Weak Argumentation Semantics and Unsafe Odd Cycles: Results and a Conjecture

Sjur K Dyrkolbotn(✉)

Department of Civil Engineering, Western Norway University of Applied Sciences, Bergen, Norway
sdy@hvl.no

Abstract. Some semantics for argumentation, including the newly introduced weakly admissible semantics, allow us to ignore attacks from arguments that are perceived as problematic. A key intuition motivating such semantics is that arguments that indirectly attack themselves may be problematic in such a way that this is justified. In this paper, we formalise this intuition and provide a class of semantics that are weakly admissible, coincide with the stable semantics on a large class of argumentation frameworks that admit stable sets, and only ignore attacks from arguments on unsafe cycles of odd length. We also show that no member of our class of semantics coincide with the semantics that takes all $\subseteq$-maximal weakly admissible sets as extensions. However, we show that this semantics satisfies an even stronger property, if the following conjecture is true: if an argumentation framework has no non-empty weakly admissible sets, then every argument lies on an unsafe odd cycle.

1 Introduction

Abstract argumentation based on argumentation frameworks in the style of [8] 'has become a popular modelling paradigm in knowledge representation and artificial intelligence. Several different semantics for argumentation have been proposed in this tradition, catering to various intuitions, applications and modelling requirements. One key issue that arises concerns the semantic status of arguments that directly or indirectly attack themselves: when should such possibly problematic arguments be regarded as capable of defeating other arguments? The traditional semantics for argumentation arguably fail to provide satisfactory answers to this question, but in [4], the authors provide a new class of semantics that looks very promising on examples. It is explicitly motivated by the idea that we should be able to ignore attacks from self-defeating arguments. But what exactly does this mean? The authors provide an informal answer, writing that "self-defeat occurs if an argument attacks itself either directly or indirectly via

Thanks to the anonymous reviewers for pointing out some relevant references and making suggestions that greatly improved the presentation of the paper.

S. Gaggl et al. (Eds.): JELIA 2023, LNAI 14281, pp. 161–175, 2023.
https://doi.org/10.1007/978-3-031-43619-2_12

an odd attack loop, unless the loop is broken up by some argument attacking the loop from outside." In this paper, we propose a formal definition based on the same intuition. We clarify what we mean by an argument attacking itself directly or indirectly and note that loops/cycles may also be broken up from the "inside", when there are additional attacks between arguments on the cycle. We then investigate what semantics for argumentation are *justified* by the key intuition at work, when formalised as a requirement that a semantics may or may not satisfy. We show that while the most permissive semantics based on weak admissibility is not justified, it is possible to define, for any well-behaved admissible semantics, a corresponding weakly admissible semantics that is justified. We show that these semantics extend the stable semantics in a reasonable way, returning only stable sets as extensions for a large class of argumentation frameworks that have no problematic odd cycles. However, we also show that no semantics in the class we define is equivalent to the weakly preferred semantics, obtained by taking all $\subseteq$-maximal weakly admissible sets as extensions. Despite this negative result, we conjecture that the weakly preferred semantics is also justified. We show that if this is true, then the weakly preferred semantics satisfies an even stronger property, whereby for every extension, every argument not included or attacked by it lies on an unbroken odd cycle.

2 Background

The basic notion is that of an argumentation framework, which mathematically speaking is nothing but a directed graph, usually assumed to be finite.

Definition 1. *An* argumentation framework *(AF) is a directed graph* $AF = (A, R)$ *where* $R \subseteq A \times A$ *is referred to as an attack relation over a finite set of arguments* A*.*

For any $AF = (A, R)$ and $S \subseteq A$, the subframework of AF induced by S is $AF\downarrow_S = (S, R \cap (S \times S))$. Moreover, for any $a \in A$ we denote by $R(a) = \{b \in A \mid (a, b) \in R\}$ the set of arguments attacked by a and by $R^-(a) = \{b \in A \mid (b, a) \in R\}$ the set of arguments that attack a. We extend the notation to sets $S \subseteq A$, so that $R(S) = \bigcup_{a \in S} R(a)$ and $R^-(S) = \bigcup_{a \in S} R^-(a)$. If $S, Q \subseteq A$, we say that S attacks Q just in case $S \cap R^-(Q) \neq \emptyset$. For any $AF = (A, R)$ and $S \subseteq A$ we let $[S]_{AF} = S \cup R(S)$. We omit the subscript when it is clear from the context. Furthermore, we denote by AF^S the subframework of AF induced by $A \backslash [S]_{AF}$, called the *reduct* of AF by S. Given S, an odd cycle in AF^S can be regarded as an odd cycle that is not broken up from the outside by S.

An (attack) walk of length n in AF is a sequence of arguments $W_{a_0,a_n} = (a_0, a_1, \ldots, a_n)$ such that $a_i \in R(a_{i-1})$ for all $1 \leq i \leq n$. If $i \neq j \Rightarrow a_i \neq a_j$ for all $0 \leq i \leq n$, the walk is an (attack) path. If $a_0 = a_n$ and $i \neq j \Rightarrow a_i \neq a_j$ for all $1 \leq i \leq n$, the walk is an (attack) cycle of length n. When n is even, the cycle is even, and when n is odd, the cycle is odd. Notice that (a, a) is an odd cycle consisting of a single argument attacking itself.

