Revisiting the Foundations of Abstract Argumentation – Semantics Based on Weak Admissibility and Weak Defense

Ringo Baumann
Department of Computer Science
Leipzig University
Germany

Gerhard Brewka
Department of Computer Science
Leipzig University
Germany

Markus Ulbricht
Department of Computer Science
Leipzig University
Germany

Abstract

In his seminal 1995 paper, Dung paved the way for abstract argumentation, a by now major research area in knowledge representation. He pointed out that there is a problematic issue with self-defeating arguments underlying all traditional semantics. A self-defeat occurs if an argument attacks itself either directly or indirectly via an odd attack loop, unless the loop is broken up by some argument attacking the loop from outside. Motivated by the fact that such arguments represent self-contradictory or paradoxical arguments, he asked for reasonable semantics which overcome the problem that such arguments may indeed invalidate any argument they attack. This paper tackles this problem from scratch. More precisely, instead of continuing to use previous concepts defined by Dung we provide new foundations for abstract argumentation, so-called weak admissibility and weak defense. After showing that these key concepts are compatible as in the classical case we introduce new versions of the classical Dung-style semantics including complete, preferred and grounded semantics. We provide a rigorous study of these new concepts including interrelationships as well as the relations to their Dung-style counterparts. The newly introduced semantics overcome the issue with self-defeating arguments, and they semantically insensitive to syntactic deletions of self-attacking arguments, a special case of self-defeat.

1 Introduction

Computational models of argumentation have received a lot of attention in AI for more than two decades now. The seminal paper boosting this interest was Dung's paper on abstract argumentation frameworks (AFs) (Dung 1995). Dung's work is based on the observation that argument evaluation, more precisely the selection of reasonable sets of arguments constituting a coherent world view, can be done without taking into account the internal structure of arguments. Arguments can be treated as abstract, atomic entities. All that is needed is information about the attack relation among such arguments. Consequently, Dung's AFs are just directed graphs. The nodes in this graph represent abstract arguments, the edges describe attacks among arguments.

Dung defines various semantics for AFs. Here, a semantics assigns to each AF a collection of extensions, that is jointly acceptable subsets of the arguments. The different semantics reflect different ways of resolving conflicts among arguments. In the meantime, several additional semantics have been defined, see (Baroni, Caminada, and Giacomin 2018) for an overview.

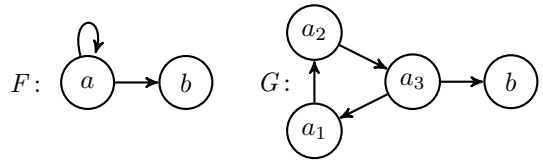
At the end of his seminal paper (Dung 1995) Dung points out an issue he left open in his approach. He writes (p. 351):

> "An interesting topic of research is the problem of self-defeating arguments as illustrated in the following example. Consider the argumentation framework $(\{A, B\}, \{(A, A), (A, B)\})$. The only preferred extension here is empty though one can argue that since A defeats itself, B should be acceptable."

The goal of this paper is to address exactly this problem. We will do so by introducing modified versions of some of the central notions in Dung's approach. Besides conflict-freeness, one of these notions is admissibility. According to Dung, a conflict-free set of arguments is admissible if it defends itself against all attackers. In other words, a set of arguments S is admissible whenever each argument attacking S from the outside is itself attacked by some element of S. An admissible set can thus not contain any argument which is attacked without at the same time being defended.

We will show in this paper that alternative and, as we claim, reasonable modifications of this and other fundamental notions of abstract argumentation exist. Our motivation is as follows. It is indeed important that a set of arguments defends itself. However, does it have to defend itself against all attackers? Isn't it sufficient to counterattack those arguments which have the slightest chance of being accepted? Let us illustrate this using the following simple examples.

Example 1.1 (Self-Defeat and Acceptance)**.** Consider the AF F. According to Dung's definition $\{b\}$ is not admissible[1] because it does not defend itself against a.

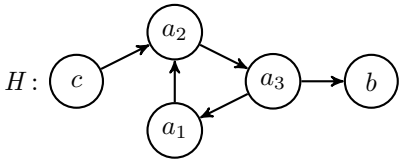


But what kind of attacker is a? Does b really have to defend itself against an argument that "kills" itself, so-to-speak? Why

[1] We refer to the background section for precise definitions.

does b have to counterattack if a itself does this job anyway? a is like a zombie, it is there but can do no harm.

Consider now G. Self-attack is only one special case of self-defeat. Why should b have to defend itself against a_3 if a_3 is among those arguments that defeat themselves indirectly, by attacking the only argument (here a_1) that could help them to defend themselves against an attacker (here a_2)? Our new notion of admissibility should classify $\{b\}$ as acceptable in both frameworks. Note that, unlike for self-attack, indirect self-defeat through an odd loop of length ≥ 3 can be broken up by additional arguments as depicted in AF H. Here an additional new argument c attacking a_2 helps a_3 to become credible again. Consequently, the attack on b should not be ignored leading to the non-acceptance of $\{b\}$.



We are aware that there exist semantics which, for instance, yield $\{b\}$ as an extension for framework F, namely naive semantics (Baroni, Caminada, and Giacomin 2018). However, naive semantics also yields extension $\{b\}$ for the self-attack free framework $F' = (\{a, b\}, \{(a, b)\})$ which we find unsatisfactory. What we are aiming for is a semantics that limits the effects of self-attacking arguments, but is as close as possible to Dung's semantics whenever there is no self-attack. Naive semantics clearly fails this requirement.

The idea underlying this paper is to weaken admissibility, requiring counterattack only against proper, that is, not directly or indirectly self-defeating arguments. To the best of our knowledge, such weaker notions of admissibility have never been investigated. Interestingly, Baroni and Giacomin (Baroni and Giacomin 2007) proposed a stronger variant called *strong admissibility* which requires that defense is ultimately rooted in the empty set. Consequently, only subsets of the grounded extensions can be strongly admissible. We will briefly discuss this concept in Section 6.

