Preferred Semantics as Socratic Discussion

Martin Caminada
University of Luxembourg
martin.caminada@uni.lu

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.

Introduction

The field of formal argumentation can be traced back to the work of Pollock (Pollock 1992; 1995), Vreeswijk (Vreeswijk 1993; 1997), and Simari and Loui (Simari and Loui 1992). 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 counter arguments, that can then themselves be attacked by counter arguments, etc.

Nowadays, much research on the topic of argumentation is based on the abstract argumentation theory of Dung (Dung 1995). 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 (Dung 1995), as well as subsequently stated approaches such as *stage* (Verheij 1996; Caminada 2010), *semi-stable* (Verheij 1996; Caminada 2006b), *ideal* (Dung, Mancarella, and Toni 2007) and *eager* semantics (Caminada 2007). 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.

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 semantics 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 lead 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 we provide an overview of the concept of preferred semantics, as it has been treated in the literature of formal argumentation. Then we provide a semi-formal analysis of Socratic discussion, based on the work of (Caminada 2004; 2008). 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. We then round off with a discussion of how similar approaches can be applied to other semantics.

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 (Dung 1995) and Caminada *et al's* reformulation of preferred semantics in terms of argument labellings (Caminada 2006a; Caminada and Gabbay 2009).

Definition 1. *An* argumentation framework *is a pair* (Ar, att) *where* Ar *is a set of arguments and* $att \subseteq Ar \times Ar$.

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.[1]

Definition 2. *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 3. *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 4. *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, like (Verheij 1996; Jakobovits and Vermeir 1999; Pollock 1995), also takes into account the arguments that are *rejected*. In this paper, we will use the particular labellings approach of Caminada (Caminada 2006a; Caminada and Gabbay 2009), 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 5. *Let (Ar, att) be an argumentation framework. A* labelling *is a (total) function $\mathcal{L}ab : Ar \longrightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$.*

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 6. *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 7. *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 in (Caminada and Gabbay 2009) 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 (Caminada and Gabbay 2009).

Socratic Argumentation

Although Dung's theory allows the internal structure of an argument to remain completely abstract, many formalisms of argumentation (such as (Vreeswijk 1993), (Caminada and Amgoud 2007), (Wu, Caminada, and Gabbay 2009) and (Prakken 2010)) 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 (Prakken 2010).

[1] 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.

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.

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 is committed by the proponent. 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 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 (Plato 1910, 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 he 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.In "The Sophist", Plato provides the following definition of the elenchus (Plato 360 BC):

> 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 the Belgian scholar Chaïm Perelman (Perelman 1982, 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 counterargument, but there are also forms of incompatibility that require argumentation beyond classical arguments.

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.

In general, one may say that many of today's interviews in which 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.

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." (Skidmore 2002, 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.

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 1979). 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 formal dialogue, consider the following example, where the proponent (P) argues that there will be a tax relief (tr) because some leading politicians made the promise to do so (pmp).

Example 1.

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

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 (Walton and Krabbe 1995).

In the above dialogue the opponent concedes the main claim, so the proponent wins the dialogue. If, during the cause 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 can lead to a fine from the EU (feu), therefore ruling out the possibility of any durable tax relief.

Example 2.

P:	*claim* tr	$C_P(\mathtt{tr})$
	"*I think that* tr."	
O:	*but-then* tr $\Rightarrow$ bd	$C_O(C_P(\mathtt{bd}))$
	"*Then you implicitly also hold that* bd."	
P:	*concede* bd	$C_P(\mathtt{tr}, \mathtt{bd})$
	"*Yes I do.*"	
O:	*but-then* bd $\Rightarrow$ feu	$C_O(C_P(\mathtt{feu}))$[2]
	"*Then you implicitly also hold that* feu."	
P:	*concede* feu	$C_P(\mathtt{tr}, \mathtt{bd}, \mathtt{feu})$
	"*Yes I do.*"	
O:	*but-then* feu $\Rightarrow \neg$tr	$C_O(C_P(\neg\mathtt{tr}))$
	"*Then you implicitly also hold that* $\neg$tr."	
P:	*concede* $\neg$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 tr then you actually also endorse 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 bd himself, so it does not hold that $C_O(\mathtt{bd})$. Furthermore, it goes too far to immediately have P committed to bd; the rule "tr $\Rightarrow$ bd" is defeasible and P may defend himself by giving a reason (an undercutter) why this

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

rule does not apply (an example of this will be treated further on). Therefore, it also does not hold that $C_P(\texttt{bd})$. The only thing that can be said regarding the but-then statement is that O claims the bd is implicitly endorsed by P. Therefore, it holds that $C_O(C_P(\texttt{bd}))$.

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).
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.

Figure 1: because and 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 $\texttt{bd} \Rightarrow \texttt{feu}$ can now be undercut.

Example 3.

P:	*claim* tr	$C_P(\texttt{tr})$
O:	*but-then* $\texttt{tr} \Rightarrow \texttt{bd}$	$C_O(C_P(\texttt{bd}))$
P:	*concede* bd	$C_P(\texttt{tr}, \texttt{bd})$
O:	*but-then* $\texttt{bd} \Rightarrow \texttt{feu}$	$C_O(C_P(\texttt{feu}))$
P:	*claim* $\neg\lceil\texttt{bd} \Rightarrow \texttt{feu}\rceil$	$C_P(\texttt{tr}, \texttt{bd}, \neg\lceil\texttt{bd} \Rightarrow \texttt{feu}\rceil)$
O:	*why* $\neg\lceil\texttt{bd} \Rightarrow \texttt{feu}\rceil$	$C_O(C_P(\texttt{feu}))$
P:	*because* $\texttt{fc} \Rightarrow \neg\lceil\texttt{bd} \Rightarrow \texttt{feu}\rceil$	$C_P(\texttt{tr}, \texttt{bd}, \neg\lceil\texttt{bd} \Rightarrow \texttt{feu}\rceil, \texttt{fc})$
O:	*retract* $C_P(\texttt{feu})$, *concede* tr	$C_O(\texttt{tr})$

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 such 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 loose the dialogue game.