If $W_{a_0,a_n} = (a_0, a_1, \ldots, a_n)$ is a walk and $W_{a_n,a_m} = (a_n, a_{n+1}, \ldots, a_m)$ is a walk of length $m-n$, then $W_{a_0,a_n} + W_{a_n,a_m} = (a_0, a_1, \ldots, a_n, a_{n+1}, \ldots, a_m)$ is a walk of length $n+(m-n) = m$. Beware that if $P_{a,b}$ is a path ending at b and $P_{b,c}$ is a path beginning at b, then $P_{a,b} + P_{b,c}$ is a walk, but not necessarily a path, since $P_{a,b}$ and $P_{b,c}$ might intersect internally. Given a set $B \subseteq A$ and a walk $W = (a_0, a_1, \ldots, a_n)$ we say that W is B-alternating if $a_i \in B \Leftrightarrow a_{i+1} \notin B$ for all $0 \leq i < n$. That is, W is B-alternating just in case every other argument on W is in B. So, for instance, if $B = \{a, c\}$, then the paths (a, b, c) and (a, b, c, d) are B-alternating, while (a, b, d) is not. Notice that a B-alternating path from B to B always has even length.

If $P = (a_0, a_1, \ldots, a_n)$ is a path, then $P_{a_i,a_{i+j}} = (a_i, a_{i+1}, \ldots, a_{i+j})$ is a subpath of P for all $0 \leq i < n$ and $j \leq n - i$. Moreover, an attack $(a_i, a_j) \in R$ with $0 \leq i, j \leq n$ and $j \neq i+1$ is called a *chord* on P. We say that a chord (a_i, a_j) on $P = (a_0, a_1, \ldots, a_n)$ *breaks* P if i and j are both even. If $C = (a_0, \ldots, a_n = a_0)$ is an odd cycle and (a_i, a_j) is a chord that breaks $P = (a_0, \ldots, a_{n-1})$, then we say that C is *safe* at a_0. It is *unsafe* at a_0 otherwise.

An argumentation semantics ς assigns, to any $AF = (A, R)$, a set of subsets of the arguments, also called ς-*extensions*, $\varsigma : A \to 2^{2^A}$. A semantics is typically defined in terms of requirements on the sets of arguments it returns as extensions. Many different semantics have been defined using various combinations of different requirements. Hence, different requirements and how they may be understood, motivated and justified in different contexts, as well as how they relate to one another mathematically, has become an important research topic in argumentation theory. Following [1], requirements are also often used to classify and compare different argumentation semantics. In this context, underlying mathematical requirements are lifted from sets of arguments to semantics and referred to as semantic *principles*. A principle corresponds to a whole class of different semantics, consisting of all semantics that only return extensions that satisfy the underlying requirement.

The most widely endorsed argumentation principle is that a semantics for argumentation should only return *conflict free* sets of arguments as extensions. Given $AF = (A, R)$ and $S \subseteq A$, we say that S is conflict free if $(S \times S) \cap R = \emptyset$. That is, S is conflict free if there are no attacks between any two elements of S. Lifting the requirement to define a class of semantics, we say that a semantics ς for argumentation is conflict free – meaning that it satisfies the principle of conflict-freeness – if for all $AF = (A, R)$ and all $S \in \varsigma(AF)$, S is conflict free.

Many semantics for argumentation, including the original ones presented in [8], satisfy another principle, namely that they only return extensions that defend themselves. Formally, a set $S \subseteq A$ defends itself just in case it attacks everything that attacks it, $R^-(S) \subseteq R(S)$. Lifting this notion from sets to semantics, we say that a semantics ς is *defensive* if for all $AF = (A, R)$ and $S \in \varsigma(AF)$, S defends itself. A set that is conflict free and defends itself is *admissible*. Lifting this notion to semantics ς, if ς is conflict free and defensive, it is an admissible semantics. Hence, notice that with this terminology there are several admissible

semantics, not just the most permissive one that always returns all admissible sets as extensions (often called the admissible semantics in the literature).

Notice that if S is admissible, $a \in S$, and $P_{a,b}$ is broken by a chord, then since S is conflict free, $P_{a,b}$ is not S-alternating. Intuitively, an S-alternating path starting at S is an unbroken path of semantic dependencies that arise when we regard S as an extension, so such paths can have no chords that break them. This also explains why an odd cycle can be broken from the inside and why we say that $C = (a_0, a_1 \ldots, a_n)$ is safe at a when the path $P = (a_0, \ldots, a_{n-1})$ is broken: attempting to include a in some extension S could not produce a sequence of semantic dependencies along C that would end up defeating a. Hence, C does not indicate that a is actually self-defeating, regardless of S and whether or not C is broken by it from the outside.

The most permissive admissible semantics is not very reasonable, most notably because it always returns $\emptyset$ as a possible extension of any AF. However, the notion of admissibility is still fundamental, since it forms the basis for a range of other semantics, often arrived at by stipulating additional principles.

Semantics that are conflict free but not defensive, allowing us to sometimes ignore attacks, are *weaker* than admissible semantics. Such semantics are not new. In fact, a whole class of semantics weaker than the admissible semantics has been introduced based on computing (maximal) conflict free sets [2]. These semantics generally do not quite match the desiderata explored in this paper, however, as they typically allow us to ignore attacks also from arguments that do not indirectly attack themselves. This is shown with examples and discussed at length in [4], so we do not go into detail. Instead, we will focus on a new class of semantics which is explicitly motivated by the intuition we formalise in this paper. The key notion is that of weak admissibility, defined as follows.

