Distinguishability in Abstract Argumentation

Isabelle Kuhlmann, Tjitze Rienstra, Lars Bengel, Kenneth Skiba, Matthias Thimm
University of Koblenz-Landau
Universitätsstraße 1
56072 Koblenz, Germany
{iskuhlmann, rienstra, larsbengel, kennethskiba, thimm}@uni-koblenz.de

Abstract

In abstract argumentation, the admissible semantics can be said to *distinguish* the preferred semantics in the sense that argumentation frameworks with the same admissible extensions also have the same preferred extensions. In this paper we present an exhaustive study of such distinguishability relationships, including those between sets of semantics. We further examine restricted classes of argumentation frameworks, such as self-attack-free and acyclic frameworks. We discuss the relevance of our results in the context of the *argumentation framework elicitation* problem.

1 Introduction

The goal of this paper is to study the notion of *distinguishability* in abstract argumentation. Given a semantics σ, two argumentation frameworks (AFs, for short) are σ-*indistinguishable* if they possess the same σ-extensions. Indistinguishability can be contrasted with *strong equivalence*: two AFs F and G are strongly equivalent under semantics σ if $F \uplus H$ and $G \uplus H$ possess the same σ-extensions *for every possible* AF H, where $\uplus$ represents the operation of merging two AFs (Oikarinen and Woltran 2011). While strong equivalence under a wide range of semantics can be nicely characterized in terms of syntactic criteria called *kernels*, indistinguishability is much harder to characterize and has received much less attention in the literature.

Our study is part of, and motivated by, our ongoing research into the *argumentation framework elicitation* problem (Kuhlmann 2021). Consider a scenario where we are faced with an agent who possesses an AF $F = (\mathsf{Arg}, R)$. The goal of AF elicitation is to discover this AF by asking the agent a number of questions. AF elicitation can be compared to *preference elicitation*, which focuses on the problem of uncovering an agent's preferences on the basis of how the agent chooses between alternatives (see (Chen and Pu 2004) for an overview). It is also related to the problem of learning AFs from sets of extensions or labelings (Riveret and Governatori 2016; Niskanen, Wallner, and Järvisalo 2019; Kido and Liao 2019). The difference is, however, that elicitation also involves choosing the questions to ask, whereas the problem of learning AFs assumes that the set of extensions or labelings is given.

Let us state the AF elicitation problem more precisely. We make the simplifying assumption that the set Arg of the agent's AF $F = (\mathsf{Arg}, R)$ is known, so that we are actually eliciting the *syntactic* structure of the attack relation R. We do so by asking questions on the *semantics* of the framework such as "Is E a σ-extension of your AF?" where $E \subseteq \mathsf{Arg}$ and σ is a semantics. The agent has what we call a *type*, which is a set X of semantics from which σ may be chosen. Thus, an agent with type X will truthfully answer, for all $E \subseteq \mathsf{Arg}$ and $\sigma \in X$, whether E is a σ-extension of her AF. For an agent with a given type, the overall goal of the elicitation task is to discover the agent's AF F by asking a minimal number of questions.

The problem is that, due to indistinguishability this goal may actually be unachievable. Consider an agent with type $\{\mathsf{co}\}$ where co stands for the complete semantics. The best we can do is discover the co-*equivalence class* to which the agent's AF belongs, which is the set of all AFs possessing the same co-extensions. We then know that the agent's actual AF is a member of this set, but we have no way to know which one it is. This raises the general question: what exactly can we discover for a given agent type? More precisely: what is the relationship between an agent's type and the equivalence class that we can discover? The goal of this paper is to provide an exhaustive answer to this question, with a focus on the conflict-free, admissible, complete, grounded, preferred, stable, and semi-stable semantics. We also consider the case where we know that the AF satisfies properties such as not having (odd-length) cycles or self-attacks which, as we will show, puts this question into a different light.