The structure of the paper is as follows. Section 2 provides the necessary background material. The main contributions of the paper are briefly summarised in the next three items:

- Introducing *weak admissibility* based on a recursive definition and showing that the motivating examples are handled as desired. It is formally proven that weak admissibility indeed extends classical admissibility and moreover, it is insensitive regarding the deletion of self-loops. A generalized version of Dungs fundamental lemma is shown. (Section 3)
- In the spirit of weak admissibility a new notion of defense, so-called *weak defense* is presented. A series of formal results show that the new notions are compatible as appreciated in the classical case, e.g. any conflict-free set is weakly admissible if and only if it weakly defends itself. (Section 4)
- Defining *weak complete semantics* based on the newly introduced notions as done in the classical theory. Provide an in-depth study of their properties including: subset relations to each other as well as to their Dung-style counterparts, universal definedness, uniqueness and the role of self-attacking arguments. (Section 5)

Finally, we discuss and compare our results and present pointers for future work. Due to limited space we present the proofs of the theorems only.

2 Background

We fix a non-finite background set $\mathcal{U}$. An argumentation framework (AF) (Dung 1995) is a directed graph $F = (A, R)$ where $A \subseteq \mathcal{U}$ represents a set of arguments and $R \subseteq A \times A$ models *attacks* between them. In this paper we consider finite AFs only (cf. (Baumann and Spanring 2015; 2017) for a consideration of infinite AFs). For $a, b \in A$, if $(a, b) \in R$ we say that a *attacks* b as well as a *attacks* (the set) E given that $b \in E \subseteq A$. Moreover, E is conflict-free in F (for short, $(E \in cf(F)$ iff for no $a, b \in E$, $(a, b) \in R$. We say a set E *classically defends* (or simply, *c-defends*) an argument a if any attacker of a is attacked by some argument of E. In this paper we consider so-called *admissible*, *complete*, *preferred* and *grounded* semantics (abbr. ad, co, pr, gr). Each semantics returns a set of sets of acceptable positions which are defined as follows (cf. (Baroni, Caminada, and Giacomin 2018) for a recent overview).

Definition 2.1. Let $F = (A, R)$ be an AF and $E \in cf(A)$.

1. $E \in ad(F)$ iff E c-defends all its elements,
2. $E \in co(F)$ iff $E \in ad(F)$ and for any x c-defended by E we have, $x \in E$,
3. $E \in pr(F)$ iff E is $\subseteq$-maximal in $co(F)$ and
4. $E \in gr(F)$ iff E is $\subseteq$-minimal in $co(F)$.

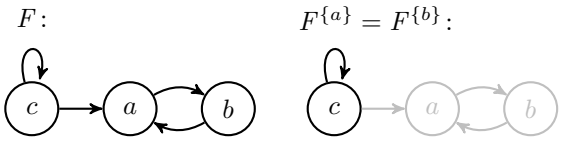
3 Weak Admissibility

Dung's definition of admissibility requires that an extension E defends itself against *all* attackers. As we already mentioned, the core idea behind our revised version of admissibility is to weaken this condition and only require defense against reasonable arguments. To formalize this approach, we use the so-called *E-reduct* of an AF which only contains the arguments which are neither in E nor attacked by E.

Definition 3.1. Let $F = (A, R)$ be an AF and let $E \subseteq A$. The E-reduct of F is the AF $F^E = (E^*, R \cap (E^* \times E^*))$ where $E^* = A \setminus (E \cup \{a \in A \mid E \text{ attacks } a\})$.

Intuitively, the E-reduct of F is the subframework of F containing those arguments whose status still needs to be decided, assuming the arguments in E are accepted. Moreover, as a matter of fact, a set E is admissible in F iff $E \in cf(F)$ and E^* contains no attacker of an argument $a \in E$. Consider therefore the following illustrating example.

Example 3.2 (Reduct and Admissibility)**.** The reduct $F^{\{b\}}$ contains only argument c which does not attack b. Hence $\{b\}$ is admissible. However, c occurs in $F^{\{a\}}$ as well. Since c attacks a, $\{a\}$ does not defend itself and is thus not admissible.



Although the $\{a\}$-reduct from the previous example contains an attacker of a, it is not necessarily an unreasonable argument. In fact, the AF F from above contains a pair a and b of conflicting arguments which is disturbed by some dummy attacking a. To formalize that c is negligible in this situation we consider the reduct $F^{\{a\}}$. Since c attacks itself, it cannot occur in any reasonable extension of $F^{\{a\}}$. However, in general, what arguments are the negligible ones in an E-reduct F^E at hand? To solve this issue, we utilize the following recursive definition.

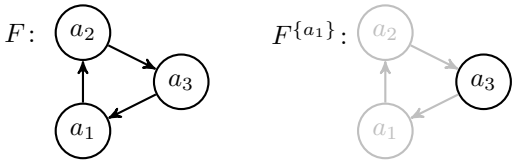
Definition 3.3. Let $F = (A, R)$ be an AF. $E \subseteq A$ is called *weakly admissible* (or *w-admissible*) in F ($E \in ad^w(F)$) iff

1. $E \in cf(F)$, and
2. for any attacker y of E we have $y \notin \bigcup ad^w(F^E)$.

The major difference between the standard definition of admissibility and our new one is that arguments do not have to c-defend themselves against *all* attackers: attackers which do not appear in any w-admissible set of the reduct can be neglected. In order to familiarize the reader with weak admissibility, let us consider some examples to see our definition at work.

