \{S \mid S \textit{ is } \subseteq\textit{-maximal in } \{C \mid C \in \gamma_{ad}(\widetilde{F}^{\mp}_A), C \subseteq \bigcap \gamma_{ss}(\widetilde{F}^{+\mp}_A)\}\};$$

$$\gamma_{sad}(\widetilde{F}_A^{-\pm}) = \{S \mid (S,S^-,S^\pm) \in \widetilde{F}^{-\pm}, \exists (S_0,S_0^-,S_0^\pm),\ldots,(S_n,S_n^-,S_n^\pm) \in \widetilde{F}^{-\pm}:$$
$$(\emptyset = S_0 \subset \cdots \subset S_n = S \wedge \forall i \in \{1,\ldots,n\} : S_i^- \subseteq S_{i-1}^\pm)\};$$
$$\gamma_{gr}(\widetilde{F}_A^{-\pm}) = \{S \mid S \in \gamma_{sad}(\widetilde{F}_A^{-\pm}), \forall (\bar{S},\bar{S}^-,\bar{S}^\pm) \in \widetilde{F}^{-\pm} : \bar{S} \supset S \Rightarrow (\bar{S}^- \setminus S^\pm) \neq \emptyset)\};$$
$$\gamma_{co}(\widetilde{F}_A^{+-}) = \{S \mid (S,S^+,S^-) \in \widetilde{F}^{+-}, (S^- \setminus S^+) = \emptyset, \forall (\bar{S},\bar{S}^+,\bar{S}^-) \in \widetilde{F}^{+-} : \bar{S} \supset S \Rightarrow (\bar{S}^- \setminus S^+) \neq \emptyset)\}.$$

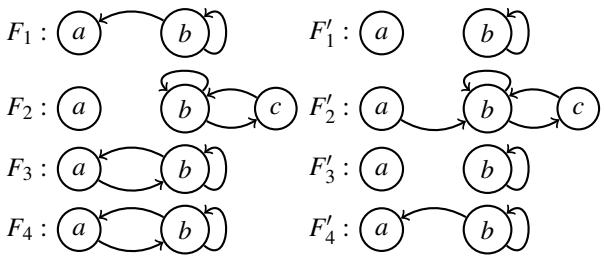
It is easy to see that the naive semantics is verifiable by the verification class induced by $\mathfrak{v}^\varepsilon$ since the naive extensions can be determined by the conflict-free sets. Stable and stage semantics, on the other hand, utilize the range of each conflict-free set in addition. Hence they are verifiable by the verification class induced by $\mathfrak{v}^+$. Now consider admissible sets. Recall that a conflict-free S set is admissible if and only if it attacks all attackers. This is captured exactly by the condition $S^\mp = \emptyset$, hence admissible sets are verifiable by the verification class induced by $\mathfrak{v}^\mp$. The same holds for preferred semantics, since we just have to determine the maximal conflict-free sets with $S^\mp = \emptyset$. Semi-stable semantics, however, needs the range of each conflict-free set in addition, see γ_{ss}, which makes it verifiable by the verification class induced by $\mathfrak{v}^{+\mp}$. Finally consider the criterion γ_{co}. The first two conditions for a set of arguments S stand for conflict-freeness and admissibility, respectively. Now assume the third condition does not hold, i.e., there exists a tuple $(\bar{S},\bar{S}^+,\bar{S}^-) \in \widetilde{F}^{+-}$ with $\bar{S} \supset S$ and $\bar{S}^- \setminus S^+ = \emptyset$. This means that every argument attacking $\bar{S}$ is attacked by S, i.e., $\bar{S}$ is defended by S. Hence S is not a complete extension, showing that $\gamma_{co}(\widetilde{F}_A^{+-}) = co(F)$ for each $F \in \mathscr{A}$. One can verify that all criteria from the list are adequate in the sense that they describe the extensions of the corresponding semantics.

The concepts of verifiability and being more informative behave correctly insofar as more informative neighborhood functions do not lead to a loss of verification capacity.