The results we obtain will provide insights into the limits of what is possible in the context of AF elicitation. Our findings complement the work on *realizability* of AFs (Dunne et al. 2015, 2016; Pührer 2020) which explores whether a set of arguments can be *realized* as an extension of an AF under a given semantics. Another related problem is presented in a study conducted by Baroni and Giacomin on *agreement* of semantics, i.e. under which conditions a set of extensions is accepted under multiple semantics (Baroni and Giacomin 2008).

The remainder of this paper is structured as follows. Section 2 provides some fundamental definitions that our work is based on. In Section 3 we introduce the theoretical groundwork as well as our results. We conclude in Section 4.

2 Preliminaries

An *argumentation framework* is a pair $F = (\mathsf{Arg}, R)$ where Arg is a finite set of arguments and R is a relation $R \subseteq \mathsf{Arg} \times \mathsf{Arg}$ (Dung 1995). An argument a is said to *attack* an argument b if $(a, b) \in R$. We say that an argument a is *defended by a set* $E \subseteq \mathsf{Arg}$ if every argument $b \in \mathsf{Arg}$ that attacks a is attacked by some $c \in E$. For $a \in \mathsf{Arg}$ we define $a^- = \{b \mid (b, a) \in R\}$ and $a^+ = \{b \mid (a, b) \in R\}$, i. e., the sets of attackers of a and the set of arguments attacked by a. For a set of arguments $E \subseteq \mathsf{Arg}$ we extend these sets by defining E^+_F and E^-_F via $E^+_F = \bigcup_{a \in E} a^+$ and $E^-_F = \bigcup_{a \in E} a^-$, respectively.

A *semantics* determines a set of *extensions* (i. e., jointly acceptable sets of arguments) of an AF. Given an AF $F = (\mathsf{Arg}, R)$, an extension $E \subseteq \mathsf{Arg}$ is called *admissible* (ad) if and only if

1. E is *conflict-free* (cf), i. e., there are no arguments $a, b \in E$ with $(a, b) \in R$, and
2. E *defends* every $a \in E$,

and is called *complete* (co) if, additionally, it satisfies

3. if E defends a then $a \in E$.

By imposing more constraints on complete extensions we can define additional types of semantics. In particular, a complete extension E is *grounded* (gr) if and only if E is $\subseteq$-minimal; is *preferred* (pr) if and only if E is $\subseteq$-maximal; is *stable* (st) if and only if $\mathsf{Arg} = E \cup E^+$; and is *semi-stable* (sst) if and only if $E \cup E^+$ is $\subseteq$-maximal among all complete extensions (Dung 1995; Caminada, Carnielli, and Dunne 2012). Note that the grounded extension is uniquely determined and that stable extensions may not exist. We define $\Sigma = \{\mathsf{cf}, \mathsf{ad}, \mathsf{co}, \mathsf{gr}, \mathsf{pr}, \mathsf{st}, \mathsf{sst}\}$ to be the set of semantics considered in this work.[1] Given $\sigma \in \Sigma$ and an AF F, we refer to the set of all σ-extensions as $\sigma(F)$. The set of all AFs is denoted $\mathcal{F}_{all}$.

3 Distinguishability

Based on the previously described AF elicitation problem we propose our theoretical groundwork. In particular, we introduce the notion of distinguishability in order to articulate whether knowing the set of extensions w.r.t. a set of semantics X also yields information regarding some other semantics $\sigma \notin X$. We first need to define X*-equivalence* between AFs, with $X \subseteq \Sigma$.

Definition 1. *Let $X \subseteq \Sigma$. Two AFs F and F' are X-equivalent if, for all $\sigma \in X$, $\sigma(F) = \sigma(F')$.*

If two AFs are $\{\sigma\}$-equivalent we will simply say that they are σ-equivalent. In the elicitation scenario we are faced with an agent with a given type $X \subseteq \Sigma$. What we are interested in is: what information can we actually gain from this agent? We formalize this using the *distinguishablity* relation defined as follows.

Definition 2. *Let $X \subseteq \Sigma$, $\sigma \in \Sigma$, and let $\mathcal{F}$ be a set of AFs. We say that X* distinguishes σ in $\mathcal{F}$ *if and only if, for every pair $F, F' \in \mathcal{F}$, if F and F' are X-equivalent then F and F' are σ-equivalent.*

[1] To simplify presentation, we also treat cf and ad as semantics.