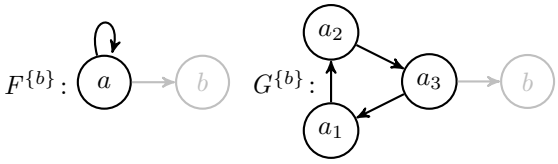
Example 3.4 (Example 3.2 ctd.)**.** In the previous example we had $F^{\{a\}} = (\{c\}, \{(c, c)\})$ where $\{c\}$ is not conflict-free and hence not w-admissible. Since the empty set is conflict-free and not attacked by any argument we deduce $ad^w(F^{\{a\}}) = \{\emptyset\}$. Now let us verify that $\{a\} \in ad^w(F)$. The following observations justify the claim. First, $\{a\} \in cf(F)$ and moreover, since $\bigcup ad^w(F^{\{a\}}) = \emptyset$ no attacker of $\{a\}$ (argument c only) may be an element of $\bigcup ad^w(F^{\{a\}})$. In anticipation of Proposition 3.7 below we mention that $\{b\}$ and $\emptyset$ are w-admissible in F too due to their admissibility.

Example 3.5 (Odd loop)**.** In order to see that no non-empty conflict-free set is w-admissible in F consider $E = \{a_1\}$.



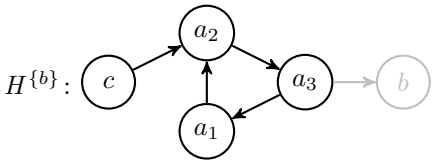
The E-reduct of F contains a_3 which attacks a_1. Since a_3 is unattacked in F^E, it is w-admissible in F^E which means in turn that $\{a_1\}$ is *not* w-admissible. Due to symmetry, no singleton is w-admissible in F justifying $ad^w(F) = \{\emptyset\}$.

Example 3.6 (Example 1.1 ctd.)**.** Let us check that the new notion of admissibility indeed handle the motivating examples as desired.



The conflict-free set $\{b\}$ is attacked by a or a_3, respectively. Neither of the attackers is contained in a w-admissible set of the associated reducts $F^{\{b\}}$ or $G^{\{b\}}$ as already seen in Examples 3.4 and 3.5. This means, we have $\{b\} \in ad^w(F)$ and $\{b\} \in ad^w(G)$ as desired.

Recall the AF H. We argued that in this case, $\{b\}$ should *not* be w-admissible since the attacker a_3 is now relevant.



Indeed, since $\{c, a_3\}$ is admissible in $H^{\{b\}}$ we deduce its w-admissibility (see Proposition 3.7 below). Hence $\{b\}$ is attacked by $a_3 \in \bigcup ad^w(H^{\{b\}})$ and thus $\{b\} \notin ad^w(G)$ as requested.

Let us proceed with some basic considerations. Since the empty set does not possess any attacker it is w-admissible in each AF. Moreover, the restriction to finite AFs guarantees the well-definedness of the recursive procedure. In other words, for any candidate set E the recursion will stop in finitely many steps. Finally, w-admissibility indeed generalizes the classical notion of admissibility as defined by Dung. This means, we do not lose any admissible set if sticking to w-admissibility as stated below.

Proposition 3.7. *For any AF F, $ad(F) \subseteq ad^w(F)$.*

Definition 2.1 presents preferred extensions as $\subseteq$-maximal complete ones. In case of classical semantics one may alternatively define an preferred extension as $\subseteq$-maximal admissible one. This second variant will now be used for the weak variant of preferred semantics.

Definition 3.8. Let $F = (A, R)$ be an AF, $E \subseteq A$ is called *weakly preferred* (or simply, *w-preferred*) in F ($E \in pr^w(F)$) iff E is $\subseteq$-maximal in $ad^w(F)$.

To illustrate the new preferred version and to compare it with its Dung-style counterpart consider again the motivating example G.

Example 3.9 (Example 3.6 ctd.)**.** We already discussed that the singleton $\{b\}$ is w-admissible without being admissible. This means, $ad(G) = \{\emptyset\} \neq \{\emptyset, \{b\}\} = ad^w(G)$. Hence, we obtain different preferred versions too, namely $pr(G) = \{\emptyset\} \neq \{\{b\}\} = pr^w(G)$.

Let us now have a look at self-attacking arguments. As extensions are required to be conflict-free by almost all reasonable AF semantics, self-attacking arguments never have a chance to be accepted. However, for no admissible-based semantics it is possible to ignore these arguments since they may influence the acceptance status of others. Our new notion of w-admissibility does support the syntactical deletion

of self-attacking arguments, i.e. the status of an argument is independent of any self-attacker.

For an AF $F = (A, R)$ we use the shorthand F° for the restriction $F_{|A^\circ}$ where $A^\circ = \{a \in A \mid (a, a) \notin R\}$. The following main theorem proves the semantical insensitivity of w-admissibility regarding self-attacker.

Theorem 3.10. *Given an AF $F = (A, R)$ and a semantics $\sigma \in \{ad, pr\}$. We have $\sigma^w(F) = \sigma^w(F^\circ)$.*

Proof. It suffices to prove the claim for $\sigma = ad$ since $ad^w(F) = ad^w(F^\circ)$ implies $pr^w(F) = pr^w(F^\circ)$.

We will prove our claim by induction over the size $|A|$ of an AF $F = (A, R)$. Remember that we consider finite AFs only. The base case is clear.

(inductive step) Let $n \in \mathbb{N}$ and assume the claim holds for any AF with at most n arguments. Consider an AF $F = (A, R)$ with $|A| = n + 1$. We will show $E \in ad^w(F)$ iff $E \in ad^w(F^\circ)$ for any $E \subseteq A$. In case of $E = \emptyset$ we immediately obtain $E \in ad^w(F)$ and $E \in ad^w(F^\circ)$. From now on we will assume $E \neq \emptyset$.