Definition 2. *Given any $AF = (A, R)$, a set of arguments $S \subseteq A$ is weakly admissible when it is conflict free and there is no set $Q \subseteq A \backslash [S]$ that attacks S in AF and is weakly admissible in AF^S.*

We will also lift this notion from sets to semantics and regard it as a principle, by saying that a semantics ς is *weakly admissible* if for all $AF = (A, R)$, if $S \in \varsigma(AF)$, then S is weakly admissible. So a semantics is said to be weakly admissible if it only returns weakly admissible sets as extensions. Notice that if S defends itself, then there is no set attacking S in AF that is also present in AF^S. Hence, every admissible S is also weakly admissible. At the level of semantics, adopting our terminology, it follows that all admissible semantics are weakly admissible.

3 Perfect Extensions of the Stable Semantics

If a semantics ς seems too permissive, for instance because $\emptyset$ is always a ς-extension, one may impose additional principles to arrive at a more restricted semantics. The most straightforward approach is to restrict ς by taking as extensions only those $S \in \varsigma(AF)$ that are $\subseteq$-maximal. This scheme yields what we

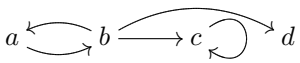
call ς-preferred semantics, which is referred to simply as the preferred semantics when ς is the most permissive admissible semantics. When ς is the most permissive weakly admissible semantics, then the ς-preferred semantics is referred to as the weakly preferred semantics.

A stronger principle than $\subseteq$-maximality is to demand that S must attack every argument not in S. Formally, for any $AF = (A, R)$ we say that $S \subseteq A$ is *dominating* if $S \cup R(S) = A$. As before, a semantics ς such that for all $AF = (A, R)$, any $S \in \varsigma(AF)$ is dominating, is called a dominating semantics.

A set that is conflict free and dominating is called a *stable set* in the literature, and the semantics that returns all stable sets in AF as extensions is called the stable semantics. A stable set unambiguously determines the semantic status of every argument in the AF, partitioning them into those arguments we accept and those we reject, which are all attacked by some argument we accept. Unfortunately, stable sets may not exist, as illustrated by the AF consisting of a single self-attacking argument.

The fact that stable sets may not exist is a key motivation for introducing weaker semantics that tolerate partial semantic verdicts. Hence, it may seem natural to require that weaker semantics are *conservative extensions* of the stable semantics, in the sense that whenever stable sets exist, the weaker semantics only returns stable sets as extensions. One way of ensuring this is to define some class of sets that include all stable sets and then choose as extensions all sets from the class that have minimal reducts. Then the stable sets are the only extensions whenever they exist, because their reducts are always empty. Following this approach starting with admissible sets yields a trivially equivalent formulation of the so-called semi-stable semantics [5], which returns as extensions all admissible sets S for which $[S] = S \cup R(S)$ is $\subseteq$-maximal.

It is not clear, however, that conservative extensions of the stable semantics yield reasonable results. Consider, for instance, the following AF:



The only stable set is $\{b\}$, which is also the only semi-stable set, having an empty reduct. It is also a preferred set, of course, but it is not the only one. The set $\{a, d\}$ is also preferred, being $\subseteq$-maximal among the admissible sets. Is it reasonable to say that d (and a) must be rejected because a prevents b from defeating the self-defeating c? This is far from obvious and will depend on what the AF is intended to model (or how it is instantiated by less abstract arguments).

Clearly, the preferred semantics is not a conservative extension, so how does it relate to the stable semantics? This can be answered formally using a concept from graph theory [11] that appears to have been largely neglected by the argumentation community. Adapting the terminology to the present setting, we say that $AF = (A, R)$ is *perfectly stable* if for all $S \subseteq A$, the subframework induced by S, $AF\downarrow_S$, has a stable set. Then we define a new argumentation principle as follows.

Definition 3. *A semantics* ς *is a* perfect extension *of the stable semantics if for all* $AF = (A, R)$ *such that* AF *is perfectly stable, we have*

$$\forall S \in \varsigma(AF) : R(S) = A \backslash S$$

That is, a semantics is a perfect extension of the stable semantics if it only returns stable sets as extensions for perfectly stable AFs.[1] Several sufficient conditions for the existence of stable sets in AF are known, most notably that a (finite) AF has a stable set if it has no odd cycles (a result that originally appeared in [13], published in 1953). This and most other sufficient conditions for the existence of stable sets ensure that AF is perfectly stable, so they also ensure that any perfect extension of the stable semantics only returns stable sets as extensions on AF. Moreover, it follows from [11] that a minimal AF that is not perfectly stable satisfies a property that is particularly interesting in the present context: all arguments a on AF lie on odd cycles. This suggests that it should be possible to define a semantics that satisfies the desiderata explored in the present paper, although how exactly to do it remains a non-trivial open question.

Before we move on, we note that the definition of a perfect extension is well matched to the concept of modularity, explored in [4] and defined as follows.

Definition 4. *An argumentation semantics* ς *is modular if for every* $AF = (A, R)$ *and* $S \subseteq A$*, if* $S \in \varsigma(AF)$ *and* $S' \in \varsigma(AF^S)$*, then* $S \cup S' \in \varsigma(AF)$*.*

Admissible sets are modular, so if S is preferred, then AF^S has no admissible set [4]. In the terminology from [4], it has a "meaningless reduct". From this it follows that the preferred semantics is a perfect extension of the stable semantics, as we now prove.