If X distinguishes σ in $\mathcal{F}_{all}$ then any two AFs with different σ extensions also have different σ' extensions for some $\sigma' \in X$. Hence, the information we can gain from an agent with type X is the σ-equivalence class—or some member of this class—of the agent's AF. To see that this is possible, consider the following simple procedure. We assume that the set Arg of arguments is known.

1. For every $E \subseteq \mathsf{Arg}$ and $\sigma \in X$ ask the agent whether E is a σ-extension of her AF and remember the answers.
2. Find an attack relation $R \subseteq \mathsf{Arg} \times \mathsf{Arg}$, such that the AF $F = (\mathsf{Arg}, R)$ agrees with the answers obtained in step 1. Because the agent is truthful, this attack relation is guaranteed to exist.

Clearly, if X distinguishes σ then the AF $F = (\mathsf{Arg}, R)$ obtained using this procedure belongs to the σ-equivalence class of the agent's AF. This procedure is clearly not efficient, since it requires asking an exponential number of questions and trying an exponential number of attack relations. The question of *how* to elicit an AF in an efficient manner will be addressed in future work, however. In this work we will only focus on distinguishability, which tells us about *what* we can elicit from a given agent type.

Let us list some basic properties of distinguishability that will be useful later.

Proposition 1 (Reflexivity)**.** *If $\sigma \in X$ then X distinguishes σ in $\mathcal{F}$.*

Proposition 2 (Monotonicity)**.** *If $X \subseteq X'$ and X distinguishes σ in $\mathcal{F}$ then X' distinguishes σ in $\mathcal{F}$.*

Proposition 3 (Transitivity)**.** *If X distinguishes σ in $\mathcal{F}$ and $\sigma \in X'$ and X' distinguishes σ' in $\mathcal{F}$ then $X \cup X' \setminus \{\sigma\}$ distinguishes σ in $\mathcal{F}$.*

An immediate consequence is:

Proposition 4. *If $\mathcal{F}' \subseteq \mathcal{F}$ and X distinguishes σ in $\mathcal{F}$ then X distinguishes σ in $\mathcal{F}'$.*

We may also know that the agent's AF satisfies certain graph-theoretical properties such as being acyclic, odd-cycle-free, or self-attack-free. This may simplify the elicitation task. We represent these classes of AFs as follows.

Definition 3. *We define the following sets of AFs:*

- $\mathcal{F}_{acyclic} = \{F \mid F \text{ is acyclic}\}$
- $\mathcal{F}_{odd\text{-}cycle\text{-}free} = \{F \mid F \text{ contains no odd cycles}\}$
- $\mathcal{F}_{self\text{-}attack\text{-}free} = \{F \mid F \text{ contains no self-attacks}\}$

The following theorems provide an overview of the results. Each theorem summarizes the distinguishability of all considered semantics w.r.t. a set of AFs $\mathcal{F}$ by listing all minimal sets of semantics $X \subseteq \Sigma$ that distinguish a semantics σ in $\mathcal{F}$. For each theorem we provide proofs of the distinguishable cases as well as counterexamples for the maximal set of semantics $X \subseteq \Sigma$ that does not distinguish a semantics σ.

Theorem 1 is concerned with the set of all AFs, $\mathcal{F}_{all}$. Here, we can only observe that $\{\mathsf{co}\}$ distinguishes gr and pr and $\{\mathsf{ad}\}$ distinguishes pr. This follows easily from the definitions (Oikarinen and Woltran 2011). Furthermore, it should be noted that there exists no other set of semantics $X \subseteq \Sigma$ that distinguishes a semantics σ in $\mathcal{F}_{all}$, if σ is not in X.