($\subseteq$) Let $E \in ad^w(F)$. Then E is conflict-free and no attacker y of E occurs in a w-admissible extension of F^E. Observe that since E is conflict-free, E in particular does not contain self-attacking arguments, so $(F^E)^\circ = (F^\circ)^E$. Moreover, since E is non-empty, F^E contains at most n arguments. By our induction hypothesis, $ad^w\left(F^E\right) = ad^w\left((F^E)^\circ\right)$. Consequently, $ad^w\left(F^E\right) = ad^w\left((F^\circ)^E\right)$. Hence E is conflict-free and no attacker y of E occurs in a w-admissible extension of $(F^\circ)^E$, that is, $E \in ad^w(F^\circ)$.

($\supseteq$) We show the contrapositive. Let $E \notin ad^w(F)$. Assume E is conflict-free. Then E must be attacked by an argument y which occurs in a w-admissible extension of F^E, say E_y. Due to the induction hypothesis, $ad^w(F^E) = ad^w\left((F^\circ)^E\right)$ implying $E_y \in ad^w((F^\circ)^E)$ as well. Consequently, E is also attacked by a w-admissible extension of $(F^\circ)^E$, i.e. $E \notin ad^w(F^\circ)$. Finally, if E is not conflict-free, then obviously $E \notin ad^w(F^\circ)$. □

The last central result we are going to present before introducing weak defense is a generalized version of Dung's fundamental lemma, which states that if E is admissible and c-defends a, then $E \cup \{a\}$ is admissible as well. It turns out that this results extends to w-admissibility. We state it here explicitly and use it later on in Section 5 to provide some technical foundations.

Theorem 3.11. *If $F = (A, R)$ is an AF, $E \in ad^w(F)$ and E c-defends a, then $E \cup \{a\} \in ad^w(F)$.*

Proof. The first observation we are going to make is that $E \cup \{a\}$ is conflict-free: Since $\{a\}$ occurs in the reduct F^E, E does not attack a. So assume $\{a\}$ attacks E. Since $\{a\}$ is unattacked in F^E, $\{a\} \in ad^w(F^E)$. Hence if $\{a\}$ attacks E, then E is not w-admissible, a contradiction.

Consider now $E \cup \{a\}$. For the sake of contradiction assume $E \cup \{a\}$ is not w-admissible. Then there is an extension $E' \in ad^w(F^{E \cup \{a\}})$ attacking $E \cup \{a\}$.

So E' is conflict-free and has no attacker in $\bigcup ad^w\left(\left(F^{E \cup \{a\}}\right)^{E'}\right)$. Observe that $E' \cup \{a\}$ is conflict-free: E' cannot attack $\{a\}$ since the latter is c-defended by E and thus unattacked in F^E. It therefore must also be unattacked in $F^{E \cup \{a\}}$. Moreover, $\{a\}$ cannot attack E' because otherwise, E' could not be an extension of $F^{E \cup \{a\}}$.

Since the reduct can be computed in any order when considering conflict-free sets, we have

$$\left(F^{E \cup \{a\}}\right)^{E'} = \left(\left(F^E\right)^{\{a\}}\right)^{E'} = \left(F^E\right)^{E' \cup \{a\}}.$$

Thus E' is conflict-free and has no attacker in $\bigcup ad^w\left(\left(F^E\right)^{E' \cup \{a\}}\right)$. Since $E' \cup \{a\}$ is conflict-free and $\{a\}$ is unattacked in F^E this means also that $E' \cup \{a\}$ is conflict-free and unattacked in $\bigcup ad^w\left(\left(F^E\right)^{E' \cup \{a\}}\right)$. So $E' \cup \{a\} \in ad^w(F^E)$.

Recall that E' attacks $E \cup \{a\}$. Since $E' \cup \{a\}$ is conflict-free, E' attacks E. Hence, $E' \cup \{a\}$ attacks E. We conclude that there is a w-admissible extension of F^E attacking E, namely $E' \cup \{a\}$. Thus, E is not w-admissible, a contradiction. □

4 Weak Defense

The classical notion of defense requires that an argument a is defended against *every* attacker. As defense is an important concept in abstract argumentation, we seek for a notion of *weak defense*, which we desire to have the following properties:

- D1. Weak defense should generalize classical defense: if a is c-defended by E, then a is w-defended by E as well.
- D2. Weak defense should be compatible with weak admissibility: if E is conflict-free, then E is w-admissible if and only if it w-defends itself.
- D3. The induced versions of complete semantics should accept at least as much as their Dung-style counterparts.

The following definition captures the idea that in order to defend arguments X, a set E does not necessarily have to counterattack each attacker y of X (as required in the classical case). Whenever y is not "serious" enough - which here means y fails to appear in at least one w-admissible set of the reduct - then y can be disregarded if two additional conditions are satisfied: y is not allowed to appear in E, and X must be a reasonably trustworthy set, meaning that the arguments of X are elements of one single w-admissible set. These two additional conditions are required to satisfy the desiderata.

Definition 4.1. Let $F = (A, R)$ be an AF. Given two sets $E, X \subseteq A$. We say E *weakly defends* (or *w-defends*) X iff for any attacker y of X we have,

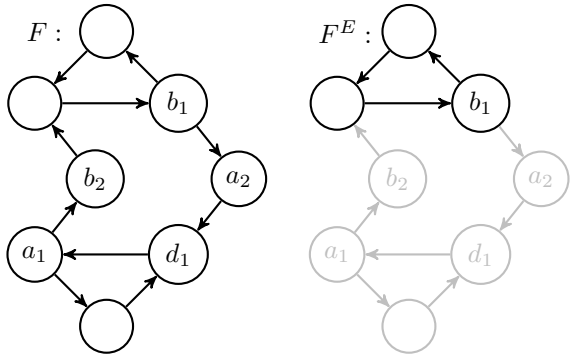
1. E attacks y, or (c-defense)
2. $y \notin \bigcup ad^w(F^E)$, $y \notin E$ and $X \subseteq X' \in ad^w(F)$.

