Preferred Semantics as Socratic Discussion

Martin Caminada[1], *University of Aberdeen,*
Department of Computing Science,
Meston Walk, Aberdeen AB24 3UE, United Kingdom
E-mail: martin.caminada@abdn.ac.uk

Wolfgang Dvořák, *University of Vienna,*
Faculty of Computer Science, Währinger Straße 29, A-1090 Austria
E-mail: wolfgang.dvorak@univie.ac.at

Srdjan Vesic[2], *CRIL - CNRS*
Rue Jean Souvraz, SP 18, F 62307 Lens Cedex, France
E-mail: vesic@cril.fr

Abstract

In abstract argumentation theory, preferred semantics has become one of the most popular approaches for determining the sets of arguments that can collectively be accepted. However, the description of preferred semantics, as it was originally stated by Dung, has a mainly technical and mathematical nature, making it difficult for lay persons to understand what the concept of preferred semantics is essentially about. In the current paper, we aim to bridge the gap between mathematics and philosophy by providing a reformulation of (credulous) preferred semantics in terms of Socratic discussion. In order to do so, we first provide a (semi-)formal treatment of some of the concepts in Socratic dialogue.

1 Introduction

The field of formal argumentation, as a branch of non-monotonic reasoning, can be traced back to the work of Pollock [39, 40], Vreeswijk [47, 48], and Simari and Loui[44]. The idea is that (nonmonotonic) reasoning can be performed by constructing and evaluating arguments, which are composed of a number of reasons for the validity of a claim. Arguments distinguish themselves from proofs by the fact that they are defeasible, that is, the validity of their conclusions can be disputed by other arguments. The question of whether a claim can be accepted therefore depends not only on the existence of an argument that supports this claim, but also on the existence of possible counterarguments, that can then themselves be attacked by counterarguments, etc.[3]

[1] The major part of the work on this paper was carried out while MC was affiliated with the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg

[2] The major part of the work on this paper was carried out while SV was affiliated with the Computer Science and Communication Research Unit at the University of Luxembourg.

[3] A different branch of argumentation theory is concerned with the dialectical process between two parties who are involved in a discussion. This kind of argumentation, referred to as *dialogue theory* in the ASPIC project [1], can be traced back to the work of Hamblin [23, 24] and Mackenzie [31, 32]. One of the aims of the current paper is

Nowadays, much research on the topic of argumentation is based on the abstract argumentation theory of Dung [18]. The central concept in this work is that of an *argumentation framework*, which is essentially a directed graph in which the arguments are represented as nodes and the attack relation is represented by the arrows. Given such a graph, one can then examine the question on which set(s) of arguments can be accepted: answering this question corresponds to defining an *argumentation semantics*. Various proposals have been formulated in this respect, like Dung's original notions of *grounded*, *complete*, *stable* and *preferred* semantics [18], as well as subsequently stated approaches such as *stage* [46, 10], *semi-stable* [46, 7], *ideal* [17] and *eager* semantics [8]. Many of these semantics, however, have originally been defined in terms of mathematical constructs like acceptability, monotonic functions, smallest fixpoints, etc. The challenge, however, is to translate the theories stated in the field of formal argumentation into a form that is easier to be understood by lay people, who do not necessarily have an immediate understanding of the mathematical constructs on which these theories are based. That is, in order for formal argumentation theories to be implemented and applied in settings with end-users, it can be beneficial if these end-users can be given at least a conceptual understanding of the underlying theories that have been implemented in the software they are working with.

As for the topic of loop-handling, it can be observed that all the above mentioned argumentation semantics coincide for argumentation frameworks that are free of loops (directed cycles). Hence, the essential difference between the various semantics is how they deal with loops. In the current paper, we examine one of the most established ways of doing so: preferred semantics. Also, we observe how to avoid loops causing an infinite discussion (basically by disallowing participants to ask the same question twice, see Section 4).

In the current paper, we provide a description that is aimed at achieving this. We focus on one of the mainstream semantics for abstract argumentation: preferred semantics. Our aim is to show that the question of whether or not an argument is in at least one preferred extension can be described in terms of a Socratic form of discussion, in which a proponent (the defender of the claim that the particular argument is in at least one preferred extension) tries to avoid being led to a contradiction by the opponent (who essentially plays the role of Socrates in a Socratic discussion).

The remaining part of this paper is structured as follows. First in Section 2 we provide an overview of the concept of preferred semantics, as it has been treated in the literature of formal argumentation. Then in Section 3 we provide a semi-formal analysis of Socratic discussion, based on the work of Caminada [5, 9]. In Section 4 we subsequently show how the notion of Socratic discussion can be applied to (credulous) preferred semantics. That is, we show that the discussion on whether or not a particular argument is in at least one preferred extension can be described as a special form of Socratic discussion. In Section 5, we then examine the role of a winning strategy, and show how it relates to the concept of an admissible set. In Section 6, we examine the computational complexity of some of the relevant decision problems and construction problems. In Section 7 we treat three other kinds of argumentation semantics (stable, ideal and grounded) and examine what are the types of discussions that these semantics can be regarded to correspond with. We then round off with a summary of the main results, and some considerations for possible applications.

to examine how these two branches of argumentation theory (NMR and dialogue) overlap.

2 Preferred Semantics

In this section, we briefly restate some of the basic definitions of preferred semantics. Our aim is to treat both Dung's original extension-based definition [18] and Caminada *et al's* reformulation of preferred semantics in terms of argument labellings [6, 13].

Definition 2.1
An *argumentation framework* is a pair (Ar, att) where Ar is a set of arguments and $att \subseteq Ar \times Ar$.

Definition 2.2
Let $B \subseteq Ar$ be a set of arguments. We define:

- $B^+ = \{A \in Ar \mid \exists A' \in B \text{ s.t. } A' \mathit{att}\ A\}$
- $B^- = \{A \in Ar \mid \exists A' \in B \text{ s.t. } A \mathit{att}\ A'\}$

In the current paper, we assume the set of arguments in the argumentation framework to be finite. We say that argument A *attacks* argument B iff $(A,B) \in att$.

An argumentation framework can be represented as a directed graph in which the arguments are represented as nodes and the attack relation is represented as arrows. In several examples throughout this paper, we will use this graph representation.

We are now ready to treat Dung's original description of preferred semantics.[4]

Definition 2.3
Let (Ar, att) be an argumentation framework.

- $\mathcal{A}rgs \subseteq Ar$ is *conflict-free* iff there exist no $A, B \in \mathcal{A}rgs$ such that A attacks B.
- $\mathcal{A}rgs \subseteq Ar$ *defends* $A \in Ar$ iff for each $B \in Ar$ that attacks A, there exists a $C \in \mathcal{A}rgs$ that attacks B.

Definition 2.4
Let (Ar, att) be an argumentation framework. $\mathcal{A}rgs \subseteq Ar$ is *admissible* iff it is conflict-free and defends each of its elements.

Definition 2.5
Let (Ar, att) be an argumentation framework. $\mathcal{A}rgs \subseteq Ar$ is a *preferred extension* iff it is a maximal (w.r.t. set inclusion) admissible set.

Where Dung's original approach of argument-based extensions focusses on the arguments that are *accepted*, the approach of argument labellings [46, 25, 40] also takes into account the arguments that are *rejected*. In this paper, we will use the particular labellings approach of Caminada [6] and Caminada and Gabbay [13], which assigns to each argument exactly one label: `in` (to indicate that the argument is accepted), `out` (to indicate that the argument is rejected) or `undec` (to indicate that one does not have an explicit opinion on whether the argument is accepted or rejected).

Definition 2.6
Let (Ar, att) be an argumentation framework. A *labelling* is a (total) function $\mathcal{L}ab : Ar \longrightarrow \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$.

[4] We use the term *defends* instead of *acceptable* since in our view, the former term is somewhat closer to the intuitions behind the concept the terms refer to.

We write $\texttt{in}(\mathcal{L}ab)$ for $\{A \mid \mathcal{L}ab(A) = \texttt{in}\}$, $\texttt{out}(\mathcal{L}ab)$ for $\{A \mid \mathcal{L}ab(A) = \texttt{out}\}$ and $\texttt{undec}(\mathcal{L}ab)$ for $\{A \mid \mathcal{L}ab(A) = \texttt{undec}\}$.

Although a labelling by itself allows for arbitrary positions on which arguments are accepted, rejected and abstained from having an opinion about, some of these positions are more reasonable than others. One possible criterion on whether a position is reasonable ("admissible") or not is whether one can explain each argument one accepts (because all attackers are rejected and hence neutralized) and whether one can explain each argument one rejects (because it has at least one attacker one accepts, causing the attacked argument out of force). This is made formal in the following definition.

Definition 2.7
Let $\mathcal{L}ab$ be a labelling of argumentation framework (Ar, att). $\mathcal{L}ab$ is an *admissible* labelling iff for each argument $A \in Ar$ it holds that:

- if $\mathcal{L}ab(A) = \texttt{in}$ then $\forall B \in Ar : (B att A \supset \mathcal{L}ab(B) = \texttt{out})$
- if $\mathcal{L}ab(A) = \texttt{out}$ then $\exists B \in Ar : (B att A \wedge \mathcal{L}ab(B) = \texttt{in})$

Definition 2.8
Let $\mathcal{L}ab$ be a labelling of argumentation framework (Ar, att). $\mathcal{L}ab$ is a *preferred* labelling iff it is an admissible labelling where $\texttt{in}(\mathcal{L}ab)$ and $\texttt{out}(\mathcal{L}ab)$ are maximal (w.r.t. set inclusion) among all admissible labellings.

From the results by Caminada and Gabbay [13] it follows that a different way to characterise a preferred labelling is as an admissible labelling where $\texttt{in}(\mathcal{L}ab)$ is maximal, or alternatively as an admissible labelling where $\texttt{out}(\mathcal{L}ab)$ is maximal. That is, for admissible labellings the maximality of the set of in-labelled arguments implies the maximality of the set of out-labelled arguments, and vice versa.

There exists a clear overlap between admissible labellings and admissible sets. An admissible set is simply the set of in-labelled arguments of an admissible labelling. Similarly, a preferred extension is simply the set of in-labelled arguments of a preferred labelling. A more detailed treatment of the overlap between labellings and extensions can be found in the work of Caminada and Gabbay [13].

3 Socratic Argumentation

Although Dung's theory allows the internal structure of an argument to remain completely abstract, many formalisms of argumentation (such as described by Vreeswijk [47], Caminada and Amgoud [12], Wu, Caminada and Gabbay [52] and Prakken [41]) regard an argument as a structured chain of rules. An argument usually begins with one or more premises — statements that are simply regarded as true by all involved parties, such as directly observable facts. After this follows the repeated application of various rules, which generate new conclusions and therefore enable the application of additional rules. An example of such an argument is as follows:

> "Sjaak probably went to the football game, since people claim his car was parked nearby the stadium, and Sjaak is known to be a football fan."

$claimed(car\_at\_stadium),\ football\_fan,$

$claimed(car\_at\_stadium) \Rightarrow car\_at\_stadium,$
$car\_at\_stadium \wedge football\_fan \Rightarrow at\_game$

Arguments are often *defeasible*, meaning that the argument by itself is not a conclusive reason for the conclusions it brings about. Whether or not an argument should be accepted depends on its possible counterarguments. For the above argument, a possible counterargument could be:

> "Sjaak did not go to the football game, since his friends claim he was watching the game with them in a bar."

$friends\_claim(at\_bar),$
$friends\_claim(at\_bar) \Rightarrow at\_bar,$
$at\_bar \rightarrow \neg at\_game$

It then depends on the relative strength of the arguments to determine which one attacks the other one [41].

Many systems for formal argumentation take arguments to be grounded in premises; that is, each rule of the argument is ultimately (directly or indirectly) based on premises only. In human argumentation, however, one can often observe arguments which are not based on premises only, but which are at least partly based on the conclusions of the other person's argument. As an illustration, consider the following example of a discussion between the opponent and proponent of a particular thesis:

> P: "Guus did not go to the game because his mobile phone record shows he was in his mother's house at the time of the game."

$phone\_record,$
$phone\_record \Rightarrow at\_mothers\_house(phone),$
$at\_mothers\_house(phone) \Rightarrow at\_mothers\_house(Guus),$
$at\_mothers\_house(Guus) \rightarrow \neg at\_game(Guus)$

> O: "Then he would not have watched the game at all, since his mother's TV has been broken for quite a while. Don't you think that's a little odd? Guus is known to be a football fan and would definitely have watched the game."

$football\_fan(Guus),$
$at\_mothers\_house(Guus) \Rightarrow \neg watch\_game(Guus),$
$football\_fan(Guus) \Rightarrow watch\_game(Guus)$

Here, the opponent takes the propositions as uttered by the proponent as a starting point and then uses these to (defeasibly) derive a contradiction, thus illustrating the (implicit) absurdity of the proponent's original argument.

Socrates and the elenchus

The idea of taking the other party's opinion and then deriving a contradiction (or something else that is undesirable to the other party) is not new. One of the first well known examples of this style of reasoning can be found in the philosophy of

Socrates, as written down by Plato. Socrates's form of reasoning — also called the elenchus — consists of letting a proponent make a statement, and then taking this statement as a starting point to derive more statements, each of which the proponent will be committed to. The ultimate aim is to let the proponent commit himself to a contradiction, which shows that the beliefs the proponent uttered in the dialogue cannot hold together and the position as a whole should therefore be rejected.

As an example of how Socrates's form of dialectical reasoning worked, consider the following dialogue, in which Socrates questions Menexenus about the nature of friendship [37, pp. 212-213]

> (...) Answer me this. As soon as one man loves another, which of the two becomes the friend? the lover of the loved, or the loved of the lover? Or does it make no difference?
>
> None in the world, that I can see, he replied.
>
> How? said I; are both friends, if only one loves?
>
> I think so, he answered.
>
> Indeed! is it not possible for one who loves, not to be loved in return by the object of his love?
>
> It is.
>
> Nay, is it not possible for him even to be hated? treatment, if I mistake not, which lovers frequently fancy they receive at the hands of their favourites. Though they love their darlings as dearly as possible, they often imagine that they are not loved in return, often that they are even hated. Don't you believe this to be true?
>
> Quite true, he replied.
>
> Well, in such a case as this, the one loves, the other is loved.
>
> Just so.
>
> Which of the two, then, is the friend of the other? the lover of the loved, whether or not he be loved in return, and even if he be hated, or the loved of the lover? or is neither the friend of the other, unless both love each other?
>
> The latter certainly seems to be the case, Socrates.
>
> If so, I continued, we think differently now from what we did before. Then it appeared that if one loved, both were friends; but now, that unless both love, neither are friends.
>
> Yes, I'm afraid we have contradicted ourselves.