Proposition 7. *If a semantics σ is x-verifiable, then σ is verifiable by all verification classes induced by some $\mathfrak{v}^y$ with $\mathfrak{v}^y \succeq \mathfrak{v}^x$.*

In order to prove unverifiability of a semantics σ w.r.t. a class induced by a certain $\mathfrak{v}^x$ it suffices to present two AFs F and G such that $\sigma(F) \neq \sigma(G)$ but, $\widetilde{F}^x = \widetilde{G}^x$ and $A(F) = A(G)$. Then the verification class induced by $\mathfrak{v}^x$ does not provide enough information to verify σ. In the following we will use this strategy to show exact verifiability. Consider a semantics σ which is verifiable by a class induced by $\mathfrak{v}^x$. If σ is unverifiable by all verifiability classes induced by $\mathfrak{v}^y$ with $\mathfrak{v}^y \prec \mathfrak{v}^x$ we have that σ is exactly verifiable by $\mathfrak{v}^x$. The following examples study this issue for the semantics under consideration.

Example 4. The complete semantics is $+-$-verifiable as seen before. The following AFs show that it is even exactly verifiable by that class.



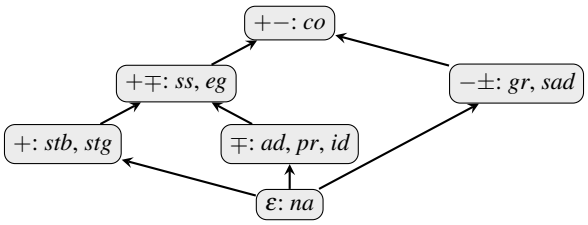
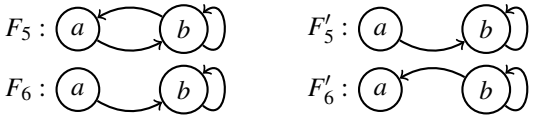


Figure 2. Semantics and their exact verification classes.



First consider the AFs F_1 and F'_1, and observe that $\widetilde{F_1}^{+\pm} = \{(\emptyset,\emptyset,\emptyset),(\{a\},\emptyset,\emptyset)\} = \widetilde{F'_1}^{+\pm}$. On the other hand F_1 and F'_1 differ in their complete extensions since $co(F_1) = \{\emptyset\}$ but $co(F'_1) = \{\{a\}\}$. Therefore complete semantics is unverifiable by the verification class induced by $\mathfrak{v}^{+\pm}$. Likewise, this can be shown for the classes induced by $\mathfrak{v}^{-\mp}$, $\mathfrak{v}^{\pm\mp}$, $\mathfrak{v}^{-\pm}$, $\mathfrak{v}^{+\mp}$, and $\mathfrak{v}^{\cap\cup}$, respectively:

- $\widetilde{F_2}^{-\mp} == \widetilde{F'_2}^{-\mp}$, but $co(F_2) = \{\{a\},\{a,c\}\} \neq \{\{a,c\}\} = co(F'_2)$.
- $\widetilde{F_3}^{\pm\mp} = \widetilde{F'_3}^{\pm\mp}$, but $co(F_3) = \{\emptyset,\{a\}\} \neq \{\{a\}\} = co(F'_3)$.
- $\widetilde{F_4}^{-\pm} = \widetilde{F'_4}^{-\pm}$, but $co(F_4) = \{\emptyset,\{a\}\} \neq \{\emptyset\} = co(F'_4)$.
- $\widetilde{F_5}^{+\mp} = \widetilde{F'_5}^{+\mp}$, but $co(F_5) = \{\emptyset,\{a\}\} \neq \{\{a\}\} = co(F'_5)$.
- $\widetilde{F_6}^{\cap\cup} = \widetilde{F'_6}^{\cap\cup}$, but $co(F_6) = \{\{a\}\} \neq \{\emptyset\} = co(F'_6)$.

Hence complete semantics is exactly verifiable by the verification class induced by $\mathfrak{v}^{+-}$.

Examples showing exact verifiability of the other semantics can be found in [19]. Figure 2 shows the resulting relation between the semantics under consideration with respect to their exact verification classes.

Theorem 8. *Every semantics which is rational is exactly verifiable by a verification class induced by one of the neighborhood functions presented in Figure 1.*

Proof. First of all note that by Lemma 6, $\mathfrak{v}^{\varepsilon}$ is the least informative neighborhood function and for every other neighborhood function $\mathfrak{v}^{x}$ it holds that $\mathfrak{v}^{\varepsilon} \preceq \mathfrak{v}^{x}$. Therefore, if a semantics is verifiable by the verification class induced by any $\mathfrak{v}^{x}$ then it is exactly verifiable by a verification class induced by some $\mathfrak{v}^{y}$ with $\mathfrak{v}^{\varepsilon} \preceq \mathfrak{v}^{y} \preceq \mathfrak{v}^{x}$. Moreover, if a semantics is exactly verifiable by a class, then it is by definition also verifiable by this class. Hence it remains to show that every semantics which is rational is verifiable by a verification class presented in Figure 1.