Theorem 1. *For all $X \subseteq \Sigma$ and $\sigma \in \Sigma$, if $\sigma \notin X$ then X does not distinguish σ in $\mathcal{F}_{all}$, unless:*

1. $\mathsf{ad} \in X$ *and* $\sigma = \mathsf{pr}$
2. $\mathsf{co} \in X$ *and* $\sigma \in \{\mathsf{gr}, \mathsf{pr}\}$

Proof. Let $F, F' \in \mathcal{F}$. If $\mathsf{ad}(F) = \mathsf{ad}(F')$ then $\mathsf{pr}(F) = \mathsf{pr}(F')$. If $\mathsf{co}(F) = \mathsf{co}(F')$ then $\mathsf{pr}(F) = \mathsf{pr}(F')$ and $\mathsf{gr}(F) = \mathsf{gr}(F')$ (Oikarinen and Woltran 2011). No other relation holds in $\mathcal{F}_{all}$ as shown by the following examples: $\sigma = \mathsf{cf}$: Example 1, $\sigma = \mathsf{ad}$: Example 2, $\sigma = \mathsf{co}$: Example 3, $\sigma = \mathsf{gr}$: Example 4, $\sigma = \mathsf{pr}$: Example 5, $\sigma = \mathsf{st}$: Example 10, $\sigma = \mathsf{sst}$: Example 11. □

In Theorem 2 we are considering the set of self-attack-free AFs. Besides the exceptions from Theorem 1 we have some additional cases. The set of semantics $\{\mathsf{cf}, \mathsf{pr}\}$ distinguishes st and sst in $\mathcal{F}_{self\text{-}attack\text{-}free}$. From Proposition 3 it follows that $\{\mathsf{cf}, \mathsf{ad}\}$ and $\{\mathsf{cf}, \mathsf{co}\}$ also distinguish st and sst. Additionally, $\{\mathsf{cf}, \mathsf{sst}\}$ distinguishes st in $\mathcal{F}_{self\text{-}attack\text{-}free}$. This means that, for example, we are able to construct an st-equivalent AF by only knowing the cf- and pr-extensions.

Theorem 2. *For all $X \subseteq \Sigma$ and $\sigma \in \Sigma$, if $\sigma \notin X$ then X does not distinguish σ in $\mathcal{F}_{self\text{-}attack\text{-}free}$, unless:*

1. $\mathsf{ad} \in X$ *and* $\sigma = \mathsf{pr}$
2. $\mathsf{co} \in X$ *and* $\sigma \in \{\mathsf{gr}, \mathsf{pr}\}$
3. $\{\mathsf{cf}, \mathsf{ad}\} \subseteq X$ *and* $\sigma \in \{\mathsf{st}, \mathsf{sst}\}$
4. $\{\mathsf{cf}, \mathsf{co}\} \subseteq X$ *and* $\sigma \in \{\mathsf{st}, \mathsf{sst}\}$
5. $\{\mathsf{cf}, \mathsf{pr}\} \subseteq X$ *and* $\sigma \in \{\mathsf{st}, \mathsf{sst}\}$
6. $\{\mathsf{cf}, \mathsf{sst}\} \subseteq X$ *and* $\sigma = \mathsf{st}$

Proof. Cases 1 and 2 follow from Theorem 1 and Proposition 4. The proof for the cases 3, 4, 5 and 6 is shown below. All other relations do not hold as shown by the corresponding examples: $\sigma = \mathsf{cf}$: Example 1, $\sigma = \mathsf{ad}$: Example 2, $\sigma = \mathsf{co}$: example 3, $\sigma = \mathsf{gr}$: Example 4, $\sigma = \mathsf{pr}$: Example 5, $\sigma = \mathsf{st}$: Examples 6 and 8, $\sigma = \mathsf{sst}$: Examples 7 and 12.

We will now prove that $\{\mathsf{cf}, \mathsf{pr}\}$ distinguishes st in $\mathcal{F}_{self\text{-}attack\text{-}free}$. Let $F = (\mathsf{Arg}, R)$ and $F' = (\mathsf{Arg}, R')$. Assume that $F, F' \in \mathcal{F}_{self\text{-}attack\text{-}free}$. Suppose that $\mathsf{cf}(F) = \mathsf{cf}(F')$ and $\mathsf{pr}(F) = \mathsf{pr}(F')$ (or $\mathsf{ad}(F) = \mathsf{ad}(F')$ or $\mathsf{co}(F) = \mathsf{co}(F')$ which implies $\mathsf{pr}(F) = \mathsf{pr}(F')$).