Socrates's method is that of asking questions. The questions, however, are often meant to direct the dialogue partner into a certain direction. It is the questions that force the dialogue partner to make certain inferences, as these seem to logically follow from the dialogue partner's own position. The inferences are not deductive, as they are usually based on common sense and what is reasonable. The inference is therefore more of a defeasible than of a purely deductive nature.

Socrates's elenchus is not meant for the derivation of new facts. On the contrary, its purpose is primarily destructive, meant to destroy someone's pretension of knowledge [34]. In "The Sophist", Plato provides the following definition of the elenchus [38]:

> They [those that apply the elenchus] cross-examine a man's words, when he thinks that he is saying something and is really saying nothing, and easily

> convict him of inconsistencies in his opinions; these they then collect by the dialectical process, and placing them side by side, show that they contradict one another about the same things, in relation to the same things, and in the same respect. He, seeing this, is angry with himself, and grows gentle towards others, and thus is entirely delivered from great prejudices and harsh notions, in a way that is most amusing to the hearer, and produces the most lasting effect to the person who is the subject of the operation.

The destruction of knowledge is best pursued by showing it to be incompatible with other knowledge, as argued by [36, p. 24]:

> How do we disqualify a fact or truth? The most effective way is to show its incompatibility with other facts and truths which are more certainly established, preferably with a *bundle* of facts and truths which we are not willing to abandon.

Of course, an obvious way to show incompatibility is by means of a classical (standard) counterargument, but there are also forms of incompatibility that require argumentation beyond classical arguments.

As an aside, there exists an alternative opinion on what the purpose of Socratic dialogue is. For instance, Walton [50] argues that one of the positive outcomes of the elenchus is that it can cause a participant to reconsider and refine his original position. This is in line with the view of Robinson [43], who states that "Plato quite evidently thinks of dialectic as a method of discovery as much as a method of teaching." The idea is that not only the participant that was refuted, but also anyone who reads the dialogue once it has been transcribed in a written form, can learn from it by seeing how refinement is needed (for instance, by accepting exceptions to general rules) regarding the initially simplistic positions that were put forward at the beginning stage of the dialogue, in order to make these positions more defensible and less open to attack by Socratic probing.[5] However, one should keep in mind that this refinement only takes place in a broader context, outside of the scope of the original dialogue itself. Since our aim is to describe the process of Socratic dialogue itself (so that we can later use it to describe the discussion game of preferred semantics) we restrict ourselves to study the more limited aim of refutation.

Some modern examples

The kind of reasoning in which one confronts the other party with the (defeasible) consequences of its statements is still widely used in modern times. Consider the following dialogue between politician P and interviewing journalist J:

P: In two years time, the waiting lists in health care will be as good as resolved.

J: Then you are actually saying that the insurance fees will be increased, because the government has already decided not to put more money into the health care system, and you have promised not to lower the coverage of the standard insurance.

[5] We thank one of the anonymous reviewers for this insight.

In general, one may say that in many of today's interviews where the interviewer takes a critical stance, the interviewer tries to force the interviewee to draw conclusions or make statements that the interviewee may wish to avoid. A similar phenomenon can be observed in legal cross-examination, as is for instance studied by Dunne *et al* [20].

In recent philosophical literature, Skidmore discusses the issue of *transcendental arguments*, which are meant to combat various forms of (philosophical) scepticism. The aim of a transcendental argument is "to locate something that the sceptic must presuppose in order for her challenge to be meaningful, then to show that from this presupposition it follows that the skeptic's challenge can be dismissed." [45, p. 121]. Skidmore gives various (rather long) examples of these kind of arguments — we will not repeat them here.

To summarize, the technique of using statements from the other party's argument against him is still common in modern times, both in popular as well as in philosophical argumentation. Therefore, the question of how these arguments can be formally modelled is a relevant one.

Analysis

Although a complete formal model of Socratic dialogue is outside the scope of the current paper, we would like to give a brief treatment of some of the conceptual issues. In the following examples of formal dialogue, we use the moves as have been described by Mackenzie [31]. To enhance the readability of the examples, we also use an explicit "concede" statement, with which a party indicates agreement with the other party. To illustrate the workings of (traditional) formal dialogue, consider the following example, where the proponent (P) argues that there will be a tax relief ($\mathtt{tr}$) because some leading politicians made the promise to do so ($\mathtt{pmp}$).

Example 3.1

P:	claim $\mathtt{tr}$ *"I think that there will be a tax relief."*	$C_P(\mathtt{tr})$
O:	why $\mathtt{tr}$ *"Why do you think so?"*	
P:	because $\mathtt{pmp} \Rightarrow \mathtt{tr}$ *"Because of the fact that the politicians made a promise."*	$C_P(\mathtt{pmp}, \mathtt{tr})$
O:	concede $\mathtt{tr}$ *"OK, you are right."*	$C_O(\mathtt{tr})$

Each move in a dialogue game consists of a speech act, like claim (for claiming a proposition), why (for questioning a proposition), because (for supporting a proposition) or concede (for admitting a proposition endorsed by the other party). A central notion in a dialogue system is that of a *commitment*. A commitment is a party's "official" standpoint in the dialogue, it is what the party is bound to defend when it is questioned or attacked [51].[6]

[6] Walton and Krabbe [51] distinguish three types of commitment: assertions, concessions and dark-side commitments. In their typology, only the assertions come with the obligation to defend them when challenged. We refer to [51] for details.

In the above dialogue the opponent concedes the main claim, so the proponent wins the dialogue. If, during the course of a dialogue, parties can confront each other with the (defeasible) consequences of their opinions, then a different dialogue may result. In the following example, we assume that a budget deficit ($\mathtt{bd}$) can lead to a fine from the EU ($\mathtt{feu}$), therefore ruling out the possibility of any durable tax relief.

Example 3.2

P:	claim $\mathtt{tr}$	$C_P(\mathtt{tr})$
	"I think that $\mathtt{tr}$*."*	
O:	but-then $\mathtt{tr} \Rightarrow \mathtt{bd}$	$C_O(C_P(\mathtt{bd}))$
	"Then you implicitly also hold that $\mathtt{bd}$*."*	
P:	concede $\mathtt{bd}$	$C_P(\mathtt{tr}, \mathtt{bd})$
	"Yes I do."	
O:	but-then $\mathtt{bd} \Rightarrow \mathtt{feu}$	$C_O(C_P(\mathtt{feu}))$[7]
	"Then you implicitly also hold that $\mathtt{feu}$*."*	
P:	concede $\mathtt{feu}$	$C_P(\mathtt{tr}, \mathtt{bd}, \mathtt{feu})$
	"Yes I do."	
O:	but-then $\mathtt{feu} \Rightarrow \neg\mathtt{tr}$	$C_O(C_P(\neg\mathtt{tr}))$
	"Then you implicitly also hold that $\neg\mathtt{tr}$*."*	
P:	concede $\neg\mathtt{tr}$	$C_P(\mathtt{tr}, \mathtt{bd}, \mathtt{feu}, \neg\mathtt{tr})$
	"Oops, you're right; I caught myself in..."	

Here, much akin to the Socratic dialogue treated earlier, the opponent wins the dialogue because the opponent forces the proponent to commit himself to an inconsistency.

A key feature in the above dialogue is the *but-then* statement, with which the opponent confronts the proponent with the defeasible consequences of the proponent's commitments. A but-then statement is a special form of claim, in which the speaker does not become committed himself to the consequent of the rule being claimed applicable. In general, in order to use a "but-then $\psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \phi$", the other party has to be committed to $\psi_1 \wedge \ldots \wedge \psi_n$. The immediate aim of a but-then statement is to commit him to ϕ as well. The final aim is then to get the other party to the point where it is obvious that his commitments are inconsistent.

Notice that the immediate effect of a but-then statement is a nested commitment, as is for instance shown on the second line of the above dialogue. Although this may appear odd at first, it is in fact the most appropriate way to describe the effects of the but-then statement in terms of commitments. When O says: "if you endorse $\mathtt{tr}$ then you actually also endorse $\mathtt{bd}$, don't you?" then what is it that O becomes committed to? The first thing to notice is that O does not necessarily endorse $\mathtt{bd}$ himself, so it does not hold that $C_O(\mathtt{bd})$. Furthermore, it goes too far to immediately have P committed to $\mathtt{bd}$; the rule "$\mathtt{tr} \Rightarrow \mathtt{bd}$" is defeasible and P may defend himself by giving a reason (an undercutter) why this rule does not apply (an example of this will be treated further on). Therefore, it also does not hold that $C_P(\mathtt{bd})$. The only thing that can be said regarding the but-then statement is that O claims the $\mathtt{bd}$ is implicitly endorsed by P. Therefore, it holds that $C_O(C_P(\mathtt{bd}))$.

[7] we no longer explicitly mention $C_O(C_P(\mathtt{bd}))$ since it already holds that $C_P(\mathtt{bd})$

An interesting question is how the style of reasoning of the "because" statement can be compared with that of the "but-then" statement (see also Figure 1):

1. With the because statement, reasoning goes *backwards*; the party being questioned tries to find reasons to support its thesis. With the but-then statement, on the other hand, reasoning goes *forward*; the party being questioned can be forced to make additional reasoning steps.
2. With the because statement, the *proponent* of a thesis (like ϕ in Figure 1) tries to find a path (or tree) from the premises to ϕ (the opponent's task is then to try to attack this path). With the but-then statement, on the other hand, it is the *opponent* of the thesis that tries to find a path (or tree).[8]
3. The path (or tree) constructed using because statements should ultimately originate from statements that are accepted to be *true* (such as premises), whereas the path constructed using but-then statements should ultimately lead to statements that are considered *false* (contradictions)
4. With a successfully constructed because path (or tree), both the proponent and opponent become committed to the propositions on the path, whereas with a successfully constructed but-then path (or tree), it is possible that only the proponent becomes committed to the propositions on the path.

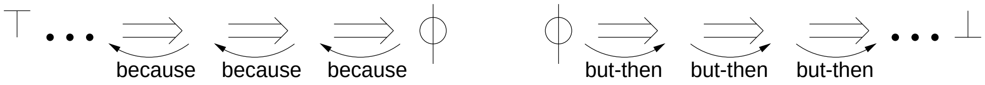


FIG. 1. "because" versus "but-then"

In the above analysis, it appears that an opponent of ϕ has two options: either trying to construct a but-then path from ϕ, or trying to prevent the proponent from successfully constructing an unattacked because path. These strategies can sometimes also be combined.

The use of a but-then statement does not automatically lead to a new commitment on the side of the other party. Sometimes, it can be successfully argued why the counterparty does not have to become committed. To illustrate why, consider again the tax-relief example, but now with the extra information that because of the current financial crisis (`fc`) the EU no longer gives any fines to member states with budget deficits. Thus, the rule `bd` $\Rightarrow$ `feu` can now be undercut.

[8] It should be noted that some of the differences between our approach and for instance the approach of [26] are related to the respective roles of the proponent and opponent. In our approach, the party that puts forward a claim is called the proponent, and the party that questions it is called the opponent. In the approach of [26, 27], something subtly different happens. A particular party called "Black" is assumed to have a particular background theory, like a philosophical system. Another party called "White" challenges this background theory by uttering a provocative thesis that is claimed to follow from the background theory of White, and that White would be eager to avoid. The idea is then to apply the dialogue game of [29, 30] with the provocative thesis as a starting point, that is, with the party that utters the provocative thesis in the role of proponent, and the party that tries to avoid it (by questioning it) in the role of opponent. However, when it comes to the feasibility of the background theory (which, in the end, is what the discussion is all about) one could argue that it is Black that should be called the proponent and White that should be called the opponent. Part of the resulting confusion can be attributed to the approach in [26, 27] of trying to reapply an *existing* dialectical formalism (such as [29, 30]) for the purpose of Socratic-style discussion.

Example 3.3

P:	claim $\mathtt{tr}$	$C_P(\mathtt{tr})$
O:	but-then $\mathtt{tr} \Rightarrow \mathtt{bd}$	$C_O(C_P(\mathtt{bd}))$
P:	concede $\mathtt{bd}$	$C_P(\mathtt{tr}, \mathtt{bd})$
O:	but-then $\mathtt{bd} \Rightarrow \mathtt{feu}$	$C_O(C_P(\mathtt{feu}))$
P:	claim $\neg\ulcorner \mathtt{bd} \Rightarrow \mathtt{feu} \urcorner$	$C_P(\mathtt{tr}, \mathtt{bd}, \neg\ulcorner \mathtt{bd} \Rightarrow \mathtt{feu} \urcorner)$
O:	why $\neg\ulcorner \mathtt{bd} \Rightarrow \mathtt{feu} \urcorner$	$C_O(C_P(\mathtt{feu}))$
P:	because $\mathtt{fc} \Rightarrow \neg\ulcorner \mathtt{bd} \Rightarrow \mathtt{feu} \urcorner$	$C_P(\mathtt{tr}, \mathtt{bd}, \neg\ulcorner \mathtt{bd} \Rightarrow \mathtt{feu} \urcorner, \mathtt{fc})$
O:	retract $C_P(\mathtt{feu})$, concede $\mathtt{tr}$	$C_O(\mathtt{tr})$

Here, the opponent again tries to construct a successful but-then path. This path, however, is undercut by the proponent. What happens next depends on the nature of the dialogue. When backtracking is allowed, the opponent may pursue another strategy. When backtracking is not allowed, the opponent loses the game.

As for the effects of the but-then statement on the commitments in the dialogue the following general remarks can be made:

1. A but-then statement is in essence a special form of a claim statement. A claim statement has as effect that a new commitment comes into existence, and this should also be the case for a but-then statement.
2. But-then statements do not in general create unnested commitments (at least, not immediately). Suppose party O utters "but-then $\psi_1 \ \wedge \ldots \wedge \ \psi_n \Rightarrow \phi$". This does of course not mean that O becomes committed to ϕ (so we do not have $C_O(\phi)$). It also does not mean that P is actually committed to ϕ (that is, we do not automatically have $C_P(\phi)$), because P may avoid commitment by successfully defending ψ_i $(1 \leq i \leq m)$ The only thing that can be said is that O feels that P is implicitly committed to ϕ (so $C_O(C_P(\phi))$), but whether P is actually committed to ϕ is still open for discussion.
3. In general, the party that makes a claim bears the responsibility of defending this claim. For instance, if P utters "claim ϕ" then upon P rests the task of defending ϕ. Similarly, if O utters "but-then $\psi_i \ \wedge \ldots \wedge \ \psi_n \ \Rightarrow \phi$" then upon O rests the task of defending $C_P(\phi)$ by making sure that P cannot avoid the conclusion ϕ. If O is unable to do so, it can lose the dialogue game.