We show the contrapositive, i.e., if a semantics is not verifiable by a verification class induced by one of the neighborhood functions presented in Figure 1 then it is not rational. Assume a semantics σ is not verifiable by one of the verification classes. This means σ is not verifiable by the verification class induced by $\mathfrak{v}^{+-}$. Hence there exist two AFs F and G such that $\widetilde{F}^{+-} = \widetilde{G}^{+-}$ and $A(F) = A(G)$, but $\sigma(F) \neq \sigma(G)$. For every argument a which is not self-attacking, a tuple $(\{a\},\{a\}^{+},\{a\}^{-})$ is contained in $\widetilde{F}^{+-}$

(and in $\widetilde{G}^{+-}$). Hence F and G have the same not-self-attacking arguments and, moreover these arguments have the same ingoing and outgoing attacks in F and G. This, together with $A(F) = A(G)$ implies that $F^l = G^l$ (see Definition 2) holds. But since $\sigma(F) \neq \sigma(G)$ we get that σ is not rational, which was to show. □

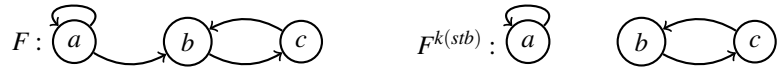
Note that the criterion giving evidence for verifiability of a semantics by a certain class has access to the set of arguments of a given AF. In fact, only the criterion for stable semantics makes use of that – it can be omitted for the other semantics.

5. Intermediate Semantics

A type of semantics which has aroused quite some interest in the literature (see e.g. [30] and [31]) are intermediate semantics, i.e. semantics which yield results lying between two existing semantics. The introduction of σ-τ-intermediate semantics can be motivated by deleting *undesired* (or add *desired*) τ-extensions while guaranteeing all reasonable positions w.r.t. σ. In other words, σ-τ-intermediate semantics can be seen as sceptical or credulous acceptance shifts within the range of σ and τ.

A natural question is whether we can make any statements about compatible kernels of intermediate semantics. In particular, if semantics σ and τ are compatible with some kernel k, is then every σ-τ-intermediate semantics k-compatible. The following example answers this question negatively.

Example 5. Recall from Theorem 1 that both stable and stage semantics are compatible with $k(stb)$, i.e. $F \equiv_E^{stb} G \Leftrightarrow F \equiv_E^{stg} G \Leftrightarrow F^{k(stb)} = G^{k(stb)}$. Now we define the following *stb*-*stg*-intermediate semantics, say *stagle* semantics: Given an AF $F = (A,R)$, $S \in sta(F)$ iff $S \in cf(F)$, $S_F^+ \cup S_F^- = A$ and for every $T \in cf(F)$ we have $S_F^+ \not\subset T_F^+$. Obviously, it holds that $stb \subseteq sta \subseteq stg$ and $stb \neq sta$ as well as $sta \neq stg$, as witnessed by the AF F:



It is easy to verify that $stb(F) = \emptyset \subset sta(F) = \{\{b\}\} \subset stg(F) = \{\{b\},\{c\}\}$. We proceed by showing that stagle semantics is not compatible with $k(stb)$. To this end consider $F^{k(stb)}$. Now, $sta(F^{k(stb)}) = \{\{b\},\{c\}\}$ witnesses $F \not\equiv^{sta} F^{k(stb)}$ and therefore, $F \not\equiv_E^{sta} F^{k(stb)}$. Since $F^{k(stb)} = (F^{k(stb)})^{k(stb)}$ we are done, i.e. stagle semantics is indeed not compatible with the stable kernel.

It is the main result of this section that compatibility of intermediate semantics w.r.t. a certain kernel can be guaranteed if verifiability w.r.t. a certain class is presumed. The provided characterization theorems generalize former results presented in [12]. Moreover, due to the abstract character of the theorems the results are applicable to semantics which may be defined in the future.

Before turning to the characterization theorems we state some implications of verifiability. In particular, under the assumption that σ is verifiable by a certain class, equality of certain kernels implies expansion equivalence w.r.t. σ.