Theorem 5. *Given any* $AF = (A, R)$*, if* AF *is perfectly stable, then every preferred set in* AF *is a stable set. Hence, the preferred semantics is a perfect extension of the stable semantics.*

Proof. Assume AF is perfectly stable and let S be a preferred set in AF. We must show that S is stable. Assume towards contradiction that it is not. Since S is conflict free, it follows that S is not dominating, so that AF^S is non-empty. Since S is perfectly stable, AF^S must then have a non-empty stable set S'. Since stable sets are admissible, S' is admissible in AF^S. Hence, by the fact that preferred sets are admissible and admissible sets are modular, it follows that $S \cup S'$ is admissible in AF, contradicting $\subseteq$-maximality of S.

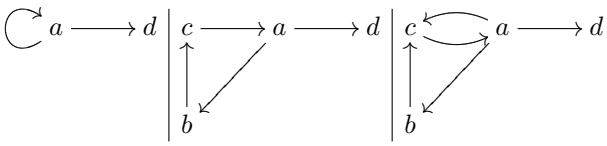
[1] The notion of a perfect extension could be made more general by explicitly taking the principle that is perfectly extended as a parameter, defining an AF to be perfectly X if all induced subdigraphs of the AF has an extension satisfying X. Then we could say that a semantics perfectly extends X, or that it satisfies the perfect extension principle for X, whenever it satisfies X for all AFs that are perfectly X. However, we only consider perfect extensions of the stable semantics in this paper, so we prefer to avoid the additional notation and terminology that the generalisation entails.

Since weakly admissible sets are modular, as shown in [4], it follows from essentially the same argument used to establish Theorem 5 that the weakly preferred semantics is *also* a perfect extension of the stable semantics. We regard this as a desirable property for an argumentation semantics to satisfy, and record it as a theorem.

Theorem 6. *The weakly preferred semantics is a perfect extension of the stable semantics.*

4 A Formal Justification for Ignoring Attacks

The informal motivation for weak admissibility presented in [3] is to provide a class of semantics that allow us to sometimes ignore attacks from arguments that attack themselves, directly or indirectly, on cycles that are not broken from the outside. On simple examples, it is verified that weakly admissible sets do indeed allow us to do this, as in the two AFs on the left below:



In the two leftmost AFs, it is easy to see that $\{d\}$ is weakly admissible, since there is no weakly admissible set from $AF^{\{d\}}$ that attacks it. Hence, we can disregard the attack from the self-defeating a, which lies on an unbroken odd cycle. By telling us to look for weakly admissible sets in the reduct AF^S, the definition of weak admissibility also seems to capture roughly the idea that we only ignore attacks from odd cycles that are not broken from the "outside" by S. However, in the rightmost AF above, $\{d\}$ is not weakly admissible, even though the odd cycle in $AF^{\{d\}}$ is not broken from the outside. It is broken from the inside, however, since it is safe at a, so we might no longer feel entitled to ignore the attack on d. Indeed, $\{a\}$ is the only non-empty weakly admissible set of this AF, despite being on an odd cycle that is not broken from the outside. It is also stable, so by Theorem 13, the weakly preferred semantics still behaves reasonably on examples like these, but not in a way that is fully explained by the informal explanation provided in [3].

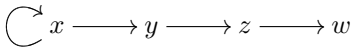
Examples like these illustrate that it is hardly intended that we should *always* ignore attacks from odd cycles that are unbroken by the outside. On the other hand, it seems quite reasonable to interpret the authors of [3] as intending that we should *only* disregard attacks from such arguments. This, at any rate, would be a very interesting descriptive property for a weak semantics to satisfy, as it would indicate that we have weakened the notion of admissibility only as much as our informal intuition warrants us to do. Based on this idea, we propose the following two semantic principles, corresponding to two possible justifications for ignoring attacks.

Definition 7. *An argumentation semantics ς is said to be*

- justified *(by unsafe odd cycles) if it is conflict free and for all $AF = (A, R)$ and all $S \in \varsigma(AF)$, every argument $a \in A \backslash [S]$ that attacks S in AF lies on an odd cycle in AF^S that is unsafe at a.*
- strongly justified *(by unsafe odd cycles) if it is conflict free and for all $AF = (A, R)$ and all $S \in \varsigma(AF)$, every argument $a \in A \backslash [S]$ lies on an odd cycle in AF^S that is unsafe at a.*

Notice that a semantics is justified whenever it is strongly justified, while the converse does not hold in general. In particular, notice that if ς is strongly justified and AF has only the empty extension under ς, then every argument in AF lies on an odd cycle. So if ς is strongly justified, then whenever there is some argument not indirectly attacking itself, there is a non-empty extension. Also notice that if ς is strongly justified and a does not lie on an odd cycle in AF, then if $S \in \varsigma(AF)$ and $R^-(a) \cap S = \emptyset$, we must have $a \in S$, since otherwise $a \in A \backslash [S]$ without being on an odd cycle in AF^S. In view of this, being strongly justified is a stronger property than being strongly complete outside odd cycles, as defined in [6], whereby $a \in S$ is only required when there is also no argument in $R^-(a)$ that lies on an odd cycle.

While being (strongly) justified is a strong property that seems desirable, it is not clear whether weakly admissible and (strongly) justified semantics exist. In the next section, we provide a class of weakly admissible semantics that are justified, before showing that they are not strongly justified. First we note that the most permissive weakly admissible semantics, taking all weakly admissible sets as extensions, is not justified. This can be shown, for instance, by the following example:



In this AF, it is clear that $\{w\}$ is weakly admissible. This follows since there is no weakly admissible set containing z, which in turn follows from the fact that there *is* a weakly admissible set containing y, since x – its only attacker – attacks itself. The fact that weakly admissible sets are not necessarily justified should not come as a great surprise. A similar phenomenon is observed for admissibility, whereby we quickly conclude that the most permissive admissible semantics is not very reasonable, despite admissibility being a fundamental notion that forms a basis for other semantics. The situation is similar, we believe, with respect to weak admissibility.