4 Preferred Semantics as Socratic Discussion

Now that the basic principles of Socratic-style discussion have been treated, we are ready to examine how these can be applied to the concept of preferred semantics. In particular, we examine the question of how to determine whether an argument is in at least one preferred extension.

The question of whether an argument is in at least one preferred extension has been studied before by Vreeswijk and Prakken [49], who defined a formal argument game to decide this. A somewhat similar game has subsequently been specified by Mackenzie [33]. Our aim is not so much to provide an entirely new approach, rather to reinterpret the existing work, like that of Vreeswijk and Prakken [49], in the context of Socratic discussion. One of the advantages of doing so is that it can help to bridge the gap between formal argumentation theory and informal human-style discussion. Note an important difference, which is that in the work of Vreeswijk and Prakken [49] it is required for the proponent (player M) to have a *winning strategy* in the discussion game, in order for the associated argument to be in a preferred extension, whereas in the current paper, as well as in the work of Caminada and Wu [14] it is sufficient just to have a *single* game won by the proponent (player M). While at the first sight this seems confusing, we will prove that the two formalisations always give the same result.

A well-known result in formal argumentation theory is that an argument is in at least one preferred extension iff it is in at least one admissible set. Furthermore, it holds that an argument is in at least one admissible set iff it is labelled `in` by at least one admissible labelling [13]. Hence, a claim that an argument is in at least one preferred extension is essentially the same as a claim that it is labelled `in` by at least one admissible labelling. In what follows, we will examine a discussion game centred around the latter claim.

The discussion game, which consists of a reinterpretation of the work of Vreeswijk and Prakken [49], has two players which we will refer to as M and S. Player M assumes the role of Menexenus, whereas player S assumes the role of Socrates. Player M starts; his task is to defend the fact that he has a reasonable position (admissible labelling) in which a particular argument is accepted (labelled `in`). Player S then tries to confront M with the consequences of M's own position, and asks for these consequences to be resolved. Player S is successful if, like Socrates, he is able to lead his discussion partner to a contradiction.

As an example of how such a discussion can take place, consider the argumentation framework of Figure 2.

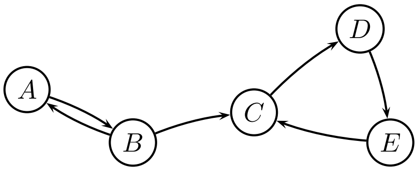


FIG. 2. An argumentation framework

Here, the player M can win the discussion game for argument D in the following way.

EXAMPLE 4.1

M: `in`(D)
"I have an admissible labelling in which D is labelled `in`*."*

S: `out`(C)
"But then in your labelling it must also be the case that D's attacker C is labelled `out`*. Based on which grounds?"*

M: $\texttt{in}(B)$
"C is labelled `out` *because B is labelled* `in`*."*

S: $\texttt{out}(A)$
"But then in your labelling it must also be the case that B's attacker A is labelled `out`*. Based on which grounds?"*

M: $\texttt{in}(B)$
"A is labelled `out` *because B is labelled* `in`*."*

As is shown in the above example, the moves of player M are statements that particular arguments are labelled `in` in M's labelling. The moves of player S, on the other hand, are meant to confront M with the consequences of his own position: "if you think that argument X is labelled `in` then you must also hold that X's attacker Y is labelled `out` in your labelling." In Section 3, we mentioned that in general, the Socratic "but-then" statement creates a nested commitment. In the particular case of the preferred semantics game, uttering $\texttt{out}(A)$ means that player S holds that player M is implicitly committed that A should be rejected. That is: $C_S(C_M(\texttt{out}(A)))$. However, it must be observed that player M has no way of avoiding this commitment, since the rule "if an argument is labelled `in` then all its attackers have to be labelled `out`" does not allow for any exceptions (the rule is strict, not defeasible). Therefore, player M has no other possibilities than to implicitly concede that A has to be labelled `out`. Hence, every out-statement (as well as every in-statement) creates an (unnested) commitment at the side of player M. Since the commitments of player M simply consist of all moves that have been uttered in the discussion so far, we do not explicitly represent them in the examples. Furthermore, the moves of player S can also be seen as *questions* about why it is legal for a particular argument Y to be labelled `out`. The moves of player M (except his first move) can then be interpreted as the *answers* to the questions of player S. Each answer follows directly to the question raised by player S. That is:

Each move of M (except the first) contains an attacker of the argument in the directly preceding move of S. (1)

Every time player M claims that an argument is labelled `in`, player S should be given the opportunity to state that as a consequence of this, player M is committed that *all* attackers of the argument are labelled `out`. The problem, however, is that each move of player S is a statement about just *one* argument. In order to deal with this problem, player S should be given the opportunity to react on the same `in`-labelled argument several times, each time confronting player M with a different `out`-labelled argument. This means that player S should be allowed to react not just on the immediately preceding move of player M, but on *any* previous move of player M.

Each move of player S contains an attacker of an argument contained in some (not necessarily the directly preceding) move of player M. (2)

Another issue is whether player S should be allowed to repeat his own moves. Recall that each move essentially contains a question ("Based on which grounds is argument Y labelled `out`?"). At the moment player S repeats one of his moves, this question has already been answered by player M, so it appears that there is no good reason to ask again. In order to avoid the discussion from going round in circles, it simply does not make sense to allow player S to repeat his moves.

Player S is not allowed to repeat his moves. (3)

On the other hand, Example 4.1 does illustrate the need for player M to be able to repeat his moves (like `in`(B)). This is because some of the questions of S (like "why is argument C `out`" and "why is argument A `out`") can have the same answer ("because argument B is `in`").

Player M is allowed to repeat his moves. (4)

The argumentation framework of Figure 2 can also be used for an example of a game won by the opponent:

Example 4.2

M: `in`(E)
"I have an admissible labelling in which E is labelled `in`*."*

S: `out`(D)
"But then in your labelling it must be the case that E's attacker D is labelled `out`*. Based on which grounds?"*

M: `in`(C)
"D is labelled `out` *because C is labelled* `in`*."*

S: `out`(E)
"But then in your labelling it must be the case that C's attacker E is labelled `out`*. This contradicts with your earlier claim that E is labelled* `in`*."*

The above example illustrates that when player S manages to use an argument uttered previously by player M, player S has won the game. After all, if player M claims an argument to be `in` and player S (still assuming the role of Socrates) subsequently manages to confront player M with the fact that in M's own position, the same argument should be labelled `out`, then player S has successfully pointed out a contradiction in M's position.

If player S uses an argument previously used by player M, then player S wins the discussion game. (5)

One can ask a similar question regarding what happens when player M uses one of the arguments previously used by player S. The fact that player S performed an `out` move means that the argument must be labelled `out` in the labelling of player M. If player M then subsequently claims that the same argument is labelled `in`, then he has directly contradicted himself.

If player M uses an argument previously used by player S, then player S wins the discussion game. (6)

There also exists a third condition under which player S wins the game. This is when player M is unable to answer one of the questions of S. This can be the case when there exists no attacker against an argument uttered by player S. Hence, player S asks why a particular argument is labelled `out` but player M is unable to come up with any attacker to be labelled `in`. In that case, player M has lost the game, for not being able to answer the critical questions of player S.

If player M cannot make a move any more, player S wins the discussion game. (7)

Similarly, one might examine what happens when it is player S who cannot make a move any more. This essentially means that player S has run out of questions. All possible relevant questions have already been asked; all relevant issues have already been raised. Moreover, player M has managed to answer all questions in a satisfactory way. Therefore, player M has survived the process of critical questioning, hence winning the discussion.

If player S cannot make a move any more, player M wins the discussion game. (8)

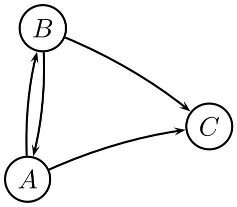


FIG. 3. An argumentation framework with floating attack

As a last illustration of the Socratic discussion game for admissible labellings, consider the argumentation framework of Figure 3. Argument C is not in any admissible set. It is illustrative to see what happens if player M tries to defend C.

EXAMPLE 4.3

M: `in`(C)
"I have an admissible labelling in which C is labelled `in`*."*

S: `out`(A)
"But then in your labelling C's attacker A must be labelled `out`*. Based on which grounds?"*

M: `in`(B)
"A is labelled `out` *because B is labelled* `in`*."*

S: `out`(B)
"But from the fact that you hold C to be `in`*, it follows that C's attacker B must be labelled* `out`*. This contradicts with your earlier claim that B is labelled* `in`*."*

The above example illustrates the need for player S to be able to respond not only to the immediately preceding move, but to any past move of player M; in the example, out(B) is a response to in(C). This is because, as we have mentioned before, for an argument to be labelled in, *all* its attackers have to be out, so player S may need to respond to a move of player M with more than one countermove.

When putting observations (1) to (8) together, we obtain the following description of the discussion game

Definition 4.4
Let (Ar, att) be an argumentation framework. An admissible discussion is a sequence of moves $[\Delta_1, \Delta_2, \ldots, \Delta_n]$ $(n \geq 0)$ such that:

- each move Δ_i $(1 \leq i \leq n)$ where i is odd is called an M-move and is of the form in(A), where $A \in Ar$
- each move Δ_i $(1 \leq i \leq n)$ where i is even is called an S-move and is of the form out(A), where $A \in Ar$
- for each S-move $\Delta_i =$ out(A) $(2 \leq i \leq n)$ there exists an M-move $\Delta_j =$ in(B) $(j < i)$ such that A attacks B
- for each M-move $\Delta_i =$ in(A) $(3 \leq i \leq n)$ it holds that Δ_{i-1} is of the form out(B), where A attacks B
- there exist no two S-moves Δ_i and Δ_j with $i \neq j$ and $\Delta_i = \Delta_j$

An admissible discussion $[\Delta_1, \Delta_2, \ldots, \Delta_n]$ is said to be *finished* iff (1) there exists no Δ_{n+1} such that $[\Delta_1, \Delta_2, \ldots, \Delta_n, \Delta_{n+1}]$ is an admissible discussion, or there exists an M-move and an S-move containing the same argument, and (2) no subsequence $[\Delta_1, \ldots, \Delta_m]$ $(m < n)$ is finished. A finished admissible discussion is won by player S if there exist an M-move and an S-move containing the same argument. Otherwise, it is won by the player making the last move (Δ_n).

We define round i as a pair $(\Delta_{2i-1}, \Delta_{2i})$.

The correctness and completeness of the game described above is stated in the following theorem.

Theorem 4.5 ([14])
Let (Ar, att) be an argumentation framework and $A \in Ar$. There exists an admissible labelling $\mathcal{L}$ with $\mathcal{L}(A) =$ in iff there exists an admissible discussion for A that is won by player M.

Theorem 4.5, together with the earlier observed facts that an argument is labelled in by an admissible labelling iff it is an element of an admissible set, and that an argument is an element of an admissible set iff it is an element of a preferred extension, implies that an argument is in a preferred extension iff player M can win the Socratic discussion game as described above. Hence, we have accomplished our goal of explaining (credulous) preferred semantics in terms of Socratic discussion.

5 Some formal properties of the admissible discussion

According to Theorem 4.5, the existence of a *single game* won by player M is sufficient for the respective argument to be in a preferred extension. This contrasts with work

[33] in which the existence of a *winning strategy* is required for an argument to be in a preferred extension. However, it turns out that for the discussion game described in the current paper (Definition 4.4) the existence of a single game won by player M coincides with the existence of a winning strategy for M. In the current section, we provide the theory that formally proves this. We first provide the preliminary notion of a tree, which will be used later to define a game tree and a winning strategy.

Definition 5.1 (Directed graph)
A directed graph is a pair (N, arr) where $arr \subseteq N \times N$. We call N the set of nodes and arr the set of arcs. A path between two nodes n and n' is a sequence of nodes $(n_1, \ldots, n_k)$ such that $n_1 = n$, $n_k = n'$, and for every $i \in \{1, \ldots, k-1\}$ it holds that $(n_i, n_{i+1}) \in arr$.

Definition 5.2 (Tree)
A tree is a directed graph in which there exists a unique node r called root, such that for any node n, there is exactly one directed path from r to n.

- If there is an arc from node n_1 to node n_2, then we say that n_1 is the parent of n_2 and that n_2 is a child of n_1.
- A node is a leaf node if and only if it has no children.
- A path $(n_1, \ldots, n_k)$ of nodes is a branch of a tree if and only if
  - $n_1 = r$, and
  - n_k is a leaf node, and
- A level of a node n is the number of nodes in the path from the root to n.

We can now define the notion of a game tree for admissible discussion. By “tree of arguments”, we mean a tree where every node is labelled by an argument from Ar.

Definition 5.3 (Correspondence between branches and discussions)
A branch b of a tree t corresponds to an admissible discussion g if and only if the sequence of labels of nodes of b, in the order from the root to the leaf node, is exactly the sequence of arguments uttered in g.

Definition 5.4 (Admissibility game tree)
A tree of arguments is an admissibility game tree for argument $A \in Ar$ if and only if it is a minimal (with respect to number of nodes) tree such that for every admissible discussion having the first move $\mathtt{in}(A)$, there exists a branch in the tree corresponding to that discussion.

The last notion we have to define is that of a pruned version of a tree. Roughly speaking, pruning consists of removing nodes from a tree. However, we still require to keep the same root and that the obtained structure is still a tree.

Definition 5.5 (A pruned version of a tree)
Let $t = (N, arr)$ and $t' = (N', arr')$ be trees. We say that t' is a pruned version of t if and only if the root of t is the same as the root of t', $N' \subseteq N$ and $arr' = arr|_{N' \times N'}$.

Now we can formalise what we mean by “winning strategy for M”. The idea is, given the game tree, to specify the move M should play in every possible node. Formally, we do this by pruning the original game tree.