Proposition 9. *For a semantics σ it holds that*

- *if σ is $+$-verifiable then $F^{k(stb)} = G^{k(stb)} \Rightarrow F \equiv_E^{\sigma} G$.*

- *if* σ *is* $+\mp$*-verifiable then* $F^{k(ad)} = G^{k(ad)} \Rightarrow F \equiv_E^\sigma G$.
- *if* σ *is* $+-$*-verifiable then* $F^{k(co)} = G^{k(co)} \Rightarrow F \equiv_E^\sigma G$.
- *if* σ *is* $-\pm$*-verifiable then* $F^{k(gr)} = G^{k(gr)} \Rightarrow F \equiv_E^\sigma G$.
- *if* σ *is* ε*-verifiable then* $F^{k(na)} = G^{k(na)} \Rightarrow F \equiv_E^\sigma G$.

We proceed with general characterization theorems. The first one states that *stb-stg*-intermediate semantics are compatible with stable kernel if $+$-verifiability is given. Consequently, stagle semantics as defined in Example 5 can not be $+$-verifiable.

Theorem 10. *Given a semantics* σ *which is* $+$*-verifiable and stb-stg-intermediate, it holds that* $F^{k(stb)} = G^{k(stb)} \Leftrightarrow F \equiv_E^\sigma G$.

Proof. ($\Rightarrow$) Follows directly from Proposition 9. ($\Leftarrow$) We show the contrapositive, i.e. $F^{k(stb)} \neq G^{k(stb)} \Rightarrow F \not\equiv_E^\sigma G$. Assuming $F^{k(stb)} \neq G^{k(stb)}$ implies $F \not\equiv_E^{stg} G$, i.e. there exists an AF H such that $stg(F \cup H) \neq stg(G \cup H)$ and therefore, $stb(F \cup H) \neq stb(G \cup H)$. Let $B = A(F) \cup A(G) \cup A(H)$ and $H' = (B \cup \{a\}, \{(a,b),(b,a) \mid b \in B\})$. It is easy to see that $stb(F \cup H') = stb(F \cup H) \cup \{\{a\}\}$ and $stb(G \cup H') = stb(G \cup H) \cup \{\{a\}\}$. Since now both $stb(F \cup H') \neq \emptyset$ and $stb(G \cup H') \neq \emptyset$ it holds that $stb(F \cup H') = stg(F \cup H')$ and $stb(G \cup H') = stg(G \cup H')$. Hence $\sigma(F \cup H') \neq \sigma(F \cup H')$, showing that $F \not\equiv_E^{stb} G$. □

The following theorems can be shown in a similar manner.

Theorem 11. *Given a semantics* σ *which is* $-\pm$*-verifiable and gr-sad-intermediate, it holds that* $F^{k(gr)} = G^{k(gr)} \Leftrightarrow F \equiv_E^\sigma G$.

Theorem 12. *Given a semantics* σ *which is* $+\mp$*-verifiable and* ρ*-ad-intermediate for any* $\rho \in \{ss, id, eg\}$*, it holds that* $F^{k(ad)} = G^{k(ad)} \Leftrightarrow F \equiv_E^\sigma G$.

Recall that complete semantics is a *ss-ad*-intermediate semantics. Furthermore, it is not characterizable by the admissible kernel as already observed in [12]. Consequently, it is not $+\mp$-verifiable (as we have shown in Example 4 with considerable effort).

6. Conclusions

In this work we have contributed to the analysis and comparison of abstract argumentation semantics. The main idea of our approach is to provide a novel categorization in terms of the amount of information required for testing whether a set of arguments is an extension of a certain semantics. The resulting notion of verification classes allows us to categorize any new semantics (given it is "rational") with respect to the information needed and compare it to other semantics. Thus our work is in the tradition of the principle-based evaluation due to Baroni and Giacomin [7] and paves the way for a more general view on semantics, their common features, and their inherent differences.

Using our notion of verifiability, we were able to show kernel-compatibility for certain intermediate semantics. Concerning concrete semantics, our results yield the following observation: While preferred, semi-stable, ideal and eager semantics coincide w.r.t. strong equivalence, verifiability of these semantics differs. In fact, preferred and ideal semantics manage to be verifiable with strictly less information.