Let $E \in \mathsf{ad}(F)$. We will first prove that $E^+_F = E^+_{F'}$. Suppose that $E^+_F \neq E^+_{F'}$. Then, either (1) $\exists a \in \mathsf{Arg} : a \in E^+_F$ and $a \notin E^+_{F'}$ or (2) $\exists a \in \mathsf{Arg} : a \in E^+_{F'}$ and $a \notin E^+_F$. In case (1), it follows from the definition of $E \in \mathsf{ad}(F)$ that $a \notin E$. Then $\nexists b \in E : (b, a) \in R'$. That means $E \cup \{a\} \in cf(F')$ and thus $E \cup \{a\} \in cf(F)$. But, since we have $\mathsf{cf}(F) = \mathsf{cf}(F')$ and $a \in E^+_F$ as well as $F, F' \in \mathcal{F}_{self\text{-}attack\text{-}free}$ it also follows that $\exists c \in E : (c, a) \in R$ and thus $E \cup \{a\} \notin cf(F)$. Case (2) follows similarly.

We will prove that $\mathsf{sst}(F) \subseteq \mathsf{sst}(F')$. Let $E \in \mathsf{sst}(F)$. Then, it follows that $E^+_F = E^+_{F'}$. Suppose $E \notin \mathsf{sst}(F')$. Since $\mathsf{pr}(F) = \mathsf{pr}(F')$ we know that $E \in \mathsf{pr}(F')$. So there must be an $E' \in \mathsf{pr}(F')$ with $E^+_{F'} \subseteq E'^+_{F'}$ and $E' \notin \mathsf{sst}(F)$. But, since $\mathsf{pr}(F) = \mathsf{pr}(F')$ and $E'^+_F = E'^+_{F'}$ it would follow that $E' \in \mathsf{sst}(F)$. Thus we have $E \in \mathsf{sst}(F')$ and we can similarly prove that $\mathsf{sst}(F) \supseteq \mathsf{sst}(F')$. We therefore have $\mathsf{sst}(F) = \mathsf{sst}(F')$.

The same procedure can be applied to prove $\{\mathsf{cf}, \mathsf{pr}\}$ distinguishes st and $\{\mathsf{cf}, \mathsf{sst}\}$ distinguishes st. □

In the odd-cycle-free case we have $\mathsf{pr} = \mathsf{st}$ (Dung 1995) and $\mathsf{st} = \mathsf{sst}$ (Caminada, Carnielli, and Dunne 2012). Hence, $\{\mathsf{ad}\}$ and $\{\mathsf{co}\}$ also distinguish st and sst in $\mathcal{F}_{odd\text{-}cycle\text{-}free}$.

Theorem 3. *For all $X \subseteq \Sigma$ and $\sigma \in \Sigma$, if $\sigma \notin X$ then X does not distinguish σ in $\mathcal{F}_{odd\text{-}cycle\text{-}free}$, unless:*

1. $\mathsf{ad} \in X$ *and* $\sigma \in \{\mathsf{pr}, \mathsf{st}, \mathsf{sst}\}$
2. $\mathsf{co} \in X$ *and* $\sigma \in \{\mathsf{gr}, \mathsf{pr}, \mathsf{st}, \mathsf{sst}\}$
3. $\mathsf{pr} \in X$ *and* $\sigma \in \{\mathsf{st}, \mathsf{sst}\}$
4. $\mathsf{st} \in X$ *and* $\sigma \in \{\mathsf{pr}, \mathsf{sst}\}$
5. $\mathsf{sst} \in X$ *and* $\sigma \in \{\mathsf{pr}, \mathsf{st}\}$