Definition 5.6 (Winning strategy for M)
Let gt be the admissibility game tree for $A \in Ar$. A winning strategy for M in that game is a tree wt such that:

- wt is a pruned version of gt
- if n is a node at odd level then n has exactly the same children in gt and wt
- if n is a node at even level then n has exactly one child in wt
- every branch of wt corresponds to an admissible game won by M.

According to the previous definition, a winning strategy is a tree which contains instructions for M, telling him what exact move to make in every possible situation which could arrive if he follows the instructions. Note however, that there may exist several winning strategies.

Let us introduce an example where argument A is in one, but not in all preferred extensions, in order to illustrate that the choice of the strategy for M is important for winning the game.

Example 5.7
Suppose the argumentation framework of Figure 4. The corresponding admissibility game tree is depicted in Figure 5. In this game, there exists a unique winning strategy for M, which is depicted in Figure 6.

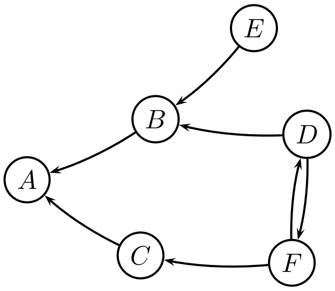


Fig. 4. Choosing strategy is important for M

The following game may take place:
M: `in`(A)
S: `out`(B)
M: `in`(D)
S: `out`(F)
M: `in`(D)
S: `out`(C)
M: `in`(F)

This represents a finished game, since F is labelled both `in` and `out`. According to Definition 4.4, S wins the game.

However, M could have won if he used another strategy. For example:

M: `in`(A)
S: `out`(B)
M: `in`(E)
S: `out`(C)
M: `in`(F)
S: `out`(D)
M: `in`(F)

Fig. 5. Admissibility game tree corresponding to Example 5.7

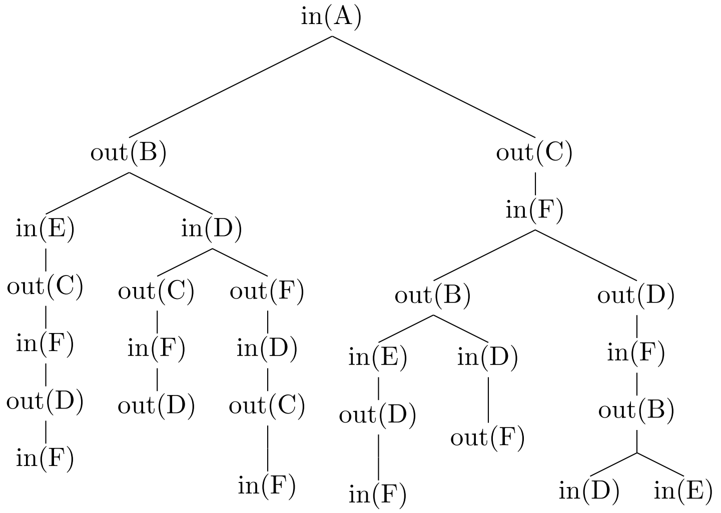
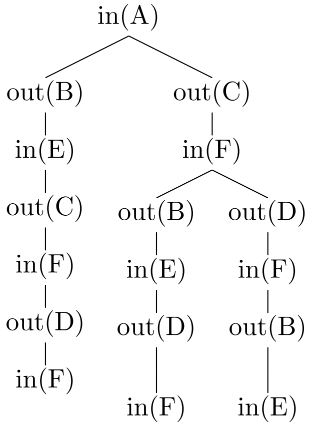


Fig. 6. A winning strategy for M corresponding to Example 5.7



We first prove a property that will be used to prove various other technical results. It states that if M won a game g, then the union of labels put on arguments by both M and S during that game, constitutes an admissible labelling.

Proposition 5.8
Let g be an admissible discussion won by M and let $\mathcal{L}ab : Ar \rightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ be a function defined as follows. For every argument $B \in Ar$:

$$\mathcal{L}ab(B) = \begin{cases} \texttt{in}, & \text{if } B \text{ was labelled } \texttt{in} \text{ during game g} \\ \texttt{out}, & \text{if } B \text{ was labelled } \texttt{out} \text{ during game g} \\ \texttt{undec}, & \text{otherwise} \end{cases}$$

Then: $\mathcal{L}ab$ is an admissible labelling.

Proof. Let us first prove that $\mathcal{L}ab$ is *well defined*. For this, it is sufficient to prove that there exists no argument $B \in Ar$ such that B was labelled both `in` and `out` during game g. From Definition 4.4, existence of such an argument would mean that S won game g, which is not the case; hence, $\mathcal{L}ab$ is well-defined.

Let us now show that $\mathcal{L}ab$ is an *admissible labelling*. Let $B \in Ar$, and $\mathcal{L}ab(B) = \texttt{in}$. From Definition 4.4, since the labelling of admissible discussion g is finished, then $\forall C \in Ar$, if $C att B$, then $\mathcal{L}ab(C) = \texttt{out}$. Let us now suppose an argument $B \in Ar$ s.t. $\mathcal{L}ab(B) = \texttt{out}$. From Definition 4.4, we conclude that $\exists C \in Ar$ such that $\mathcal{L}ab(C) = \texttt{in}$ and $C att B$. From those two facts, we conclude that $\mathcal{L}ab$ is an admissible labelling. ■

The previous result allows us to define a labelling corresponding to a game won by M.

Definition 5.9 (Associated labelling)
Let g be an admissible discussion won by M, and let $\mathcal{L}ab$ be a labelling defined as in Proposition 5.8. The labelling $\mathcal{L}ab$ is called the associated labelling of admissible discussion g.

It is clear that each winning strategy relies on an admissible set. However, there may be different winning strategies relying on the same admissible set. We would like to be able to formally represent all of them at once. For this purpose, we define a roadmap. Informally speaking, a roadmap is just a tree containing all the winning strategies based on the same admissible set.

Definition 5.10 (Roadmap)
Let $\mathcal{A}rgs \subseteq Ar$ be an admissible set and $A \in \mathcal{A}rgs$. Let gt be the admissibility game tree for $A \in Ar$. Then, we say that rm is a roadmap associated to $\mathcal{A}rgs$ and A if and only if it is a maximal (with respect to number of nodes) tree such that:

- rm is a pruned version of gt
- if n is a node at odd level then n has exactly the same children in gt and rm
- if n is a node at odd level and n corresponds to a move $\texttt{in}(B)$, then $B \in \mathcal{A}rgs$
- every branch of rm corresponds to an admissible game won by M.

For a roadmap associated to $\mathcal{A}rgs$ and A, we use the notation $AsRM(\mathcal{A}rgs, A)$.

Note that the only difference in the definition of a winning strategy and a roadmap is the third item: in a winning strategy, every node at even level has exactly one child, in order to make the playing algorithm of M deterministic, while in a roadmap, every node at even level can have one child or multiple children, but they all have to be in the admissible set $\mathcal{A}rgs$. Thus, from every roadmap, we can construct a winning strategy, by simply keeping one (arbitrary) child of every node at even level.

Also, for any pair $(\mathcal{A}rgs, A)$, if $\mathcal{A}rgs$ is an admissible set and $A \in \mathcal{A}rgs$, then there exists a unique associated roadmap $AsRM(\mathcal{A}rgs, A)$. As another important fact, note that for a given pair $(\mathcal{A}rgs, A)$, there exists an associated roadmap if and only if there exists an associated winning strategy. We already know how to convert a roadmap to a winning strategy. Conversely, if we are given a winning strategy, it is sufficient to select the admissible set it is based on, and to construct a corresponding roadmap relying on that set.

We can show that for any given pair $(\mathcal{A}rgs, A)$, where $\mathcal{A}rgs$ is an admissible set containing A, if $|\mathcal{A}rgs^-|$ is minimal, then every branch of the associated roadmap is won by M and has length $2 \cdot |\mathcal{A}rgs^-| + 1$.

Proposition 5.11
Let $\mathcal{A}rgs \subseteq Ar$ be a set such that

1. $\mathcal{A}rgs$ is admissible
2. $A \in \mathcal{A}rgs$
3. there is no other set $\mathcal{A}rgs'$ satisfying (1) and (2) s.t. $|\mathcal{A}rgs'^-| < |\mathcal{A}rgs^-|$

Then:

1. each branch of $AsRM(\mathcal{A}rgs, A)$ is won by M
2. each branch of $AsRM(\mathcal{A}rgs, A)$ has the length $2 \cdot |\mathcal{A}rgs^-| + 1$.

Proof. The first part of the proposition holds from the fourth item of Definition 5.10. We now prove the second statement. Note that M is only allowed to utter arguments from $\mathcal{A}rgs$. Also, for every argument uttered by M, S must utter all its attackers. Now we prove that since $|\mathcal{A}rgs^-|$ is minimal, every argument from $\mathcal{A}rgs^-$ will be uttered during every admissible discussion in which M plays according to a branch from $AsRM(\mathcal{A}rgs, A)$. We proceed by reductio ad absurdum. Let there exist an admissible discussion in which M plays according to a branch from $AsRM(\mathcal{A}rgs, A)$ such that there exists an argument $E \in \mathcal{A}rgs^-$ such that S does not utter E during the discussion. Let $\mathcal{A}rgs'$ be the set of arguments uttered by M during this discussion. It is trivial that $\mathcal{A}rgs' \subseteq \mathcal{A}rgs$. Also, for every $D \in \mathcal{A}rgs$, if $E att D$ then $D \notin \mathcal{A}rgs'$. We also must have $A \in \mathcal{A}rgs'$. Consequently, $|\mathcal{A}rgs'^-| < |\mathcal{A}rgs^-|$. Contradiction. Thus, it must be that every argument from $\mathcal{A}rgs^-$ is uttered by S during every discussion where M plays according to a branch of $AsRM(\mathcal{A}rgs, A)$. Since S must utter all the attackers of all the arguments uttered by M, and S cannot repeat his moves, then S will move exactly $|\mathcal{A}rgs^-|$ times. Since M plays first, and M has an attacker of every argument uttered by S, M will play $|\mathcal{A}rgs^-| + 1$ times. Thus, the discussion has $2 \cdot |\mathcal{A}rgs^-| + 1$ moves. ∎

One particular consequence of Proposition 5.11 is that once a minimal admissible set $\mathcal{A}rgs$ containing A has been fixed, the length of discussion does not depend on the particular moves of player S, as long as player M follows the associated roadmap when choosing his moves.

6 Computational Aspects

Suppose the aim is to apply the admissible discussion game to convince the user that a particular argument is in an admissible set. In general, there might be several different strategies for doing so. For instance, in the example of Figure 7, there are two winning strategies for argument A, corresponding to admissible sets $\{A, C\}$ and $\{A, D\}$. The first winning strategy consists of a single discussion: $\texttt{in}(A), \texttt{out}(B), \texttt{in}(C), \texttt{out}(E), \texttt{in}(C)$. The second winning strategy consists of a single discussion: $\texttt{in}(A), \texttt{out}(B), \texttt{in}(D)$. Although both winning strategies do the job of showing that an argument is in an admissible set, the second strategy has less moves than the first one.

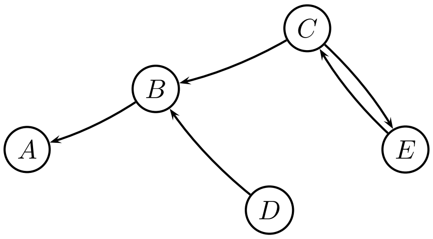


Fig. 7. How to find the shortest discussion?

When applying a discussion game in a human-computer setting, it can make sense to try to minimise the number of moves in the game in order not to have the user interact with the system unnecessarily long. From the results in the previous section, we know that the length of the discussion (more precisely, the length of the discussion within a winning strategy or roadmap) is $2 \cdot |\mathcal{A}rgs^-| + 1$, where $\mathcal{A}rgs$ is the associated admissible set. That is, a minimal length discussion corresponds to an admissible set $\mathcal{A}rgs$ with minimal $|\mathcal{A}rgs^-|$. Therefore, it makes sense to try to find an admissible set $\mathcal{A}rgs$ that contains the argument we are interested in and where $|\mathcal{A}rgs^-|$ is minimal. In the following we will call an admissible set $\mathcal{A}rgs$ minimal admissible for an argument A if $A \in \mathcal{A}rgs$ and $|\mathcal{A}rgs^-|$ is minimal among all these sets. In the current section, we show the complexity of the related problems.

We first consider the following two decision problems, verifying that a given discussion is of minimal length and deciding whether there is a discussion within a given bound on the length of the discussion.

Definition 6.1 (Verification problem)
Given: An argumentation framework (Ar, att) an argument $A \in Ar$ and a set $\mathcal{A}rgs \subseteq Ar$.
Question: Is $\mathcal{A}rgs$ minimal admissible for A?

Definition 6.2 (Existence problem)
Given: An argumentation framework (Ar, att), an argument $A \in Ar$ and an integer k.
Question: Is there a set $\mathcal{A}rgs$ that is minimal admissible for A with $|\mathcal{A}rgs^-| \leq k$?

It is well-known that verifying an admissible set, i.e. a witness that an argument is credulously accepted, can be done in polynomial time (see e.g. [19]), but as we show next, adding minimality of $|\mathcal{A}rgs^-|$ makes even the Verification problem coNP-complete.

We first show that the problem can be solved in coNP.

Lemma 6.3
The Verification Problem is in coNP.

Proof. The *membership* in the class coNP is by a guess and check algorithm for falsifying a set to be a minimal admissible set. Given $\mathcal{A}rgs$ the algorithm first checks whether $\mathcal{A}rgs$ is an admissible set and $A \in \mathcal{A}rgs$, and if not the algorithm immediately terminates. Then the algorithm guesses a set $\mathcal{A}rgs'$ and checks whether $\mathcal{A}rgs'$ is an admissible set and $A \in \mathcal{A}rgs'$ (both can be done in polynomial time). Now it is sufficient to test whether $|\mathcal{A}rgs'^{-}| < |\mathcal{A}rgs^{-}|$, which can also be done in polynomial time. Hence falsifying a minimal admissible set is in NP, and thus verifying it is in coNP. ∎

To show hardness we introduce a reduction from the 3-SAT problem which is an extension of the standard reduction for abstract argumentation [21].