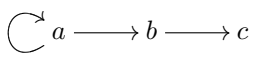
5 A Class of Justified Semantics Based on Admissible Sets

To arrive at a class of justified semantics, we will start by defining a class of semantics that is *more* permissive than weakly admissible semantics. Then we will define a restrictive class of weakly admissible semantics that ignore attacks

only from arguments that are not acceptable under the corresponding permissive semantics. We will then prove that the more restrictive class of semantics is weakly admissible and that taking its $\subseteq$-maximal sets yields a justified semantics that is also a perfect extension of the stable semantics. The first definition, giving rise to the permissive class of semantics, is the following.

Definition 8. *Given any $AF = (A, R)$ and a semantics ς, we say that $S \subseteq A$ is ς-plausible if it is conflict free and there is no $Q \in \varsigma(AF^S)$ that attacks S.*

Notice that if $S \subseteq A$ is admissible, then S is also ς-plausible. This is trivial, since S is conflict free and is not attacked by *any* argument from AF^S, since it defends itself. Also notice that if we take ς to be the most permissive weakly admissible semantics, consisting of all weakly admissible sets, then S is ς-plausible if, and only if, it is weakly admissible. So for the weakly admissible semantics, there is no difference between being a ς-extension and being ς-plausible. This is not true for semantics based on admissible sets. In fact, since weakly admissible sets are admissible, it follows that if ς is admissible, then S is ς-plausible whenever S is weakly admissible. So ς-plausibility behaves similarly to the weakly admissible semantics on simple motivating examples. It also seems to have independent interest as a natural dual of ς. However, ς-plausibility for admissible ς is too permissive to be justified. This is illustrated by the fact that both $\{b\}$ and $\{c\}$ is ς-plausible in the following AF, whenever ς is admissible:



This example also demonstrates that ς-plausible sets are not modular, so they will fail to provide justified and perfect extensions of the stable semantics. However, as it turns out, the doubly dual notion obtained by demanding non-existence of ς-plausible attackers *does* yield such semantics.

Definition 9. *For any semantics ς and any $AF = (A, R)$: if $S \subseteq A$ is conflict free and S is not attacked in AF by any ς-plausible set from AF^S, we say that S is ς-reasonable.*

A ς-reasonable semantics is any semantics that only returns ς-reasonable sets as extensions. As before, if S is admissible, then it is trivially ς-reasonable, for any ς. We also note the following property.

Proposition 10. *For any admissible ς and any $AF = (A, R)$, if $S \subseteq A$ is ς-reasonable, then it is ς-plausible.*

Proof. Let $S \subseteq A$ be ς-reasonable and assume towards contradiction that it is not ς-plausible. Then there is some admissible set S' in AF^S that attacks S in AF. Since S' is admissible, it is not attacked in AF^S by any set from $(AF^S)^{S'}$. Hence, S' is trivially ς-plausible, contradicting the fact that S is ς-reasonable.

So for admissible ς, we have that every admissible set it ς-reasonable and that every ς-reasonable set is ς-plausible. Moreover, it is easy to show that any ς-reasonable set is weakly admissible.

Proposition 11. *For any admissible semantics ς and any $AF = (A, R)$, if $S \subseteq A$ is ς-reasonable, then S is weakly admissible.*

Proof. Let $S \subseteq A$ be ς-reasonable and assume towards contradiction that S is not weakly admissible. Then there is some weakly admissible set S' in AF^S such that S' attacks AF^S. However, since S is ς-reasonable, there is some admissible set S'' in $(AF^S)^{S'}$ that attacks S' in AF^S. But then S'' is also a weakly admissible set in $(AF^S)^{S'}$ that attacks S' in AF^S, contradicting the fact that S' is weakly admissible.

We now show the less obvious result that ς-reasonable sets are in fact modular whenever ς is admissible and modular.

Proposition 12. *For any admissible and modular ς and any $AF = (A, R)$, if $S \subseteq A$ is ς-reasonable in AF and S' is ς-reasonable in AF^S, then $S \cup S'$ is ς-reasonable in AF.*

Proof. Assume towards contradiction that $S \cup S'$ is not conflict free. Since $S' \subseteq (A \backslash [S])$ is not attacked by S, it follows that S' attacks S in AF. Since S is ς-reasonable, this means that S' is not ς-plausible, but since S' is ς-reasonable, this contradicts Proposition 10. So $S \cup S'$ is conflict free. Assume towards contradiction that there is some conflict free Q that is ς-plausible in $AF^{S \cup S'} = (AF^S)^{S'}$ and attacks $S \cup S'$ in AF. If Q attacks S', this contradicts the fact that S' is ς-reasonable in AF^S. Hence, Q does not attack S'. It follows that $Q \cup S'$ is a conflict free set from AF^S that attacks S in AF. Since S is ς-reasonable, there is a ς-extension K in $(AF^S)^{Q \cup S'} = (AF^{S \cup S'})^Q$ that attacks $Q \cup S'$ in AF^S. Since Q is ς-plausible in $(AF^{S \cup S'})$, K does not attack Q. Hence, $K \cup Q$ is a conflict free set from $AF^{S \cup S'}$ that attacks S' in AF^S. Assume towards contradiction that there is some ς-extension L in $(AF^{S \cup S'})^{K \cup Q} = ((AF^{S \cup S'})^Q)^K$ that attacks $K \cup Q$. Then since K is a ς-extension in $(AF^{S \cup S'})^Q$ and ς is modular, it follows that $K \cup L$ is a ς-extension in $(AF^{S \cup S'})^Q$ that attacks Q, contradicting the fact that Q is ς-plausible. Hence, $K \cup Q$ is ς-plausible in $AF^{S \cup S'} = (AF^S)^{S'}$, contradicting the fact that S' is ς-reasonable.