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 2000), who defined a formal argument game to decide this. A somewhat similar game has subsequently been specified by (Modgil and Caminada 2009). Our aim is not so much to provide an entirely new approach, rather to reinterpret the existing work of (Vreeswijk and Prakken 2000) in the context of Socratic discussion.

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 (Caminada and Gabbay 2009). 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 2000), 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.

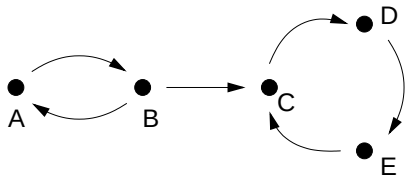

Figure 2: An argumentation framework

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

Example 4.

M: $\texttt{in}(D)$
"I have an admissible labelling in which D is labelled `in`*."*

S: $\texttt{out}(C)$
"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." 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)

At the other hand, Example 4 does illustrate the need for player M to be able to repeat his moves (like $\texttt{in}(B)$). 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 5.

M: $\texttt{in}(E)$
"I have an admissible labelling in which E is labelled `in`*."*

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

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

S: $\texttt{out}(E)$
"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.

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

Similarly, one might examine what happens when it is player S who cannot make a move anymore. This essentially means that player S has ran 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 anymore, player M wins the discussion game. (8)

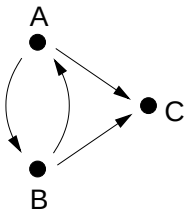

Figure 3: An argumentation framework with floating attack

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 6.

M: $\texttt{in}(C)$
"I have an admissible labelling in which C is labelled `in`*."*

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

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

S: $\texttt{out}(B)$
"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, $\texttt{out}(B)$ is a response to $\texttt{in}(C)$. 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 8. *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 Δ_i ($1 \leq i \leq n$) where i is odd (which is called an M-move) is of the form $\texttt{in}(A)$, where $A \in Ar$*
- *each Δ_i ($1 \leq i \leq n$) where i is even (which is called an S-move) is of the form $\texttt{out}(A)$, where $A \in Ar$*
- *for each S-move $\Delta_i = \texttt{out}(A)$ ($2 \leq i \leq n$) there exists an M-move $\Delta_j = \texttt{in}(B)$ ($j < i$) such that A attacks B*
- *for each M-move $\Delta_i = \texttt{in}(A)$ ($3 \leq i \leq n$) it holds that Δ_{i-1} is of the form $\texttt{out}(B)$, 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).*

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

Theorem 1 ((Vreeswijk and Prakken 2000; Caminada and Wu 2009))**.** *Let (Ar, att) be an argumentation framework and $A \in Ar$. There exists an admissible labelling $\mathcal{L}$ with $\mathcal{L}(A) = \texttt{in}$ iff there exists an admissible discussion for A that is won by player M.*

Theorem 1, 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. Hence, we have accomplished our goal of explaining (credulous) preferred semantics in terms of Socratic discussion.

Discussion

The approach of describing a particular argumentation semantics by means of Socratic discussion is not limited to

preferred semantics. A slightly altered version of the here described discussion game can also be applied in the context of stable semantics. Basically, the idea is to give player S the freedom also to ask player M for his opinion on arguments that are not directly related to those previously mentioned by player M (Caminada and Wu 2009). As for grounded semantics, however, the situation is fundamentally different. Since an argument is in the grounded extension iff it is labelled `in` by every complete labelling (Caminada and Gabbay 2009) the aim of the game seems to convince a sceptical discussion partner that he has no other choice than to accept the argument to be labelled `in`, also in his own (complete) labelling. Hence, the grounded discussion game (Prakken and Sartor 1997; Caminada 2004; Modgil and Caminada 2009) appears to resemble a more "traditional" persuasion discussion than to resemble the Socratic type of discussion treated in the current paper.

References

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

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

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

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

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

Caminada, M. 2006b. Semi-stable semantics. In Dunne, P., and Bench-Capon, T., eds., *Computational Models of Argument; Proceedings of COMMA 2006*, 121–130. IOS Press.

Caminada, M. 2007. Comparing two unique extension semantics for formal argumentation: ideal and eager. In Dastani, M. M., and de Jong, E., eds., *Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC 2007)*, 81–87.

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

Caminada, M. 2010. An algorithm for stage semantics. In *Proceedings of the Third International Conference on Computational Models of Argument (COMMA 2010)*. (in print).

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

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

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

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

Modgil, S., and Caminada, M. 2009. Proof theories and algorithms for abstract argumentation frameworks. In Rahwan, I., and Simari, G., eds., *Argumentation in Artificial Intelligence*. Springer. 105–129.

Perelman, C. 1982. *The Realm of Rhetoric*. Notre Dame, Indiana: University of Notre Dame Press. translated by William Kluback.

Plato. 1910. Lysis. In Rhys, E., ed., *Socratic Discourses by Plato and Xenophon*. London: J.M. Dent & Sons ltd.

Plato. 360 BC. Sophist. translated by Benjamin Jowett.

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

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

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

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

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

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

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

Vreeswijk, G., and Prakken, H. 2000. 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, 239–253. Berlin: Springer Verlag.

Vreeswijk, G. 1993. Studies in defeasible argumentation. *PhD thesis at Free University of Amsterdam*.

Vreeswijk, G. 1997. Abstract argumentation systems. *Artificial Intelligence* 90:225–279.

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

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