Reduction 6.4
Given a 3-CNF formula φ over variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ as a set $\mathcal{C}$ of clauses, where each clause $C \in \mathcal{C}$ is a set over atoms and negated atoms (denoted by $\bar{X}$). The argumentation framework $\mathtt{AF}_\varphi = (Ar, att)$ with

$$\begin{aligned} Ar &= \mathcal{X} \cup \bar{\mathcal{X}} \cup \mathcal{C} \cup \{T, \bar{T}, B, G\} \cup \{F_i \mid 1 \leq i \leq |\mathcal{X}| + |\mathcal{C}| + 1\} \\ att &= \{(X, \bar{X}), (\bar{X}, X) \mid X \in \mathcal{X}\} \cup \{(L, C) \mid L \in C, C \in \mathcal{C}\} \cup \\ &\quad \{(C, T) \mid C \in \mathcal{C}\} \cup \\ &\quad \{(F_i, \bar{T}), (\bar{T}, F_i), (F_i, F_i) \mid 1 \leq i \leq |Ar| + |att| + 1\} \cup \\ &\quad \{(T, B), (\bar{T}, B), (B, B), (B, G)\} \end{aligned}$$

where $\bar{\mathcal{X}} = \{\bar{X} \mid X \in \mathcal{X}\}$ and $F_i, T, \bar{T}, B, G$ are fresh arguments.

Reduction 6.4 is illustrated in Figure 8 for the 3-CNF φ with clauses $C_1 = \{X_1, X_2, X_3\}$, $C_2 = \{\bar{X}_2, \bar{X}_3, \bar{X}_4\}$, and $C_3 = \{\bar{X}_1, X_2, X_4\}$. The idea behind Reduction 6.4 is that we are interested in admissible sets containing G and there are only two ways to defend G against the attack from argument B. First, we can add $\bar{T}$ to the admissible set. This results in the admissible set $\{G, T\}$ which has $|Ar| + |att| + 3$ attackers. Second we can add T to the admissible set. But by the structure of $\mathtt{AF}_\varphi$ the argument T can only be defended if φ is satisfiable. If this is the case this gives an admissible set with at most $|Ar| + |att| + 2$ attackers. Thus we then have that $\{G, T\}$ is minimal admissible if and only if φ is unsatisfiable. We make this argument formal by the following lemma.

Lemma 6.5
For each 3-CNF formula φ it holds that φ is satisfiable iff $\mathcal{A}rgs = \{\bar{T}, G\}$ is not a minimal admissible set for $\mathtt{AF}_\varphi$.

Proof. We first show that $\{\bar{T}, G\}$ is always an admissible set. First it is conflict-free as there is no attack between $\bar{T}$ and G. Further we have that $\mathcal{A}rgs^{-} = \{F_i \mid 1 \leq i \leq |\mathcal{X}| + |\mathcal{C}| + 1\} \cup \{T, B\}$ and each of these arguments is attacked by $\bar{T}$. Hence $\{\bar{T}, G\}$ is admissible.

⇒: Let us assume φ is satisfiable and let $M \subseteq \mathcal{X}$ be a model of φ. Then it is easy to verify that the set $\mathcal{A}rgs' = \{T, G\} \cup M \cup \{\bar{X} \mid X \in \mathcal{X} \setminus M\}$ is conflict free in

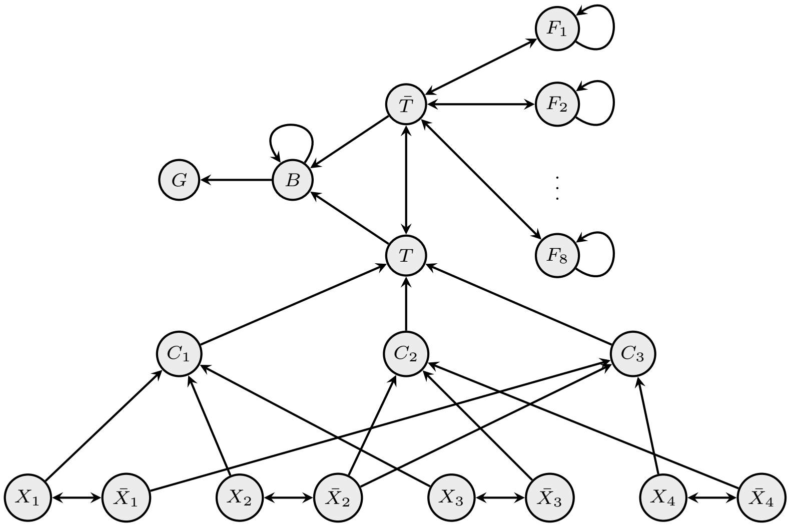


FIG. 8: The argumentation framework $\mathtt{AF}_\varphi$, as defined in Reduction 6.4, for a CNF formula φ with clauses $C_1 = \{X_1, X_2, X_3\}$, $C_2 = \{\bar{X}_2, \bar{X}_3, \bar{X}_4\}$, and $C_3 = \{\bar{X}_1, X_2, X_4\}$.

$\mathtt{AF}_\varphi$. The set $\mathcal{A}rgs'^{-}$ is given by $\mathcal{X} \setminus M \cup \{\bar{X} \mid X \in M\} \cup \mathcal{C} \cup \{\bar{T}, B\}$. Now (i) the arguments in the set $\mathcal{X} \setminus M \cup \{\bar{X} \mid X \in \mathcal{X} \cap M\}$ are attacked by their duals in $\mathcal{A}rgs'$; (ii) by the assumption that M is a model each $C \in \mathcal{C}$ is attacked by an argument in $M \cup \{\bar{X} \mid X \in \mathcal{X} \setminus M\}$; and (iii) $\bar{T}, B$ are attacked by T. Now as the set $\mathcal{A}rgs'$ attacks all arguments in $\mathcal{A}rgs'^{-}$ it is also admissible.

Now let us compare $\mathcal{A}rgs'^{-}$ with $\mathcal{A}rgs^{-}$. By construction of $\mathtt{AF}_\varphi$ we have that $\mathcal{A}rgs^{-} = \{F_i \mid 1 \leq i \leq |\mathcal{X}| + |\mathcal{C}| + 1\} \cup \{T, B\}$. Now as $|\mathcal{A}rgs'^{-}| = |\mathcal{X}| + |\mathcal{C}| + 2$ and $|\mathcal{A}rgs^{-}| = |\mathcal{X}| + |\mathcal{C}| + 3$ the set $\mathcal{A}rgs$ is not minimal admissible for G.

$\Leftarrow$: Let us assume that $\{\bar{T}, G\}$ is not minimal admissible for G. As $\{\bar{T}, G\}$ is admissible there also exists a minimal admissible set $\mathcal{A}rgs'$ for G. As $|(.)^{-}|$ is a monotonic operator $\mathcal{A}rgs'$ cannot be a superset of $\{\bar{T}, G\}$ and thus we can conclude that $\bar{T} \notin \mathcal{A}rgs'$. Now as $\mathcal{A}rgs'$ has to defend G against B and B is only attacked by itself, T and $\bar{T}$ we obtain that $T \in \mathcal{A}rgs'$. Furthermore, $\mathcal{A}rgs'$ must defend T against all $C \in \mathcal{C}$ and thus, by construction of $\mathtt{AF}_\varphi$, we have $\mathcal{A}rgs' \cap \mathcal{X}$ is a model of φ, i.e. φ is satisfiable. ∎

Together with the previous results we obtain that the Verification Problem is complete for the class coNP.

THEOREM 6.6
The Verification Problem is coNP complete.

PROOF. The membership in the class coNP is by Lemma 6.3. Hardness follows from the facts that Reduction 6.4 (i) can be performed in polynomial time and (ii) by Lemma 6.5 reduces the coNP-hard problem of 3-UNSAT to the Verification Problem. ∎

We now turn to the Existence Problem.

Theorem 6.7
The Existence Problem is NP-complete.

Proof. The membership is again by a guess and check algorithm. Simply guess a set $\mathcal{A}rgs \subseteq Ar$ and test whether it is admissible, $a \in \mathcal{A}rgs$ and $|S^-| \leq k$.

Hardness follows from the fact that for $k = |Ar|$ this problem coincides with credulous acceptance w.r.t. admissible sets, i.e. with the problem of deciding whether an argument is in at least one admissible set, which is well-known to be NP-complete [16, 19]. ∎

The above results already show that computing a minimal admissible set is computationally hard. However the above studied decision problems do not fully cover the characteristics of constructing a minimal admissible set for a given argument. In the following we give results about the functional complexity of computing the length of the minimal discussions and constructing a minimal admissible set. More precisely, we show that our problem is closely related to MAX-SAT, the problem of determining the maximal number of simultaneously satisfiable clauses of a CNF formula.

When dealing with function problems we use a more general notion of reductions, so called metric reductions.

Definition 6.8 (metric reductions)
A metric reduction from a (function) problem A to a (function) problem B is a pair (R, T) of polynomial time computable functions such that

- if x is an instance of A then $R(x)$ is an instance of B; and
- if x is an instance of A and z a correct output of B w.r.t. R(x) then $T(x, z)$ is a correct output of $A(x)$.

In the following, we consider function problems concerning minimal admissible sets. First, we are interested in the minimal length of a discussion supporting an argument A, or in other terms the minimum $|\mathcal{A}rgs^-|$ for all minimal admissible sets $\mathcal{A}rgs$ for A.

Definition 6.9 (Minimum Discussion Length Problem)
Given: An argumentation framework (Ar, att) and an argument $A \in Ar$.
Task: Compute the minimum $|\mathcal{A}rgs^-|$ for all minimal admissible sets $\mathcal{A}rgs$ for A.

Further we are interested in constructing a minimal admissible set for an given argument:

Definition 6.10 (Minimal Admissible Set Problem)
Given: An argumentation framework (Ar, att) and an argument $A \in Ar$.
Task: Compute a minimal admissible set $\mathcal{A}rgs$ for A.

Clearly the second problem is the harder one, as whenever a minimal admissible set $\mathcal{A}rgs$ for A is given one can easily compute $|\mathcal{A}rgs^-|$.

Next, let us formally define the two corresponding problems for MAX-SAT, and then review known complexity results for them.

Definition 6.11 (MAX-SAT)
Given: A formula in CNF (or, equivalently, a set of clauses).
Task: Compute the maximal number of clauses simultaneously satisfied by a truth assignment.

Definition 6.12 (MAX-SAT Assignment)
Given: A formula in CNF (or, equivalently, a set of clauses).
Task: Find a truth assignment satisfying a maximal number of clauses.

Next let us briefly recapitulate the concept of oracle machines and complexity classes defined on top of them. Let $\mathcal{C}$ denote some complexity class. By a $\mathcal{C}$-oracle machine we mean a polynomial time Turing machine which can access an oracle that decides a given (sub)-problem in $\mathcal{C}$ within one step. We denote the class of function problems, that can be solved by such machines, as $FP^{\mathcal{C}}$ if the underlying Turing machine is deterministic. Concretely, we will consider the class $FP^{\mathrm{NP[log\,n]}}$, i.e. the class of functions that can be computed in polynomial time with a logarithmic number of calls to an NP-oracle.

Theorem 6.13 ([28])
MAX-SAT is $FP^{\mathrm{NP[log\,n]}}$-complete.

However, to the best of the authors' knowledge there is no explicit complexity classification for the above MAX-SAT Assignment problem in terms of metric reductions[9].

We are now prepared for treating the functional complexity of our problems. We will show that the MAX-SAT problems are of the same complexity as the corresponding versions of our minimal admissible set problem. To prove this we provide two reductions. The first one constructs an argumentation framework out of 3-CNF formula while the other one works in the reverse direction and builds a propositional formula out of a given argumentation framework. Then given that the problems are reducible to each other we obtain that they are of the same complexity. We first present the reduction from 3-CNF formulas to argumentation frameworks.

Reduction 6.14
For a 3-CNF formula φ over variables $\mathcal{X}$ given as a set $\mathcal{C}$ of clauses, where the argumentation framework $\mathtt{AF}_\varphi = (Ar, att)$ is given by

$$\begin{aligned}
Ar &= \mathcal{X} \cup \bar{\mathcal{X}} \cup \mathcal{C} \cup \mathcal{C}' \cup \{T\} \cup \{G_{i,C} \mid 1 \leq i \leq |\mathcal{X}|+1, C \in \mathcal{C}\}\\
att &= \{(X,\bar{X}), (\bar{X},X) \mid X \in \mathcal{X}\} \cup \{(L,C) \mid L \in C, C \in \mathcal{C}\} \cup\\
&\quad \{(C,T), (C',C) \mid C \in \mathcal{C}\} \cup\\
&\quad \{(G_{i,C},C'), (C',G_{i,C}), (G_{i,C},G_{i,C}) \mid 1 \leq i \leq |\mathcal{X}|+1, C \in \mathcal{C}\}
\end{aligned}$$

where $\bar{\mathcal{X}} = \{\bar{X} \mid X \in \mathcal{X}\}$, $\mathcal{C}' = \{C' \mid C \in \mathcal{C}\}$ and $T, G_{i,c}$ are fresh arguments.

An example of Reduction 6.14 is given in Figure 9. Before proving the correctness of the reduction we first discuss the intuition behind it. We are interested in the argument T which is attacked by the arguments C_i encoding the clauses of φ. Thus an admissible set has to attack all C_i. This can be done by arguments $X_i, \bar{X}_i$ corresponding to a (partial) truth assignment of φ or by the additional arguments C_i'. But each argument C_i' has $|\mathcal{X}|+1$ many attackers and thus a minimal admissible set minimizes the number of such arguments. Hence it also maximizes the number of clauses satisfied by the (partial) truth assignment corresponding to the $X_i, \bar{X}_i$ arguments in the set.

[9] There are several complexity classifications concerning the approximability of MAX-SAT, see e.g. [35, 15].

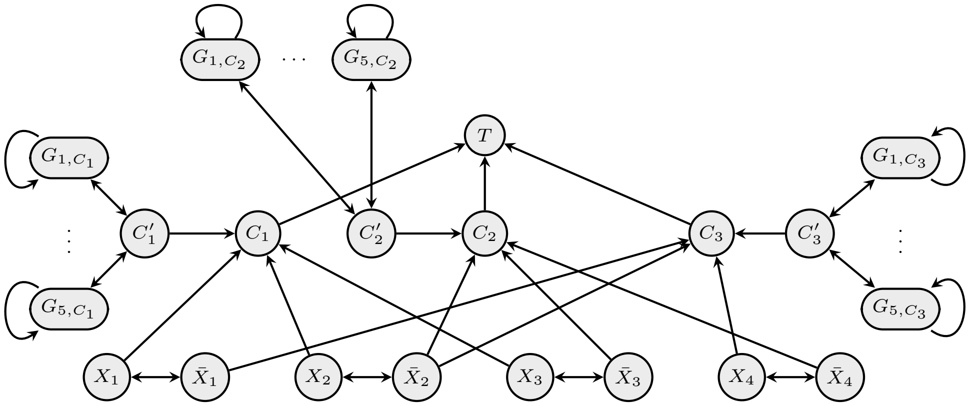


Fig. 9: The argumentation framework $\mathtt{AF}_\varphi$, as constructed by Reduction 6.14, for a CNF formula φ with clauses $C_1 = \{X_1, X_2, X_3\}$, $C_2 = \{\bar{X}_2, \bar{X}_3, \bar{X}_4\}$, and $C_3 = \{\bar{X}_1, X_2, X_4\}$.