Note that desideratum D1 is fulfilled due to the first option of the definition above. Consider the motivating example G.

Example 4.2 (Example 3.9 ctd.)**.** Let us verify that $\emptyset$ w-defends $\{b\}$. Since b is attacked by a_3, $\{b\}$ is not c-defended by $\emptyset$. However, the following three conditions are met: i) trivially, $a_3 \notin \emptyset$, ii) since $G^\emptyset = G$ we deduce that $a_3 \notin \bigcup ad^w\left(G^\emptyset\right) = \bigcup ad^w(G) = \{b\}$, and iii) $\{b\} \subseteq \{b\} \in ad^w(G)$. Thus, the second option of Definition 4.1 is satisfied. Moreover, $\{b\}$ w-defends itself.

Consider now a more involved example which does not possess any non-trivial admissible or complete extension.

Example 4.3. The reader may verify that $ad^w(F) = \{\emptyset, \{a_1\}, \{a_2\}, \{a_1, a_2\}, \{b_1\}, \{b_2\}, \{b_1, b_2\}\}$. For example, take $E = \{a_1, a_2\}$. Since $ad^w(F^E) = \{\emptyset\}$ we infer that the remaining attacker b_1 of E is not contained in any w-admissible set of F^E. Thus, $E \in ad^w(F)$.



Let us verify now that $\emptyset$ w-defends $\{a_1\}$ and $\{b_1\}$, but not $\{a_1, b_1\}$. Since $\emptyset$ does not c-defend $\{a_1\}$, we require the second item of Definition 4.1: i) clearly, the attacker d_1 of a_1 is not contained in $\emptyset$, ii) since $F = F^\emptyset$ we deduce $d_1 \notin \bigcup ad^w(F^\emptyset) = \{a_1, a_2, b_1, b_2\}$, and iii) $\{a_1\}$ is a subset of $\{a_1, a_2\}$. Thus, $\emptyset$ w-defends $\{a_1\}$. Due to symmetry, $\{b_1\}$ is w-defended as well. However, since a_1 and b_1 do not occur in the same w-admissible extension of F, $\{a_1, b_1\}$ is not w-defended by $\emptyset$. Furthermore, facing all mentioned results it can be easily seen that E w-defends itself. The same applies to $\{b_1, b_2\}$ for symmetry reasons.

Let us prove some basic relations involving c-defense, w-defense as well as w-admissibility.

Proposition 4.4. *Given an AF $F = (A, R)$ and two sets $E, X \subseteq A$. We have*

1. *If E c-defends X, then E w-defends. X*
   (w-defends weakens c-defends)
2. *If E w-defends X and $E \in cf(F)$, then $X \in cf(F)$.*
   (conflict-free transfer)
3. *If $E \in ad^w(F)$, then E w-defends E.*
   (a w-adm set w-defends itself)
4. *If $E \in cf(F)$ and E w-defends E, then $E \in ad^w(F)$.*
   (a conflict-free set which w-defends itself is w-adm)

We mention that desideratum D2 is an immediate consequence of the last two items of the proposition above.

5 Weak Complete Semantics

Many argumentation semantics can be defined using Dung-style admissibility and classical defense. For instance, complete semantics requires that a set E has to be admissible and any argument x being c-defended by E has to be contained in E (cf. Definition 2.1). Since any admissible set c-defends itself, complete semantics can be equivalently expressed in terms of set defense as follows: First, E has to be admissible and secondly, any superset X of E being c-defended by E has to be contained in E as shown below.

Proposition 5.1. *Let $F = (A, R)$ be an AF. If E is admissible, then E is complete iff for any set $E \subseteq X$ defended by E, we have $X \subseteq E$ as well.*

We now introduce the weak variant of complete semantics by using the equivalent characterization shown above.

Definition 5.2. Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is called *weakly complete* (or just, *w-complete*) in F ($E \in co^w(F)$) iff $E \in ad^w(F)$ and for any X, s.t. $E \subseteq X$ and X w-defended by E, we have $X \subseteq E$.

We obtain the following non-trivial connection between w-complete and w-preferred semantics which holds for the classical theory too, namely subset maximizing w-admissibility is nothing else than subset maximizing w-completeness. This coincidence can be seen as a further indication that w-admissibility and w-defense are compatible even in a strong formal sense.

Theorem 5.3. *Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is w-preferred iff it is $\subseteq$-maximal in $co^w(F)$.*

Proof. ($\subseteq$) Striving for a contraction we assume $E \in pr^w(F)$, i.e. E is $\subseteq$-maximal in $ad^w(F)$ and E is not $\subseteq$-maximal in $co^w(F)$. The latter condition may hold for two reasons:

1. $E \notin co^w(F)$ or
2. $E \in co^w(F)$ but not $\subseteq$-maximal in $co^w(F)$.

In the following we show that both options are impossible.

1. $E \notin co^w(F)$. Since $E \in ad^w(F)$ is given we deduce there is an X, s.t. $E \subseteq X$ and X w-defended by E. Due to Statement 3 in Proposition 4.4 we have $E \subsetneq X$. Assume there is at least one attacker y of X, s.t. the second condition of Definition 4.1 holds. Hence, there is an X', s.t. $X \subseteq X' \in ad^w(F)$ contradicting the $\subseteq$-maximality of E in $ad^w(F)$ since $E \subsetneq X \subseteq X'$. Consequently, any attacker y of X is counter-attacked by E (first condition of Definition 4.1). This means, X is c-defended by E which implies that X is c-defended by X too (monotonicity of c-defense). Moreover, $X \in cf(F)$ is given (Statement 2, Proposition 4.4) which finally leads to $X \in ad^w(F)$ (Statement 4, Proposition 4.4) contradicting the $\subseteq$-maximality of E in $ad^w(F)$.
2. $E \in co^w(F)$ but not $\subseteq$-maximal in $co^w(F)$. Hence, there is an other $E' \in co^w(F)$, s.t. $E \subsetneq E'$. By definition of w-complete extensions we have $E' \in ad^w(F)$ which contradicts the $\subseteq$-maximality of E in $ad^w(F)$.