As with the preferred and weakly preferred semantics, modularity of ς implies that the ς-reasonable preferred semantics is a perfect extension of the stable semantics.

Theorem 13. *When ς is admissible and modular, then the ς-reasonable preferred semantics is a perfect extension of the stable semantics.*

Next, we will need a non-trivial graph-theoretic property of admissible sets, namely that if Q is a minimal such set containing a, then all arguments in Q have Q-alternating paths to a. To our knowledge, the following statement and proof of this fact is new, but the result is a variation of theorems from [11], regarding the closely related concepts of kernels and semi-kernels from graph theory (for more on the link between argumentation and kernel theory, see [10]). We remark that minimal non-empty admissible sets have also been studied independently

in argumentation theory [14,16], where such sets are referred to as *initial sets*. Hence, the following graph-theoretic result may well be of broader interest to the argumentation community.

Theorem 14. *For any $AF = (A, R)$ and any admissible set $S \subseteq A$ with $a \in S$: if Q is a $\subseteq$-minimal admissible set such that $Q \subseteq S$ and $a \in Q$, then for all $b \in Q$ there is a Q-alternating path $P_{b,a}$ from b to a in AF.*

Proof. Assume that Q is a $\subseteq$-minimal admissible set satisfying $Q \subseteq S$ and $a \in Q$. Let K be the set of all $b \in Q$ such that there is a Q-alternating path $P_{b,a}$ from b to a in AF. Clearly, we have $a \in K$, witnessed by the empty path. We are done if we can show that K is admissible, since then $K = Q$ by $\subseteq$-minimality of Q. Since Q is conflict free, K is conflict free. To show that K defends itself, assume c attacks K in AF at $d \in K \cap R(c)$. We must show that K defends d against c. Since $d \in K$, there is a Q-alternating path $P_{d,a}$ from d to a. Since Q is admissible and $d \in Q$, there must be some $e \in Q$ attacking c. Then $(e, c) + (c, d) + P_{d,a}$ is a walk from e to a. There are three cases. Case i) e occurs on $P_{d,a}$. In this case, the sub-path of $P_{d,a}$ from e to a is a Q-alternating path from e to a, so $e \in K$ as desired. Case ii) e does not occur on $P_{d,a}$, but c occurs on $P_{d,a}$. In this case, let $P_{c,a}$ denote the sub-path of $P_{d,a}$ from c to a. Then $(e, c) + P_{c,a}$ is a Q-alternating path from e to a, so $e \in K$ as desired. Case iii) neither e nor c occurs on $P_{d,a}$. Then $(e, c) + (c, d) + P_{d,a}$ is a Q-alternating path from e to a, so $e \in K$ in this case as well.

Notice that all Q-alternating paths from Q to Q have even length, since every other argument from such a path is from Q. It follows that we are now in a position to prove that the ς-reasonable preferred semantics is in fact justified by unsafe odd cycles.

Theorem 15. *For any admissible and modular ς and any $AF = (A, R)$: if $S \subseteq A$ is a $\subseteq$-maximal ς-reasonable set and $a \in A \backslash [S]$ attacks S, then a lies on an odd cycle that is unsafe at a.*

Proof. Assume $S \subseteq A$ is ς-reasonable and that $a \in A \backslash [S]$ attacks S. Since S is ς-reasonable, $\{a\}$ is not ς-plausible. If $(a, a) \in R$, the proof is done, so assume this is not the case. Then since $\{a\}$ is conflict free but not ς-plausible, there is a ς-extension $K \in (AF^S)^{\{a\}}$ with some $b \in K$ that attacks a. Since K is ς-plausible and S is ς-reasonable, K does not attack S. Since ς is admissible, K is an admissible set. Hence, we let Q be a $\subseteq$-minimal admissible set in $(AF^S)^{\{a\}}$ with $Q \subseteq K$ and $b \in Q$. Note that Q does not attack S and that Q is trivially ς-reasonable in $(AF^S)^{\{a\}}$, since it is admissible there. Since S is a $\subseteq$-maximal ς-reasonable set, it then follows from Proposition 12 that Q is not admissible in AF^S. Hence, there is some $c \in R(a)$ that attacks some $d \in Q$. By Theorem 14, there is a Q-alternating path $P_{d,b}$ from d to b in $(AF^S)^{\{a\}}$. Let f be the first occurrence of an argument from Q on $P_{d,b}$ that attacks a. Then $P_{d,f}$ is a path from d to f that is also a Q-alternating path from $(AF^S)^{\{a\}}$. Since a and c do not occur on $P_{d,f}$, $C = (a, c) + (c, d) + P_{d,f} + (f, a)$ is an odd cycle. If C is unsafe