Definition 6.15
For a CNF formula φ and a partial truth assignment α we use $c_\varphi(\alpha)$ to denote the number of clauses satisfied by α and $\mathrm{dom}(\alpha)$ to denote the domain of α, i.e. the set of the atoms $X \in \mathcal{X}$ that α assigns to a truth value.

We first relate (partial) truth assignments to admissible sets.

Lemma 6.16
Given a (partial) truth assignment α then the set $\mathcal{A}rgs = \{T\} \cup \{X \mid \alpha(X) = 1\} \cup \{\bar{X} \mid \alpha(X) = 0\} \cup \{C' \mid \alpha(C) \neq 1\}$ is admissible in $\mathtt{AF}_\varphi$ and $|\mathcal{A}rgs^-| = |\mathcal{C}| + |\,\mathrm{dom}(\alpha)| + (|\mathcal{C}| - c_\varphi(\alpha))(|\mathcal{X}| + 1)$.

Proof. It is easy to check that $\mathcal{A}rgs$ is conflict-free and thus it remains to show that each argument in $\mathcal{A}rgs$ is defended by $\mathcal{A}rgs$. First consider an argument $A \in \{X \mid \alpha(X) = 1\}$. Then A is only attacked by $\bar{A}$. As also A attacks $\bar{A}$, we obtain that A is defended by $\mathcal{A}rgs$. By symmetry also each $A \in \{\bar{X} \mid \alpha(X) = 0\}$ is defended. Next, consider $A \in \{C' \mid \alpha(C) \neq 1\}$. Again as all incoming attacks are mutual attacks, A defends itself. Finally, consider the argument T which is attacked by all $C \in \mathcal{C}$. By construction we have that if $\alpha(C) = 1$ then there is either a $X \in C$ with $\alpha(X) = 1$ attacking C or a $\bar{X} \in C$ with $\alpha(X) = 0$ attacking C. In both cases, C is attacked by $\mathcal{A}rgs$. Otherwise if $\alpha(C) \neq 1$ then $C' \in \mathcal{A}rgs$ and as C' attacks C we obtain that $\mathcal{A}rgs$ attacks all $C \in \mathcal{C}$ and thus defends T.

Next, consider $|\mathcal{A}rgs^-|$. As mentioned before, $C \subseteq \mathcal{A}rgs^-$ as each $C \in \mathcal{C}$ attacks the argument T. Then for each argument X / $\bar{X}$ that is assigned by α the dual argument $\bar{X}$ / X goes to $\mathcal{A}rgs^-$. That adds $|\,\mathrm{dom}(\alpha)|$ arguments to the set. Finally, for each unsatisfied clause C, we get the $|\mathcal{X}| + 1$ attackers of C. As there are $|\mathcal{C}| - c_\varphi(\alpha)$ unsatisfied clauses this adds $(|\mathcal{C}| - c_\varphi(\alpha))(|\mathcal{X}| + 1)$ arguments to $\mathcal{A}rgs^-$. ∎

Next we relate admissible sets to (partial) truth assignments.

Lemma 6.17
Given an admissible set $\mathcal{A}rgs$ of $\mathtt{AF}_\varphi$ with $T \in \mathcal{A}rgs$ and the (partial) truth assignment α, with $\alpha(X) = 1$ if $X \in \mathcal{A}rgs$ and $\alpha(X) = 0$ if $\bar{X} \in \mathcal{A}rgs$ then for each $C \in \mathcal{C}$, $\alpha(C) = 1$ if $C' \notin \mathcal{A}rgs$. Moreover there is an admissible set $\mathcal{A}rgs'$ with $T \in \mathcal{A}rgs' \subseteq \mathcal{A}rgs$ and $|\mathcal{A}rgs'^-| = |\mathcal{C}| + |\operatorname{dom}(\alpha)| + (|\mathcal{C}| - c_\varphi(\alpha))(|\mathcal{X}| + 1)$.

Proof. First notice that for an admissible set it cannot be the case that both $X \in \mathcal{A}rgs$ and $\bar{X} \in \mathcal{A}rgs$ as they attack each other. Thus the partial truth assignment α is well-defined. As $T \in \mathcal{A}rgs$, the set $\mathcal{A}rgs$ attacks all $C \in \mathcal{C}$. Now consider a $C \in \mathcal{C}$ such that $C' \notin \mathcal{A}rgs$. Then, by construction of $\mathtt{AF}_\varphi$, C' is either attacked by an $X \in \mathcal{A}rgs$ with $X \in C$ or by an $\bar{X} \in \mathcal{A}rgs$ with $\bar{X} \in C$. In both cases, $\alpha(C) = 1$.
Next consider the set $\mathcal{A}rgs' = \mathcal{A}rgs \setminus \{C' \mid \alpha(C) = 1\}$. By Definition $T \in \mathcal{A}rgs' \subseteq \mathcal{A}rgs$. Now consider the cardinality of the set $\mathcal{A}rgs'^-$. As in Lemma 6.16, $\mathcal{C} \subseteq \mathcal{A}rgs'^-$ and for each argument X / $\bar{X}$ that is assigned by α the dual argument $\bar{X}$ / X goes to $\mathcal{A}rgs'^-$. Also for each $C' \in \mathcal{A}rgs$ we get the $|\mathcal{X}| + 1$ attackers $G_{i,C}$ of C'. As there are $|\mathcal{X}| - c_\varphi(\alpha)$ unsatisfied clauses, by the above observation, there are as many $C' \in \mathcal{A}rgs'$. Hence this adds at least $(|\mathcal{X}| - c_\varphi(\alpha))(|\mathcal{X}| + 1)$ arguments to $\mathcal{A}rgs'^-$. ∎

Together Lemma 6.16 and Lemma 6.17 give us a correspondence between partial truth assignments and admissible sets containing T.

The following lemma exploits this correspondence to show that Reduction 6.14 is a valid reduction from the MAX-SAT problems to our minimal admissible set problems.

Lemma 6.18
Consider a 3-CNF φ and the argumentation framework $\mathtt{AF}_\varphi$ from Reduction 6.14.

1. For a minimal admissible set $\mathcal{A}rgs$ the truth assignment α, with $\alpha(X) = 1$ if $X \in \mathcal{A}rgs$ and $\alpha(X) = 0$ otherwise, satisfies the maximal number of clauses of φ.
2. Given a minimal admissible set $\mathcal{A}rgs$ then the number of simultaneously satisfiable clauses is given by $|C| - \left\lfloor \frac{|\mathcal{A}rgs^-| - |\mathcal{C}|}{|\mathcal{X}|+1} \right\rfloor$.

Proof. Using lemmas 6.16 and 6.17 and the fact that $|\operatorname{dom}(\alpha)| < |\mathcal{X}| + 1$ we get that an admissible set is minimal iff $c_\varphi(\alpha)$ is maximal for the corresponding partial truth assignment α, i.e. the minimal admissible sets corresponds to an optimal partial truth assignments. So given a minimal admissible set for T, by Lemma 6.17 we can construct a truth-assignment satisfying the maximal number of clauses of φ. This shows (1).

For a minimal admissible set $\mathcal{A}rgs$ by Lemma 6.17 we have $|\mathcal{A}rgs^-| = |\mathcal{C}| + |\operatorname{dom}(\alpha)| + (|\mathcal{C}| - c_\varphi(\alpha))(|\mathcal{X}|+1)$ which is equivalent to $c_\varphi(\alpha) = |\mathcal{C}| - \frac{|\mathcal{A}rgs^-| - |\mathcal{C}| - \operatorname{dom}(\alpha)}{|\mathcal{X}|+1}$. To eliminate the $\operatorname{dom}(\alpha)$ term we use that $|\operatorname{dom}(\alpha)| \leq |\mathcal{X}|$ and obtain $c_\varphi(\alpha) = |\mathcal{C}| - \left\lfloor \frac{|\mathcal{A}rgs^-| - |\mathcal{C}|}{|\mathcal{X}|+1} \right\rfloor$. This shows (2). ∎

For the reverse direction we give a reduction from an arbitrary minimal admissible set instance to a MAX-SAT instance.

Reduction 6.19
Given an argumentation framework (Ar, att) and an argument $\beta \in Ar$, we build the

following CNF formula

$$\varphi_{(Ar,att),\beta} = x_\beta \wedge \bigwedge_{(A,B)\in att} (\neg x_A \vee \neg x_B) \wedge \bigwedge_{(B,A)\in att} (\neg x_A \vee \bigvee_{(C,B)\in att} x_C) \wedge \tag{6.1}$$

$$\bigwedge_{(A,B)\in att} (x'_A \vee \neg x_B) \wedge \bigwedge_{A\in Ar} (\neg x'_A \vee \bigvee_{(A,B)\in att} x_B) \wedge \tag{6.2}$$

$$\bigwedge_{A\in Ar} \neg x'_A \tag{6.3}$$

Moreover we make $|Ar| + 1$ many copies of each clause in (6.1) and(6.2).

With the clauses (6.1) we encode that the set $\mathcal{A}rgs = \{A \in Ar \mid x_A = 1\}$ must be admissible and contain α. The clauses (6.2) introduce additional variables x'_A encoding that $A \in \mathcal{A}rgs^-$. We will use the clauses (6.3) to minimize the set $\mathcal{A}rgs^-$. The copies of the clauses in (6.1) and (6.2) implement a higher weighting of these clauses and encode that we do not want that any of these clauses is violated. We exemplify Reduction 6.19 for the argumentation framework in Figure 7 (for convenience we omit repetitions of clauses):

$$\begin{aligned}\varphi_{(Ar,att),A} = {} & x_A \wedge (\neg x_A \vee \neg x_B) \wedge (\neg x_B \vee \neg x_C) \wedge (\neg x_B \vee \neg x_D) \wedge (\neg x_C \vee \neg x_E) \wedge \\ & (\neg x_A \vee x_C \vee x_D) \wedge (\neg x_B \vee x_E) \wedge (\neg x_B) \wedge (\neg x_C \vee x_C) \wedge (\neg x_E \vee x_E) \wedge \\ & (x'_B \vee \neg x_A) \wedge (x'_C \vee \neg x_B) \wedge (x'_D \vee \neg x_B) \wedge (x'_E \vee \neg x_C) \wedge (x'_C \vee \neg x_E) \wedge \\ & \neg x'_A \wedge (\neg x'_B \vee x_A) \wedge (\neg x'_C \vee x_B \vee x_E) \wedge (\neg x'_D \vee x_B) \wedge (\neg x'_E \vee x_C) \wedge \\ & \neg x'_A \wedge \neg x'_B \wedge \neg x'_C \wedge \neg x'_D \wedge \neg x'_E\end{aligned}$$

Lemma 6.20

For an optimal satisfying truth assignment α of $\varphi_{(Ar,att),\beta}$ and c_α the number of satisfied clauses, either (i) $\mathcal{A}rgs = \{A \in Ar \mid \alpha(x_A) = 1\}$ is a minimal admissible set of (Ar, att) containing β and $|\mathcal{A}rgs^-| = |Ar| - (c_\alpha - m \cdot (|Ar| + 1))$, or (ii) $c_\alpha < m \cdot (|Ar| + 1)$ and there is no admissible set containing β, with m being the number of clauses in (6.1) and (6.2), i.e. $m = (3 \cdot |att| + |Ar| + 1)$.

Proof. First, $\mathcal{A}rgs$ satisfies the clauses in (6.1) iff it is admissible and contains the argument β (cf. [2, 22]). Next we show that $\{A \in Ar \mid \alpha(x'_A) = 1\} = \mathcal{A}rgs^-$ iff the clauses in (6.2) are satisfied.

- Assume $\{A \in Ar \mid \alpha(x'_A) = 1\} = \mathcal{A}rgs^-$: For each attack (A, B) either $B \notin \mathcal{A}rgs$ or $A \in \mathcal{A}rgs^-$ and thus the clause $(x'_A \vee \neg x_B)$ is satisfied.

  If $A \in \mathcal{A}rgs^-$ then there exits an B with $(A, B) \in attack$ and $B \in \mathcal{A}rgs$. Thus $\neg x'_A \vee \bigvee_{(A,B)\in att} x_B)$ is satisfied. Finally all the clauses in (6.2) are satisfied.
- Now assume the clauses in (6.2) are satisfied. We first show that $\{A \in Ar \mid \alpha(x'_A) = 1\} \subseteq \mathcal{A}rgs^-$. So consider an argument A with $\alpha(x'_A) = 1$. As the clause $(\neg x'_A \vee \bigvee_{(A,B)\in att} x_B)$ is satisfied we obtain that there exits an B with $(A, B) \in attack$ and $B \in \mathcal{A}rgs$. Thus $A \in \mathcal{A}rgs^-$.

  It remains to show that $\{A \in Ar \mid \alpha(x'_A) = 1\} \supseteq \mathcal{A}rgs^-$. To this end consider $A \in \mathcal{A}rgs^-$. There exists an B with $(A, B) \in attack$ and $B \in \mathcal{A}rgs$. Now as the clause $(x'_A \vee \neg x_B)$ is satisfied we obtain $\alpha(x'_A) = 1$.

To sum up, if α satisfies (6.1) and (6.2) then $\mathcal{A}rgs$ is an admissible set containing β such that $\alpha(x'_A) = 1$ iff $A \in \mathcal{A}rgs^-$.

Next consider the number of satisfied clauses c_α. If there is an admissible set containing β then there is also an assignment satisfying all clauses (6.1) and (6.2). As we have $|Ar| + 1$ copies of these clauses the optimal assignment then satisfied at least $m \cdot (|Ar| + 1)$ clauses. On the other side if an assignment misses one of the clause it misses all $(|Ar| + 1)$ copies and as there are only $|Ar|$ clauses in (6.3) such an assignment clearly satisfies less than $m \cdot (|Ar| + 1)$ clauses.