($\supseteq$) Assume now that E is $\subseteq$-maximal in $co^w(F)$ without being $\subseteq$-maximal in $ad^w(F)$. Due to w-completeness of E we deduce $E \in ad^w(F)$. Since E is assumed to be not $\subseteq$-maximal in $ad^w(F)$ we deduce (due to finite assumption) the existence of a proper superset E' of E being $\subseteq$-maximal

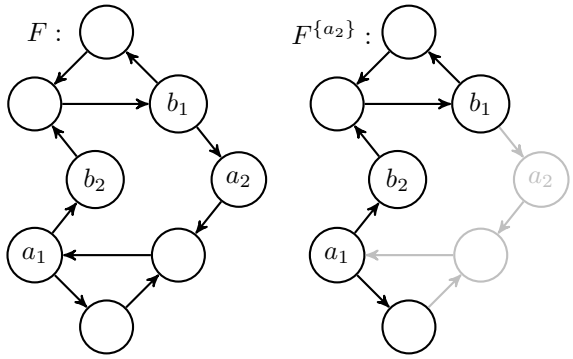
in $ad^w(F)$, i.e. $E' \in pr^w(F)$. Now applying the already shown direction ($\subseteq$) we deduce that E' is $\subseteq$-maximal in $co^w(F)$ in contradiction to the assumed $\subseteq$-maximality of E in $co^w(F)$. □

Now let us introduce the remaining prominent actor in the field of complete semantics, namely the weak version of grounded semantics.

Definition 5.4. Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is called *weakly grounded* (or *w-grounded*) in F ($E \in gr^w(F)$) iff E is $\subseteq$-minimal in $co^w(F)$.

The following example illustrates w-complete extensions. Note that by Definition 5.2 w-complete extensions can be found among w-admissible ones only.

Example 5.5 (Example 4.3 ctd.)**.** We have already seen that $\emptyset$ w-defends $\{a_1\}$. Hence, $\emptyset \notin co^w(F)$ since $\{a_1\} \not\subseteq \emptyset$. We show that $\{a_2\}$ w-defends $\{a_1, a_2\}$ and is thus not w-complete. Since $\{a_2\}$ classically defends a_1, there is only one relevant attacker, namely b_1. We check the requirements of Definition 4.1, item 2: i) $b_1 \notin \{a_2\}$ is clear, ii) $ad^w\left(F^{\{a_2\}}\right) = \{a_1\}$ as the reader may straightforwardly verify, so $b_1 \notin \bigcup ad^w\left(F^{\{a_2\}}\right)$, and iii) $\{a_1, a_2\}$ itself is w-admissible as we saw in Example 4.3.



Due to symmetry, $\{b_2\} \notin co^w(F)$ either. One may verify that $co^w(F) = \{\{a_1\}, \{a_1, a_2\}, \{b_1\}, \{b_1, b_2\}\}$. Hence, $gr^w(F) = \{\{a_1\}, \{b_1\}\}$.

The reader may have surprisingly noticed that the example considered above possesses two w-grounded extensions which is impossible for the uniquely defined classical notion of grounded semantics. Let us approach this feature carefully by starting to examine the fundamental intersemantic relations between the newly introduced weak versions as well as their classical counterparts.

The following proposition shows that well-known subset relation of Dung-style semantics transfer to their weak versions. For the sake of completeness we add one of the most important argumentation semantics, namely the *stable semantics* (*stb*) (Dung 1995) as well as a further variation of classical admissibility captured by so-called *strong admissible sets* (ad^s) firstly introduced in (Baroni, Caminada, and Giacomin 2018).

Proposition 5.6. *Let σ and τ be semantics. Then $\sigma(F) \subseteq \tau(F)$ for any F iff there is a directed path from σ to τ in the following graph:*

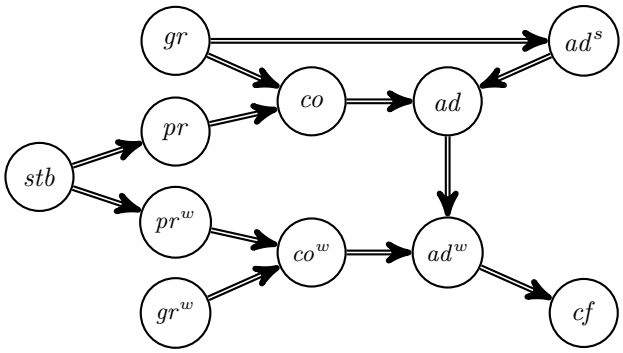


Figure 1: Subset Relations

We proceed with stating the universal definedness of any newly introduced semantics.

Proposition 5.7. *Given an AF $F = (A, R)$ and a semantics $\sigma \in \{ad, pr, co, gr\}$. We have, $\sigma^w(F) \neq \emptyset$.*

According to Proposition 5.6 we already know that neither are w-complete (w-preferred) extension necessarily complete (preferred) ones nor vice versa. Nevertheless, both versions are not totally unrelated as shown below. More precisely, the following proposition shows that one can find weak extensions by augmenting a corresponding classical one.

Proposition 5.8. *Given an AF $F = (A, R)$ and a semantics $\sigma \in \{co, pr\}$. If $E \in \sigma(F)$, then there is an $E^w \in \sigma^w(F)$, s.t. $E \subseteq E^w$.*

For complete semantics we may even show the converse result, i.e. a weak complete extensions always contains a classical complete part. The proof of this result is based on the observation that w-complete extensions contain all their classically defended arguments. Since this result is interesting on its own, we state it explicitly here.