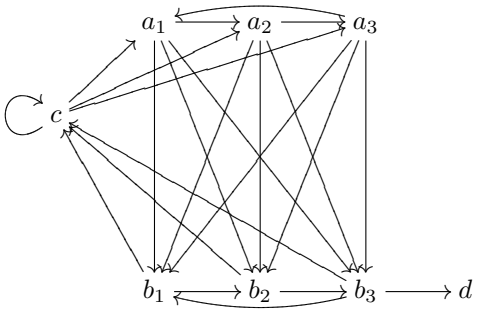
at a we are done, so assume C is safe at a. Then let $P_{a,f}$ be the sub-path of C from a to f. Since C is safe at a, there is a chord (x, y) on $P_{a,f}$ such that the sub-paths from a to x and from a to y along $P_{a,b}$ are both even. Assume towards contradiction that $x \neq y$. Since $Q \subseteq A \backslash [S \cup \{a\}]$ and x attacks y, $x \neq a$. Moreover, by our choice of f there is no argument from Q on $P_{d,f}$ that attacks a. Hence, $y \neq a$. So both x and y are in Q, contradicting the fact that Q is conflict free. So $x = y = a$. Then a attacks itself and we are done.

5.1 A Remark on Strongly Undisputed Sets

As pointed out by one of the reviewers, there is a close connection between ς-plausible and ς-reasonable sets and so-called *undisputed* and *strongly undisputed* sets, as recently introduced in [15]. In fact, when ς is the most permissive admissible semantics, then it is easy to see that the undisputed sets of AF are its ς-plausible preferred sets while the strongly undisputed sets are its ς-reasonable preferred sets. Hence, the present paper generalises the two notions, while showing how to define them without having $\subseteq$-maximality built in from the start. Moreover, the results proven about strongly undisputed sets in [15], including results on complexity which we have not addressed, carry over to preferred ς-reasonable preferred semantics when ς is the most permissive admissible semantics. Conversely, it follows from Theorem 15 that the strongly undisputed semantics is justified by unsafe odd cycles.

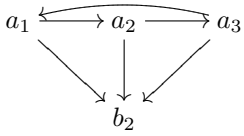
6 A Counterexample and a Conjecture

It is natural to ask about the relationship between ς-reasonable preferred sets and weakly preferred sets for admissible and modular ς. On simple examples, they behave the same way, so it is tempting to think that they might be equivalent for some admissible and modular ς. If this was true, it would mean that the recursive scheme of weak admissibility is redundant and that the nature of $\subseteq$-maximal weakly admissible sets could be described in a more succinct way in terms of admissible sets. However, the following example shows that this is not the case:



In this AF, we have two odd cycles of length 3, namely $C_a = (a_1, a_2, a_3, a_1)$ and $C_b = (b_1, b_2, b_3, b_4)$, as well as a self-attacking c and a much more innocent-looking d. All arguments on C_a attack all arguments on C_b, all arguments on

C_b attack c, and c attacks all arguments on C_a. The argument d, meanwhile, is only attacked by b_3 and attacks no argument. What is its semantic status? It is only attacked by b_3, which is on an odd cycle, so a justified semantics is entitled to ignore the attack on d, making $\{d\}$ a possible extension. However, no ς-reasonable semantics allows us to accept d when ς is admissible.
To verify that d cannot be accepted, notice that b_3 is not attacked by any admissible set from the reduct, $AF^{\{b_3\}}$:



Clearly, $AF^{\{b_3\}}$ has no admissible set, so it has no admissible set attacking b_3. This means that $\{b_3\}$ is ς-plausible in $AF^{\{d\}}$ for all admissible ς, which in turn implies that $\{d\}$ is not ς-reasonable. In fact, AF has no non-empty ς-reasonable extension for any admissible ς, as the reader can verify. The weakly preferred semantics, by contrast, provides $\{d\}$ as the unique weakly preferred extension of the AF above. This is because b_3 is attacked by a weakly admissible set from $AF^{\{b_3\}}$, namely $\{b_2\}$. Hence, we have proven the following result about the weakly preferred semantics.

Proposition 16. *The weakly preferred semantics is not equivalent to any ς-reasonable semantics for which ς is admissible.*

The counterexample also shows that while ς-reasonable semantics for admissible ς are justified, they are not strongly justified: the counterexample has no non-empty ς-reasonable set, yet it has an argument that is not on any (odd) cycle. We believe the weakly preferred semantics *is* in fact strongly justified, but we have been unable to prove it so far. Hence, we leave it as a conjecture.

Conjecture 17. *The weakly preferred semantics is strongly justified.*

The challenging part is to prove that the weakly preferred semantics is justified. If it is, then it is not hard to prove that it is also strongly justified, using the following simple lemma.

Lemma 18. *If the weakly preferred semantics is justified and AF has no non-empty weakly preferred set, then every argument a in AF lies on an odd cycle that is unsafe at a.*

Proof. Assume $AF = (A, R)$ has no non-empty weakly preferred set and let $a \in A$ be arbitrary. We must show that a lies on an odd cycle from AF that is unsafe at a. So define the AF $M = (A \cup \{b\}, R \cup \{(a, b)\})$ for some fresh $b \notin A$. Clearly, $\{b\}$ is weakly preferred since $M^{\{b\}} = AF$ has no non-empty weakly admissible set. Hence, if the weakly preferred semantics is justified, then a is on an odd cycle in AF that is unsafe at a.