Hence, if the best assignment satisfies less than $m \cdot (|Ar| + 1)$ clauses then we know that that there is no minimal admissible set wrt. β. If the best assignment α satisfies at least $m \cdot (|Ar| + 1)$ clauses then we know that $\mathcal{A}rgs$ is an admissible set containing β. Now as all assignments corresponding to admissible sets containing β satisfy the clauses in (6.1) and (6.2), we know that α is the one satisfying a maximal number of clauses in (6.3). But as (6.2) are satisfied this is equivalent to minimizing the set $\mathcal{A}rgs^-$. Hence, $\mathcal{A}rgs$ is minimal admissible. The equation $|\mathcal{A}rgs^-| = |Ar| - (c_\alpha - m \cdot (|Ar| + 1))$ is immediate by the fact hat all clauses in (6.1) and (6.2) are satisfied. ■

We are now ready to state our theorems.

Theorem 6.21
The Minimum Discussion Length Problem is of the same complexity as MAX-SAT and thus is $FP^{\mathrm{NP}(\log \mathrm{n})}$-complete.

Proof. First we have the following reduction from MAX-SAT to the Minimum Discussion Length Problem. First build $\mathtt{AF}_\varphi$ from *Reduction* 6.14. Solving this returns $|\mathcal{A}rgs^-|$ for a minimal admissible set $\mathcal{A}rgs$. By Lemma 6.18 (2) we can then compute the number of simultaneously satisfiable clauses by $|C| - \left\lfloor \frac{|\mathcal{A}rgs^-| - |\mathcal{C}|}{|\mathcal{X}|+1} \right\rfloor$. Both the construction of $\mathtt{AF}_\varphi$ and the final calculation can be clearly performed in polynomial time.

Second we have a reduction from the Minimum Discussion Length Problem to MAX-SAT. First, build the formula $\varphi_{(Ar,att),\beta}$ from Reduction 6.19. Then compute c_α the number of clauses satisfied by a maximal satisfying assignment α. If $c_\alpha < m \cdot (|Ar| + 1)$, by Lemma 6.20, we know that there is no such set. Otherwise, by Lemma 6.20, we can easily compute the cardinality of $|\mathcal{A}rgs^-|$ using the number of satisfied clauses c_α as $|\mathcal{A}rgs^-| = |Ar| - (c_\alpha - m \cdot (|Ar| + 1))$. Again the construction of $\mathtt{AF}_\varphi$ and the final calculation can be performed in polynomial time. ■

Theorem 6.22
The Minimal Admissible Set Problem is of the same complexity as the MAX-SAT Assignment problem.

Proof. A reduction from MAX-SAT to the Minimum Discussion Length Problem: First build $\mathtt{AF}_\varphi$ from *Reduction* 6.14. Solving this returns a minimal admissible set $\mathcal{A}rgs$. By Lemma 6.18 (1) the truth assignment α, with $\alpha(X) = 1$ if $X \in \mathcal{A}rgs$ and $\alpha(X) = 0$ otherwise, satisfies the maximal number of clauses of φ. Both the construction of $\mathtt{AF}_\varphi$ and the construction of α can be clearly performed in polynomial time.

A reduction from Minimum Discussion Length Problem to MAX-SAT. First, build the formula $\varphi_{(Ar,att),\beta}$ from Reduction 6.19. Then compute a maximal satisfying assignment α. Finally construct $\mathcal{A}rgs = \{A \in Ar \mid \alpha(x_A) = 1\}$ which, by Lemma 6.20, is a minimal admissible set. Again the construction of $\mathtt{AF}_\varphi$ and the construction of $\mathcal{A}rgs$ can be performed in polynomial time. ∎

Let us briefly relate the complexity of computing a minimal discussion supporting an argument with computing an arbitrary discussion supporting an argument. As mentioned before the latter is equivalent to credulous reasoning with preferred semantics which lies on the NP-layer, i.e. credulous acceptance is NP-complete and computing an admissible set is in FNP (the function variant of NP), and thus is of the same complexity as SAT. Our results show that adding the requirement of minimality of discussion length increases complexity to the level of MAX-SAT.

Finally, the complexity results also guide a way to a possible implementation. One can use the encoding as propositional formula provided in Reduction 6.19 and then use one of the available maxsat-solvers [10] to compute a minimal admissible set for the considered argument.

7 Other Semantics

Preferred semantics is not the only semantics that can be expressed as a particular type of semi-natural discussion. We now briefly discuss three other semantics (stable, ideal and grounded) and their associated discussion games.

Stable Semantics

The question of how to express stable semantics as structured discussion has been treated by Caminada and Wu [14]. Basically, the idea is to take the discussion game for preferred semantics, as described in the current paper, and allow player S (who assumes the role of Socrates) to use one additional type of move: `question`.

To illustrate the role of the `question` move, consider again the argumentation framework of Figure 2 (page 12). Here, there are two preferred labellings:

- $\mathcal{L}ab_1$ with $\mathtt{in}(\mathcal{L}ab_1) = \{A\}$, $\mathtt{out}(\mathcal{L}ab_1) = \{B\}$ and $\mathtt{undec}(\mathcal{L}ab_1) = \{C, D, E\}$
- $\mathcal{L}ab_2$ with $\mathtt{in}(\mathcal{L}ab_2) = \{B, D\}$, $\mathtt{out}(\mathcal{L}ab_2) = \{A, C, E\}$ and $\mathtt{undec}(\mathcal{L}ab_2) = \emptyset$

Of these two preferred labellings, only $\mathcal{L}ab_2$ is also a stable labelling.[11] So although argument A is labelled `in` by a preferred labelling (A is an element of a preferred extension), A is not labelled `in` by any stable labelling (it is not an element of any stable extension).

To see why A is labelled `in` by at least one admissible labelling, consider the following discussion.

[10] See `http://maxsat.ia.udl.cat/solvers/` for an overview of maxsat-solvers.

[11] We recall that a stable labelling is an admissible labelling without any arguments that are labelled `undec`, just like a stable extension is an admissible set (or equivalently, a conflict-free set) that attacks each argument that is outside of it.

M: `in`(A)
"I have an admissible labelling where A is labelled `in`*."*
S: `out`(B)
"Then in your labelling, argument B must be labelled `out`*. Based on which grounds?"*
M: `in`(A)
"B is labelled `out` *because A is labelled* `in`*"*

The point is, however, that once it has been committed that A is labelled `in` and B is labelled `out`, it is not possible any more to label the remaining arguments such that final result is a stable labelling. This can be seen as follows. Suppose C is labelled `in`. Then E must be labelled `out`, so D should be labelled `in`, which means that C would be labelled `out`. Contradiction. Similarly, suppose that C is labelled `out`. Then E must be labelled `in`, so D should be labelled `out`, so C should be labelled `in`. Again, contradiction.

As is argued by Caminada and Wu [14], the discussion game for stable semantics requires an additional type of move: `question`. To illustrate the role of this new move, imagine a politician being interviewed for TV. At first, the discussion may be about financial matters (say, whether the banking system should be nationalized). Then, the discussion may be about the consequences of the politician's opinion ("If you accept to nationalize the banks, then you must reject the possibility to improve healthcare, because there will not be enough money left to do so."). However, at some moment, the interviewer could choose to totally change topic ("By the way, what are your opinions about abortion?"). It is this change of topic that is enabled by the `question` move.

For the discussion game for stable semantics, the idea is to use the `question` move to involve those arguments that have never been uttered before so that we can explicitly label all arguments of the argumentation framework. By questioning an argument (`question`(A)), the opponent asks the proponent to give an explicit opinion on whether A should be labelled `in` or `out`. If the proponent thinks that A should be labelled `in` then it should respond with `in`(A). If the proponent thinks that A should be labelled `out` then it should respond with `in`(B) where B is an attacker of A.

As an example of the use of the discussion game and the role of the `question` move, consider again the argumentation framework depicted in Figure 2. Suppose player M would like to start a discussion about A.

M: `in`(A)
"I have a stable labelling in which A is labelled `in`*."*
S: `out`(B)
"Then in your labelling, A's attacker B must be labelled `out`*. Based on which grounds?"*
M: `in`(A)
"B is labelled `out` *because A is labelled* `in`*."*
S: `question`(C)
"What about C?"
M: `in`(C)

"C is labelled `in`*."*

S: `out`(E)

"Then C's attacker E must be labelled `out`*. Based on which grounds?"*

M: `in`(D)

"E is labelled `out` *because D is labelled* `in`*."*

S: `out`(C)

"Then D's attacker C has to be labelled `out`*. This contradicts with your earlier claim that C is labelled* `in`*."*

Player S wins the discussion because he managed to point out a contradiction in the position of player M. A similar contradiction would have been pointed out if player M had responded to `question`(C) by `in`(E). Hence, once player M starts the discussion with `in`(A), he has no way of winning it any more when player S is allowed to use the `question` move. Hence, argument A is not in a stable extension.

For a fully formal account of the stable semantics discussion game, as well as for proofs of its correctness and completeness, we refer to Caminada and Wu's work [14]. For now, the main point we want to convey is that the type of discussion that is associated with stable semantics is essentially Socratic, with one little twist: the player who assumes the role of Socrates is allowed to change topic.

Ideal Semantics

An ideal set of arguments [17] is an admissible set that is a subset of each preferred extension.[12] It has been proven that the maximal ideal set is unique and is also a complete extension [17]. An alternative way to characterise an ideal set is as an admissible set that is not attacked by any other admissible set. This clears the way of expressing ideal semantics in terms of Socratic discussion. Basically, the discussion whether an argument (say A) is in an ideal set consists of two phases. In the first phase, one runs the standard Socratic discussion game, as is described in the current paper. This is to determine whether the argument is in an admissible set. Then, in the second phase of the discussion, one needs to determine whether this set is attacked by another admissible set. This is done by again running the discussion game for each of the arguments that were rejected (labelled `out`) during the first phase of the discussion, this time trying to defend (label `in`) the argument.

As an example, consider again the argumentation framework of Figure 2. Now consider the question of whether argument D is in an ideal set. The first phase of the discussion would be like Example 4.1. Then, in the second phase of the discussion, one has to try to find an argument that was labelled `out` during the first phase (say A) and can be defended in a new Socratic discussion game. Such a game would be as follows.

M: `in`(A)

"I have a reasonable position (admissible labelling) in which A is accepted (labelled `in`*)."*

S: `out`(B)

"Then in your position, argument B must be rejected (labelled `out`*). Based on which grounds?"*

[12] A treatment of ideal semantics in terms of argument labellings is given in [11].

M: `in`(A)
"B is rejected (labelled `out`*) because A is accepted (labelled* `in`*)."*

Hence, we have an admissible set (labelling) that attacks the admissible set (labelling) found during the first phase, so the admissible set (labelling) of the first phase is not an ideal set (labelling).[13]

The overall procedure for ideal semantics puts an extra burden on the proponent of the argument. Not only does he have to win the standard Socratic discussion game in the first phase, but he has to win it in such a way[14] that the resulting position (labelling) cannot be argued against in the second phase. Hence, the idea is to build a reasonable position that cannot be attacked by any reasonable position.

Grounded Semantics

Whereas the discussion games of stable and ideal semantics, as treated in the previous sections, are essentially based on the standard Socratic discussion game, the discussion game for grounded semantics appears to have a fundamentally different nature. Perhaps the best way of explaining this is from the perspective of complete labellings [6, 13]. We recall that, given an argumentation framework (Ar, att), a complete labelling is an argument labelling $\mathcal{L}ab$ such that for each argument $A \in Ar$ it holds that:

- if $\mathcal{L}ab(A) = \texttt{in}$ then $\forall B \in Ar : (B att A \supset \mathcal{L}ab(B) = \texttt{out})$
- if $\mathcal{L}ab(A) = \texttt{out}$ then $\exists B \in Ar : (B att A \wedge \mathcal{L}ab(B) = \texttt{in})$
- if $\mathcal{L}ab(A) = \texttt{undec}$ then $\neg\forall B \in Ar : (B att A \supset \mathcal{L}ab(B) = \texttt{out})$ and $\neg\exists B \in Ar : (B att A \wedge \mathcal{L}ab(B) = \texttt{in})$

Notice that the first two conditions are those of an admissible labelling. Hence, whereas an admissible labelling requires one to explain everything one accepts (because all attackers are rejected) and everything one rejects (because there is an attacker that is accepted), a complete labelling also requires one to explain everything one abstains from having an explicit opinion about (because there are insufficient reasons to accept it, and insufficient reasons to reject it). Hence, the overall idea of a complete labelling is that one has to be able to explain everything one accepts, everything one rejects and everything one abstains from having an explicit opinion about.

If one regards a complete labelling as a reasonable position one can take in the presence of the conflicting information represented in the argumentation framework, then two questions become relevant:
(1) Is there *at least one* reasonable position (complete labelling) in which the argument is accepted (labelled `in`) ?
(2) Does the argument have to be accepted (labelled `in`) in *every* reasonable position[15]

[13] In fact, for the argumentation framework of Figure 2, the only ideal set is the empty set.

[14] Since an argument can be element of more than one admissible set, there can be different ways to win the Socratic discussion game.

[15] The question "does the argument have to be accepted in every reasonable position" becomes trivial when we equate a reasonable position with an admissible labelling, since it is always possible to construct an admissible labelling by abstaining on everything (labelling each argument `undec`). Hence, one needs an additional clause that puts restrictions on when one is allowed to abstain. Such a restriction is provided by the concept of a complete labelling (third clause).

(complete labelling) ?
Questions (1) and (2) are fundamentally different, also from the perspective of formal discussion. To resolve whether there exists at least one reasonable position in which a particular argument is accepted, one can have one of the parties adopt the argument and the other party critically questioning whether the resulting position is in fact reasonable. To resolve whether an argument has to be accepted in *every* reasonable position, on the other hand, one can try to convince a sceptical but rational party that he cannot avoid accepting the argument. Hence, whereas the discussion around question (1) is of the form "I'm being reasonable when I accept that...", the discussion around question (2) is of the form "If you are being reasonable then you have to accept that..."

As we have seen, the discussion around question (1) is essentially of a Socratic nature, where the proponent (player M) claims that his position is reasonable, whereas the opponent (player S) tries to show that the position is *not* reasonable, by leading proponent to refutation. Standard argumentation theory states that an argument is labelled `in` by at least one complete labelling iff it is labelled `in` by at least one admissible labelling [13]. Therefore, to determine whether an argument is labelled `in` by at least one complete labelling, one can simply run the standard Socratic discussion game as was described in Section 4.