Proposition 5.9. *Let $F = (A, R)$ be an AF. If $E^w \in co^w(F)$ and E^w c-defends a, then $a \in E^w$.*

Proposition 5.10. *Let $F = (A, R)$ be an AF. If $E^w \in co^w(F)$, then there is an $E \in co(F)$, s.t. $E \subseteq E^w$.*

One may wonder whether the result above does hold for preferred semantics too. We mention that this is not the case (consider AF F depicted in Example 3.2).

Let us turn now to w-grounded semantics. In the classical case the grounded extension is uniquely determined and moreover, it meets the requirement of a $\subseteq$- least complete extension (although introduced as a $\subseteq$-minimal one only). As an analogy to Propositions 5.8 and 5.10 we obtain the following revealing result for w-grounded semantics. Any w-grounded extension contains at least the arguments which can be traced back (via c-defence) to undisputable arguments, i.e. arguments which are not questioned/attacked by any other argument. In other words, although w-grounded extensions are not uniquely determined they always contain a common core, namely the classical grounded extension G and moreover (due to universal definedness), G can always be extended to a w-grounded extension.

Proposition 5.11. *Given $F = (A, R)$ and $\{G\} = gr(F)$.*

1. *If* $G^w \in gr^w(F)$*, then* $G \subseteq G^w$ *and*
2. *there is a set* $G^w \in gr^w(F)$*, s.t.* $G \subseteq G^w$*.*

According to Proposition 5.8 and the above shown results we have finally shown that the weak versions of complete semantics do not reject any argument accepted by their Dung-style counterparts. This means, desideratum D3 is indeed fulfilled.

As a by-product we obtain that w-complete extension always contain the classical grounded part. This property might be interesting for finding w-complete extensions since grounded semantics can be computed in polynomial time (Dvorák and Dunne 2018).

Corollary 5.12. *For any* $F = (A, R)$*, semantics* $\sigma \in \{gr, pr, co\}$ *and* $gr(F) = \{G\}$*. If* $E^w \in \sigma^w(F)$*, then* $G \subseteq E^w$*.*

In Theorem 3.10 we observed that self-attacking arguments do not influence w-admissible extension of an AF. This result even extends to complete semantics. The key feature for showing this assertion is the following proposition stating that in the crucial cases w-defends is not influenced by self-attacking arguments.

Proposition 5.13. *Let* $F = (A, R)$ *be an AF. Given* $E, X \subseteq A$*, s.t.* $E \in cf(F)$ *and* $X \subseteq A^\circ$*. Then.* E *w-defends* X *in* F *iff* E *w-defends* X *in* F°*.*

Having Proposition 5.13 at hand we are now able to prove the main theorem stating the independence of self-attacking arguments regarding w-complete as well as w-grounded extensions.

Theorem 5.14. *Given an AF* $F = (A, R)$ *and a semantics* $\sigma \in \{co, gr\}$*. We have* $\sigma^w(F) = \sigma^w(F^\circ)$*.*

Proof. It suffices to prove the claim for $\sigma = co$ since $co^w(F) = co^w(F^\circ)$ implies $gr^w(F) = gr^w(F^\circ)$.
($\subseteq$) Let $E \in co^w(F)$. By definition, $E \in ad^w(F)$ and for any X, s.t. $E \subseteq X$ and X w-defended by E, we have $X \subseteq E$. First observe that $E \in ad^w(F^\circ)$ (Theorem 3.10). So let $X \subseteq E$. Assume X is not w-defended by E in F. We have to show that X is not w-defended in F°, either. Since E is conflict-free, this is guaranteed by Proposition 5.13 whenever X does not contain self-attacking arguments. Otherwise X cannot be defended in F° due to *conflict-free transfer* from Proposition 4.4.
($\supseteq$) We show the contrapositive, i.e. assume $E \notin co^w(F)$. We may assume $E \in ad^w(F)$, otherwise we apply Theorem 3.10 to obtain $E \notin co^w(F^\circ)$. Let $E \subseteq X$ be w-defended by E with $X \not\subseteq E$. We deduce that X is conflict-free due to *conflict-free transfer*. It thus does not contain any self-attacking arguments. By Proposition 5.13, E is also w-defended in F°. To summarize, X satisfies $E \subseteq X$, is w-defended by E in F°, but $X \not\subseteq E$. So $E \notin co^w(F^\circ)$, either concluding the proof. □

6 Summary and Related Work

Based on the motivation that the effect of self-defeating arguments should be limited, we developed in this paper new semantics for abstract argumentation frameworks which are based on weakened versions of admissibility and defense. We introduced refined versions of complete, preferred and grounded semantics and thoroughly studied their properties and interrelationships. The new semantics lead to more informative extensions and provide us with reasonable sets of arguments where classical Dung-style semantics fail, see for instance Example 4.3. Moreover, in contrast to their classical counterparts, they also allow for certain syntactic manipulations (deletion of self-attacking arguments) without affecting the meaning of the frameworks. This unique feature can be used as preprocessing step for any system engaged with the solution of central reasoning problems for abstract argumentation (Dvorák et al. 2019). Beside this specific use case we expect that our new semantics are widely applicable: it will be interesting to investigate whether approaches which directly or indirectly rest upon Dung semantics - like the treatment of *preferences* in (Amgoud and Cayrol 1998), the *assumption-based approach* to argumentation (Toni 2014), or the *ASPIC*$^+$ framework (Modgil and Prakken 2014), to name a few prominent examples - can benefit from our concepts.

There are numerous papers studying notions of self-defeat in argumentation. Such studies can be traced back at least to J. Pollock (Pollock 1987), one of the pioneers of computational models of argumentation. In his study of argument-based defeasible reasoning, which preceded Dung's seminal paper, he proposed a semantics similar to grounded semantics. This semantics considers self-defeat via self-attack, but not via odd loops of length ≥ 3. We will focus in this section on approaches which, explicitly or implicitly, modify some of the notions underlying Dung's work.