The difficulty comes when we try to prove that an argument from the reduct attacking a weakly preferred set must be on an odd cycle. This is made challenging by the fact that we need to keep track of the parity of paths, without having Theorem 14 to help us. So we think the best proof strategy is to first try to establish its analogue for weakly admissible sets, if possible. Note that we could weaken the definition of a strongly justified semantics by saying that ς is weakly justified when for all $AF = (A, R)$, if $\varsigma(AF) = \{\emptyset\}$, then for all $a \in A$ there is an odd *walk* from a to a. Then it is relatively straightforward to prove by induction on the size of AF that the weakly preferred semantics is weakly justified. We omit the details for space reasons, but note that while this property goes some way towards justifying the weakly preferred semantics in formal terms, it is a much weaker property than being strongly justified.

7 Conclusion

We have provided a formal definition of the intuition that if we ignore an attack from argument a then a should be part of an unbroken odd cycle. We provided a class of semantics satisfying this requirement, showing that they are also weakly admissible and agree with the stable semantics on a large class of AFs that have stable sets. We also conjectured that the weakly preferred semantics satisfies an even stronger property, namely that whenever S is weakly preferred and $a \notin [S]$, then a is part of an unbroken odd cycle.

In future work, we would like to prove our conjecture, or find a counterexample to it. We would also like to explore the new class of semantics introduced here in further depth, as they seem to be of independent interest. It seems clear, for instance, that our notion of ς-plausibility is closely related to the labelling-based semantics explored in [9]. These labelling-based semantics should also be investigated further, not just as argumentation semantics, but as systems of three-valued logic and theories in such systems. They seem to arise from introducing an interesting conditional, whereby $a \rightarrow b$ is true just in case a is neither true nor undecided when b is false, in which case $(a \rightarrow b) \rightarrow (\neg b \rightarrow \neg a)$ is no longer valid in the presence of undecidedness. It is also interesting to explore applications of the new semantics we introduce, for instance in the context of legal argumentation, by combining them with the work done on modelling shifting proof burdens in [12]. Moreover, it would be natural to classify the new semantics in a more comprehensive way with respect to the principles investigated in [7]. Finally, we would like to generalise our results to infinite AFs, where absence of odd cycles is no longer sufficient for the existence of stable sets.

References

1. Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. Artif. Intell. **171**(10–15), 675–700 (2007)
2. Baroni, P., Giacomin, M., Guida, G.: SCC-recursiveness: a general schema for argumentation semantics. Artif. Intell. **168**(1–2), 162–210 (2005)

3. Baumann, R., Brewka, G., Ulbricht, M.: Revisiting the foundations of abstract argumentation - semantics based on weak admissibility and weak defense. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), pp. 2742–2749. AAAI Press (2020)
4. Baumann, R., Brewka, G., Ulbricht, M.: Shedding new light on the foundations of abstract argumentation: Modularization and weak admissibility. Artif. Intell. **310**, 103742 (2022)
5. Caminada, M.W.A., Carnielli, W.A., Dunne, P.E.: Semi-stable semantics. J. Log. Comput. **22**(5), 1207–1254 (2012)
6. Cramer, M., van der Torre, L.: SCF2 - an argumentation semantics for rational human judgments on argument acceptability. In: Beierle, C., Ragni, M., Stolzenburg, F., Thimm, M. (eds.) Proceedings of the 8th Workshop on Dynamics of Knowledge and Belief (DKB-2019) and the 7th Workshop KI & Kognition (KIK-2019)co-located with 44nd German Conference on Artificial Intelligence (KI 2019), Kassel, Germany, September 23, 2019. CEUR Workshop Proceedings, vol. 2445, pp. 24–35. CEUR-WS.org (2019). https://ceur-ws.org/Vol-2445/paper_3.pdf
7. Dauphin, J., Rienstra, T., van der Torre, L.: A principle-based analysis of weakly admissible semantics. In: Prakken, H., Bistarelli, S., Santini, F., Taticchi, C. (eds.) Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4–11, 2020. Frontiers in Artificial Intelligence and Applications, vol. 326, pp. 167–178. IOS Press (2020)
8. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
9. Dvorák, W., Rienstra, T., van der Torre, L., Woltran, S.: Non-admissibility in abstract argumentation. In: Toni, F., Polberg, S., Booth, R., Caminada, M., Kido, H. (eds.) Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14–16 September 2022. Frontiers in Artificial Intelligence and Applications, vol. 353, pp. 128–139. IOS Press (2022)
10. Dyrkolbotn, S.K., Walicki, M.: Propositional discourse logic. Synthese **191**(5), 863–899 (2014)
11. Galeana-Sánchez, H., Neumann-Lara, V.: On kernels and semikernels of digraphs. Discret. Math. **48**(1), 67–76 (1984)
12. Kampik, T., Gabbay, D.M., Sartor, G.: A comprehensive account of the burden of persuasion in abstract argumentation. J. Log. Comput. **33**(2), 257–288 (2023)
13. Richardson, M.: Solutions of irreflexive relations. Ann. Math. **58**(3), 573–590 (1953). http://www.jstor.org/stable/1969755
14. Thimm, M.: Revisiting initial sets in abstract argumentation. Argument Comput. **13**(3), 325–360 (2022)
15. Thimm, M.: On undisputed sets in abstract argumentation. In: The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23), pp. 6550–6557. AAAI Press (2023)
16. Xu, Y., Cayrol, C.: Initial sets in abstract argumentation frameworks. In: Ågotnes, T., Liao, B., Wáng, Y.N. (eds.) Proceedings of the 1st Chinese Conference on Logic and Argumentation (CLAR 2016), Hangzhou, China, April 2–3, 2016. CEUR Workshop Proceedings, vol. 1811, pp. 72–85. CEUR-WS.org (2016). https://ceur-ws.org/Vol-1811/paper6.pdf