As for question (2), it has to be observed that the discussion around the issue "If you are being reasonable then you have to accept that..." is of a fundamentally different form. Instead of merely having to avoid being refuted, the proponent faces the more challenging task of actually *convincing* the opponent to accept the argument in question. In terms of the typology of Walton and Krabbe [51], such a discussion would be regarded as *persuasion*. The idea is to persuade the opponent that, by adopting the rationality conditions of a complete labelling, the opponent also has to accept the argument in question, even when being maximally sceptical.

As an example of how such a discussion could take place, consider an argumentation framework with arguments A, B and C, where A attacks B and B attacks C.

Prop: `in`(C)
"C has to be accepted in every reasonable position, therefore also in your position"

Opp: `out`(B)?
"Are you sure? Perhaps there is a reasonable position in which C's attacker B is not rejected. Why does B always have to be rejected?"

Prop: `in`(A)
"B has to be rejected because its attacker A has to be accepted, since A doesn't have any attackers itself."

After the third move, the opponent has to admit (concede) that A has to be accepted, and that therefore B has to be rejected and A accepted. Hence, the proponent wins the discussion game.

As an example of a discussion that the proponent is unable to win, consider an argumentation framework with two arguments, A and B, where A attacks B, and B attacks A.

Prop: in(A)
"A has to be accepted in every reasonable position, therefore also in your position."

Opp: out(B)?
"Are you sure? Perhaps there is a reasonable position in which A's attacker B is not rejected. Why does B always have to be rejected?"

After the opponent's move, the proponent could of course reply with in(A) again, but this would mean the discussion could go on perpetually, without the proponent ever convincing the opponent.[16] Hence, it is not possible for the proponent to convince the opponent. This result is different than for the Socratic discussion game, in which the proponent (player M) wins after just three moves. This illustrates that, at least in abstract argumentation theory, the treshold is lower for maintaining one's own position is merely reasonable, rather than for actually convincing the other party of one's own position.

Standard argumentation theory says that an argument is labelled in by every complete labelling iff it is labelled in by the grounded labelling [6, 13].[17] Therefore, we can use the grounded game as defined by many researchers [42, 5, 33] to determine whether an argument is labelled in by every complete labelling. In fact, the two example discussions above are actually instances of this discussion game. In general, just like the preferred game can be interpreted as a particular form of Socratic discussion, the grounded game can be interpreted as a particular form of persuasion discussion.

Apart from the conceptual difference, there also exists an important technical difference between the preferred game and the grounded game. Whereas for the preferred game, it is sufficient that there exists at least one preferred game won by the proponent (player M) for an argument to be in a preferred extension (or equivalently, in an admissible set or a complete extension), for the grounded game the existence of a single game won by the proponent is not enough. In fact, the proponent needs to have a *winning strategy* for the grounded game, in order for the argument to be in the grounded extension (or equivalently, to be labelled in by every complete labelling). Conceptually, this is a bit odd, since whether or not one persuades the other party should depend on the actual discussion only, and not on the discussions that could have been taken place. Hence, an interesting research question is whether it is possible to reformulate the grounded game in such a way that a single discussion is sufficient.[18] Ideally, the precise rules of such a game should naturally follow from the nature of a persuasion dialogue, just like the rules of the preferred game can easily be explained by examining the concept of Socratic dialogue ("Don't ask the same question twice", "It's possible for different questions to have the same answer", etc). A paper that addresses these issues has recently been published [3].

[16] This is why in the discussion game [42, 5, 33] the proponent is disallowed to repeat his moves.

[17] We recall that the grounded labelling is the (unique) complete labelling whose set of in-labelled arguments is minimal (w.r.t. set-inclusion) among all complete labellings [6, 13], just like the grounded extension is the (unique) minimal complete extension [18]. We also recall that the set of in-labelled arguments of the grounded labelling is the grounded extension [6, 13].

[18] It has to be mentioned that the original preferred game [49] also requires a winning strategy, and that it were Caminada and Wu [14] who reformulated this game such that this requirement could be dropped.

8 Discussion

In the current paper, we contributed to bridging the gap between formal Dung-style argumentation theory and the kind of informal discussion that one observes in conversational arguments. To the newcomer in abstract argumentation theory, it may sometimes appear that the field is mainly about things like fixpoints and graph theory. However, as we have pointed out in the current paper, it is very well possible to reinterpret abstract argumentation theory in the form of semi-natural discussion. Whereas classical logic is concerned mainly with what is *true*[19], argumentation theory, in this view, is about what can be defended in rational discussion. Different argumentation semantics, in essence, represent different ideas about the nature of rational discussion and their associated proof standards. An overview is presented in Table 1.

Semantics	type of discussion (credulous acceptance)
preferred, complete and admissible	Socratic discussion
stable	Socratic discussion in which Socrates can change topic
ideal	Socratic discussion whose results cannot be argued against in another Socratic discussion
grounded	persuasion discussion

Table 1. Argumentation semantics and their associated discussion games

The connection between abstract argumentation theory and discussion games is relevant for more than just theoretical reasons. For an argumentation-based expert system, it can be hugely beneficial if it were able to explain its answers not in terms of monotonic functions and fixpoints, but by means of (semi-) natural discussion that the user is intuitively already familiar with. After all, Socratic dialogue, as we mentioned earlier, goes back to classical antiquity, and is essentially still in use for purposes like legal cross-examination and critical interviews, and similar observations can be made regarding persuasion dialogue. Ideally, if an argumentation-based expert system provides an answer that is not immediately understood by the user, then the user should be able to do the same as when disagreeing with another person: start a discussion.[20]

Funding

For Martin Caminada, the major part of the work on this paper was carried out while he was affiliated with the Interdisciplinary Centre for Security, Reliability and Trust at the University of Luxembourg. His research was supported by the National

[19] It must be mentioned, however, that even for classical logic, dialectical interpretations do exist (see for instance the work of Lorenzen and Lorenz [30]). However, these never gained as much popularity as the traditional model-based semantics.

[20] There are a few other open issues that would still need to be dealt with in order for argumentation to be truly beneficial for the purpose of logical inference. We refer to the paper of Caminada and Wu [4] for an overview.

Research Fund, Luxembourg (LAAMI project), as well as by the Engineering and Physical Sciences Research Council (EPSRC, UK), grant ref. EP/J012084/1 (SAsSY project).

For Srdjan Vesic, the major part of the work on this paper was carried out while he was affiliated with the Computer Science and Communication Research Unit at the University of Luxembourg. He worked on this paper during the tenure of an ERCIM Alain Bensoussan Fellowship Programme, which is supported by the Marie Curie Co-funding of Regional, National and International Programmes (COFUND) of the European Commission. During this time, he was also funded by the National Research Fund, Luxembourg. When finishing the work on this paper, he was a CRNS researcher affiliated with CRIL.

Acknowledgements

We would like to thank Douglas Walton, Erik Krabbe and Richard Booth for their useful comments.

References

[1] ASPIC-consortium. Deliverable D2.5: Draft formal semantics for ASPIC system, June 2005.

[2] Ph Besnard and S. Doutre. Checking the acceptability of a set of arguments. In J.P. Delgrande and T. Schaub, editors, *Proc. 10th International Workshop on Non-Monotonic Reasoning (NMR04)*, pages 59–64, 2004.

[3] M. Caminada and M. Podlaszewski. A persuasion dialogue for grounded semantics. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA'12)*, pages 105–116, 2012.

[4] M. Caminada and Y. Wu. On the limitations of abstract argumentation. In Patrick de Causmaecker, Joris Maervoet, Tommy Messelis, Katja Verbeeck, and Tim Vermeulen, editors, *Proceedings of the 23rd Benelux Conference on Artificial Intelligence (BNAIC 2011)*, pages 59–66, 2011.

[5] M.W.A. Caminada. *For the sake of the Argument. Explorations into argument-based reasoning.* Doctoral dissertation Free University Amsterdam, 2004.

[6] M.W.A. Caminada. On the issue of reinstatement in argumentation. In M. Fischer, W. van der Hoek, B. Konev, and A. Lisitsa, editors, *Logics in Artificial Intelligence; 10th European Conference, JELIA 2006*, pages 111–123. Springer, 2006. LNAI 4160.

[7] M.W.A. Caminada. Semi-stable semantics. In P.E. Dunne and TJ.M. Bench-Capon, editors, *Computational Models of Argument; Proceedings of COMMA 2006*, pages 121–130. IOS Press, 2006.

[8] M.W.A. Caminada. Comparing two unique extension semantics for formal argumentation: ideal and eager. In Mohammad Mehdi Dastani and Edwin de Jong, editors, *Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2007)*, pages 81–87, 2007.

[9] M.W.A. Caminada. A formal account of socratic-style argumentation. *Journal of Applied Logic*, 6(1):109–132, 2008.

[10] M.W.A. Caminada. An algorithm for stage semantics. In M. Giacomin P. Baroni, F. Cerutti and G.R. Simari, editors, *Proceedings of the Third International Conference on Computational Models of Argument (COMMA 2010)*, pages 147–158. IOS Press, 2010.

[11] M.W.A. Caminada. A labelling approach for ideal and stage semantics. *Argument & Computation*, 2:1–21, 2011.

[12] M.W.A. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

[13] M.W.A. Caminada and D.M. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109–145, 2009. Special issue: new ideas in argumentation theory.

[14] M.W.A. Caminada and Y. Wu. An argument game of stable semantics. *Logic Journal of IGPL*, 17(1):77–90, 2009.

[15] P. Crescenzi and L. Trevisan. Max np-completeness made easy. *Theor. Comput. Sci.*, 225(1-2):65–79, 1999.

[16] Y. Dimopoulos and A Torres. Graph theoretical structures in logic programs and default theories. *Theoretical Computer Science*, 170:209–244, 1996.

[17] P. M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15):642–674, 2007.

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

[19] P. E. Dunne and T. J. M. Bench-Capon. Coherence in Finite Argument Systems. *Artificial Intelligence*, 141(1–2):187–203, October 2002.

[20] P. E. Dunne, Sylvie Doutre, and Trevor Bench-Capon. Discovering inconsistencies through examination dialogues. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

[21] P. E. Dunne and M. Wooldridge. Complexity of abstract argumentation. In I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 85–104. Springer, Berlin, 2009.

[22] W. Dvořák, M. Järvisalo, J. P. Wallner, and S. Woltran. Complexity-sensitive decision procedures for abstract argumentation. *Artificial Intelligence*, 206(0):53 – 78, 2014.

[23] C. L. Hamblin. *Fallacies.* Methuen, London, UK, 1970.

[24] C. L. Hamblin. Mathematical models of dialogue. *Theoria*, 37:130–155, 1971.

[25] H. Jakobovits and D. Vermeir. Robust semantics for argumentation frameworks. *Journal of logic and computation*, 9(2):215–261, 1999.

[26] E. Krabbe. Formal dialectics as immanent criticism of philosophical systems. In E.M. Barth and J.L. Martens, editors, *Argumentation Approaches to Theory Formation*, pages 233–243, Amsterdam, 1982. John Benjamins.

[27] E. Krabbe. *Studies in Dialogical Logic.* PhD thesis, Rijksuniversiteit Groningen, 1982.

[28] M. W. Krentel. The complexity of optimization problems. In Juris Hartmanis, editor, *Proceedings of the 18th Annual ACM Symposium on Theory of Computing, May 28-30, 1986, Berkeley, California, USA*, pages 69–76. ACM, 1986.

[29] K. Lorenz. Die dialogische rechtfertigung der effectiven logic. In F. Kambartel and J. Mittelstrass, editors, *Zum normativen Fundament der Wissenschaft*, pages 250–280, Frankfurt am Main, 1973.

[30] P. Lorenzen and K. Lorenz. Dialogische logik. *Wissenschaftliche Buchgesellschaft, Darmstadt*, 1978.

[31] J. D. Mackenzie. Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.

[32] J. D. Mackenzie. Four dialogue systems. *Studia Logica*, 51:567–583, 1990.

[33] S. Modgil and M.W.A. Caminada. Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G.R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 105–129. Springer, 2009.

[34] L. Nelson. *De Socratische Methode.* Uitgeverij Boom, Amsterdam, 1994.

[35] C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *J. Comput. Syst. Sci.*, 43(3):425–440, 1991.

[36] Chaim Perelman. *The Realm of Rhetoric.* University of Notre Dame Press, Notre Dame, Indiana, 1982. translated by William Kluback.

[37] Plato. Lysis. In E. Rhys, editor, *Socratic Discourses by Plato and Xenophon.* J.M. Dent & Sons ltd., London, 1910.

[38] Plato. Sophist, 360BC. translated by Benjamin Jowett.

[39] J.L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.

[40] J.L. Pollock. *Cognitive Carpentry. A Blueprint for How to Build a Person.* MIT Press, Cambridge, MA, 1995.

[41] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[42] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7:25–75, 1997.

[43] R. Robinson. *Plato's Earlier Dialectic*. Oxford University Press, Oxford, 1962.

[44] G.R. Simari and R.P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53:125–157, 1992.

[45] J. Skidmore. Skepticism about practical reasoning: transcendental arguments and their limits. *Philosophical Studies*, 109:121–141, 2002.

[46] B. Verheij. Two approaches to dialectical argumentation: admissible sets and argumentation stages. In J.-J.Ch. Meyer and L.C. van der Gaag, editors, *Proceedings of the Eighth Dutch Conference on Artificial Intelligence (NAIC'96)*, pages 357–368, 1996.

[47] G.A.W. Vreeswijk. Studies in defeasible argumentation. *PhD thesis at Free University of Amsterdam*, 1993.

[48] G.A.W. Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90:225–279, 1997.

[49] G.A.W. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proceedings of the 7th European Workshop on Logic for Artificial Intelligence (JELIA-00)*, number 1919 in Springer Lecture Notes in AI, pages 239–253, Berlin, 2000. Springer Verlag.

[50] D. Walton. *Dialog Theory for Critical Argumentation*. John Benjamins Publishing Company, Amsterdam, 2007.

[51] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Series in Logic and Language. State University of New York Press, Albany, NY, USA, 1995.

[52] Y. Wu, M.W.A. Caminada, and D.M. Gabbay. Complete extensions in argumentation coincide with 3-valued stable models in logic programming. *Studia Logica*, 93(1-2):383–403, 2009. Special issue: new ideas in argumentation theory.