As mentioned in the Introduction, in (Baroni and Giacomin 2007) a stronger variant of admissibility called *strong admissibility* is proposed. Their purpose was to highlight some of the intrinsic properties of grounded semantics and they showed (among other things) that Dung's grounded extension is the $\subseteq$-greatest strongly admissible set. In further studies, strong admissibility was used to investigate the computational behaviour of *discussion-based proof procedures* (Caminada 2014) and its *strong equivalence* as well as *verifiability* was analyzed (Baumann, Linsbichler, and Woltran 2016). None of these papers discusses options to modify classical semantics based on a new notion of admissibility.

Another relevant line of research regarding basic requirements of argumentation semantics are *conflict-tolerant* approaches. As the name suggests, such semantics may return extensions which are not necessarily conflict-free (Arieli 2012). An interesting example is weighted argument systems as defined in (Dunne et al. 2011). Here each attack is assigned a numerical weight and conflicts within extensions are allowed as long as a certain predefined inconsistency budget is not exceeded. Although this also leads to extensions which are not admissible in Dung's original sense, it is substantially different from what we do. Rather than allowing conflicts to a certain extent, we limit the effects of a specific type of arguments but insist on conflict-freeness of extensions. The two approaches thus address complementary issues. They certainly could be combined in a straightforward manner. Another conflict-tolerant approach is so-called *graded semantics* (Grossi and Modgil 2015). This approach relies on the syntactic structure of AFs only. More precisely, in order to obtain

generalized versions of Dung's semantics it takes the number of attackers and defenders into account. Checking to which extent both approaches can be combined is on our agenda.

We also plan a systematic study of further properties of our semantics like the ones discussed in (Baroni, Caminada, and Giacomin 2018) and (van der Torre and Vesic 2018). A prominent example is SCC-recursiveness. Table 1 in the latter paper describes 15 different semantics for argumentation frameworks together with principles they satisfy. It turns out that all these semantics either satisfy admissibility (each extension is admissible) or naivety (each extension is a maximal conflict-free set). As an initial observation we would like to mention that our semantics (intentionally) do not belong to any of these two categories, which shows that they significantly differ from existing semantics.[2] Furthermore, there is the recently introduced research direction of inconsistency in abstract argumentation theory which deals with possible repairs of AFs that prevent one from drawing any plausible conclusion, i.e. no argument is accepted (Baumann and Ulbricht 2018; Ulbricht 2019). It will be interesting to see how weak admissibility-based semantics behave in this regard.

In addition, we plan to study computational aspects of the new semantics (complexity and algorithms) as well as the possibility of transferring the new concepts of admissibility and defense to further classical semantics. It will also be interesting to see whether the ideas underlying our approach can be fruitfully applied to other AI formalisms like for instance answer set programming (Brewka, Eiter, and Truszczynski 2011). We believe the concepts introduced here may turn out to be the basis for a completely new line of fundamental research in abstract argumentation.

Acknowledgements

We thank DFG (Deutsche Forschungsgemeinschaft) for funding this work (project number 406289255 and project BR 1817/7-2).

References

Amgoud, L., and Cayrol, C. 1998. On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, UAI*, 1–7.

Arieli, O. 2012. Conflict-tolerant semantics for argumentation frameworks. In *Proceedings of 13th European Conference in Logics in Artificial Intelligence, JELIA,*, 28–40.

Baroni, P., and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence* 171:675–700.

Baroni, P.; Caminada, M.; and Giacomin, M. 2018. Abstract argumentation frameworks and their semantics. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications.

Baumann, R., and Spanring, C. 2015. Infinite argumentation frameworks - on the existence and uniqueness of extensions. In *Essays Dedicated to Gerhard Brewka on the Occasion of His 60th Birthday*, volume 9060, 281–295. Springer.

Baumann, R., and Spanring, C. 2017. A study of unrestricted abstract argumentation frameworks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, 807–813.

Baumann, R., and Ulbricht, M. 2018. If nothing is accepted - repairing argumentation frameworks. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR*, 108–117.

Baumann, R.; Linsbichler, T.; and Woltran, S. 2016. Verifiability of argumentation semantics. In *Proceedings of the 6th International Conference of Computational Models of Argument, COMMA*, 83–94.

Brewka, G.; Eiter, T.; and Truszczynski, M. 2011. Answer set programming at a glance. *Communications of the ACM* 54(12):92–103.

Caminada, M. 2014. Strong admissibility revisited. In *Computational Models of Argument - Proceedings of COMMA 2014*, 197–208.

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(2):321–357.

Dunne, P. E.; Hunter, A.; McBurney, P.; Parsons, S.; and Wooldridge, M. J. 2011. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence* 175(2):457–486.

Dvorák, W., and Dunne, P. E. 2018. Computational problems in formal argumentation and their complexity. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications.

Dvorák, W.; Järvisalo, M.; Linsbichler, T.; Niskanen, A.; and Woltran, S. 2019. Preprocessing argumentation frameworks via replacement patterns. In *Logics in Artificial Intelligence - 16th European Conference, JELIA 2019*, 116–132.

Grossi, D., and Modgil, S. 2015. On the graded acceptability of arguments. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI*, 868–874.

Modgil, S., and Prakken, H. 2014. The *ASPIC*$^+$ framework for structured argumentation: a tutorial. *Argument & Computation* 5(1):31–62.

Pollock, J. L. 1987. Defeasible reasoning. *Cognitive Science* 11(4):481–518.

Toni, F. 2014. A tutorial on assumption-based argumentation. *Argument & Computation* 5(1):89–117.

Ulbricht, M. 2019. *Understanding Inconsistency – A Contribution to the Field of Non-monotonic Reasoning*. Ph.D. Dissertation, Leipzig University.

van der Torre, L., and Vesic, S. 2018. The principle-based approach to abstract argumentation semantics. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications.

[2] We thank L. van der Torre for pointing this out to us.