For future work we envisage an extension of the notion of verifiability classes in order to categorize semantics not captured by the approach followed in this paper, such as *cf2* [23].

References

[1] Loui, R.P.: Defeat among arguments: a system of defeasible inference. Computational Intelligence **14** (1987) 100–106

[2] Pollock, J.L.: Defeasible reasoning. Cognitive Science **11** (1987) 481–518

[3] Prakken, H., Vreeswijk, G.: Logics for defeasible argumentation. In: Handbook of Philosophical Logic. Dordrecht (2002) 219–318

[4] Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., Toni, F.: Special issue: Tutorials on structured argumentation. Argument and Computation **5** (2014) 1–117

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

[6] Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowledge Eng. Review **26** (2011) 365–410

[7] Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. Artif. Intell. **171** (2007) 675–700

[8] Jakobovits, H., Vermeir, D.: Robust semantics for argumentation frameworks. JLC **9** (1999) 215–261

[9] Arieli, O.: Conflict-tolerant semantics for argumentation frameworks. In: Proc. JELIA. (2012) 28–40

[10] Grossi, D., Modgil, S.: On the graded acceptability of arguments. In: Proc. IJCAI. (2015) 868–874

[11] Verheij, B.: Two approaches to dialectical argumentation: admissible sets and argumentation stages. In: Proc. NAIC. (1996) 357–368

[12] Oikarinen, E., Woltran, S.: Characterizing strong equivalence for argumentation frameworks. Artif. Intell. **175** (2011) 1985–2009

[13] Baumann, R.: Characterizing equivalence notions for labelling-based semantics. In: Proc. KR. (2016) 22–32

[14] Maher, M.J.: Eqivalences of logic programs. In: Proc. ICLP. (1986) 410–424

[15] Lifschitz, V., Pearce, D., Valverde, A.: Strongly equivalent logic programs. ACM Trans. Comput. Log. **2** (2001) 526–541

[16] Turner, H.: Strong equivalence for causal theories. In: Proc. LPNMR. (2004) 289–301

[17] Truszczynski, M.: Strong and uniform equivalence of nonmonotonic theories - an algebraic approach. Ann. Math. Artif. Intell. **48** (2006) 245–265

[18] Caminada, M.: Strong admissibility revisited. In: Proc. COMMA. (2014) 197–208

[19] Baumann, R., Linsbichler, T., Woltran, S.: Verifiability of argumentation semantics. In: Proc. NMR. (2016) Available at `http://arxiv.org/abs/1603.09502`.

[20] Caminada, M., Carnielli, W.A., Dunne, P.E.: Semi-stable semantics. JLC **22** (2012) 1207–1254

[21] Dung, P.M., Mancarella, P., Toni, F.: Computing ideal sceptical argumentation. Artif. Intell. **171** (2007) 642–674

[22] Caminada, M.: Comparing two unique extension semantics for formal argumentation: Ideal and eager. In: Proc. BNAIC. (2007) 81–87

[23] Baroni, P., Giacomin, M., Guida, G.: SCC-Recursiveness: A general schema for argumentation semantics. Artif. Intell. **168** (2005) 162–210

[24] Gaggl, S.A., Woltran, S.: The cf2 argumentation semantics revisited. JLC **23** (2013) 925–949

[25] Baumann, R., Brewka, G.: Analyzing the equivalence zoo in abstract argumentation. In: Proc. CLIMA. (2013) 18–33

[26] Baumann, R., Brewka, G.: The equivalence zoo for Dung-style semantics. JLC (2015)

[27] Baumann, R., Woltran, S.: The role of self-attacking arguments in characterizations of equivalence notions. JLC: Special Issue on Loops in Argumentation (2014)

[28] Baroni, P., Dunne, P.E., Giacomin, M.: On the resolution-based family of abstract argumentation semantics and its grounded instance. Artif. Intell. **175** (2011) 791–813

[29] Dvořák, W., Linsbichler, T., Oikarinen, E., Woltran, S.: Resolution-based grounded semantics revisited. In: Proc. COMMA. (2014) 269–280

[30] Baroni, P., Giacomin, M.: Comparing argumentation semantics with respect to skepticism. In: Proc. ECSQARU. (2007) 210–221

[31] Nieves, J.C., Osorio, M., Zepeda, C.: A schema for generating relevant logic programming semantics and its applications in argumentation theory. Fundam. Inform. **106** (2011) 295–319