Proof. Cases 1 to 5 follow from Theorem 1 and the fact that $\mathsf{pr} = \mathsf{st} = \mathsf{sst}$ in $\mathcal{F}_{odd\text{-}cycle\text{-}free}$. All other relations do not hold as evidenced by the corresponding counterexamples: $\sigma = \mathsf{cf}$: Example 1, $\sigma = \mathsf{ad}$: Example 2, $\sigma = \mathsf{co}$: Example 3, $\sigma = \mathsf{gr}$: Example 4, $\sigma \in \{\mathsf{pr}, \mathsf{st}, \mathsf{sst}\}$: Example 8. □

In the acyclic case, the $\mathsf{co}, \mathsf{pr}, \mathsf{gr}, \mathsf{st}$ and sst semantics coincide and thus distinguish each other. Apart from that, they are also distinguished by ad. Only cf and ad are not distinguished by any of the other semantics.

Theorem 4. *For all $X \subseteq \Sigma$ and $\sigma \in \Sigma$, if $\sigma \notin X$ then X does not distinguish σ in $\mathcal{F}_{acyclic}$, unless:*

- $X \cap \Sigma \setminus \{\mathsf{cf}\} \neq \emptyset$ *and* $\sigma \in \{\mathsf{co}, \mathsf{pr}, \mathsf{gr}, \mathsf{st}, \mathsf{sst}\}$.

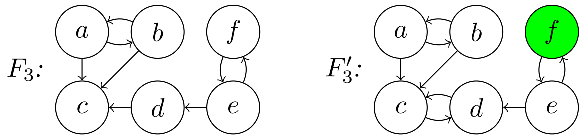
Proof. Follows from the fact that the $\mathsf{co}, \mathsf{pr}, \mathsf{gr}, \mathsf{st}$ and sst semantics coincide in acyclic AFs. Furthermore, ad distinguishes pr (Theorem 1) and thus (Proposition 3) also the others. That $\{\mathsf{cf}\}$ does not distinguish anything is shown by Example 9. That cf and ad are not distinguished is shown by Examples 1 and 2. □

Below, we list several counterexamples used in the proofs of the above theorems. The examples consist of two AFs F and F' belonging to a set of AFs $\mathcal{F}$ as defined in Definition 3. In each example we have a set of semantics $X \subseteq \Sigma$ for which F and F' have the same extensions (i.e. they are *X-equivalent*), as well as a semantics $\sigma \in \Sigma$ under which they have different extensions. We highlight one extension that exists in F' but not in F under σ semantics using green color. It follows that X does not distinguish σ in the set of AFs $\mathcal{F}$.

Example 1. *Consider the acyclic AFs* $F_1 = (\{a, b, c\}, \{(a, b)\})$ *and* $F'_1 = (\{a, b, c\}, \{(c, b)\})$. *Then the* ad, co, gr, pr, st *and* sst *extensions of* F_1 *and* F'_1 *are the same but the extensions under* cf *semantics are different.*

Example 2. *Consider the acyclic AFs* $F_2 = (\{a, b, c\}, \{(a, b), (b, c)\})$ *and* $F'_2 = (\{a, b, c\}, \{(c, b), (b, a)\})$. *Then the* cf, co, gr, pr, st *and* sst *extensions of* F_2 *and* F'_2 *are the same but the extensions under* ad *semantics are different.*

Example 3. *Consider the odd-cycle-free AFs* F_3 *and* F'_3.

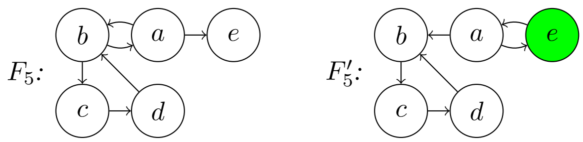


Then the cf, ad, gr, pr, st *and* sst *extensions of F_3 and F_3' are the same but the extensions under* co *semantics are different.*

Example 4. *Consider the odd-cycle-free AFs F_4 and F_4'.*

Then the cf, ad, pr, st *and* sst *extensions of F_4 and F_4' are the same but the extensions under* gr *semantics are different.*

Example 5. *Consider the self-attack-free AFs F_5 and F_5'.*

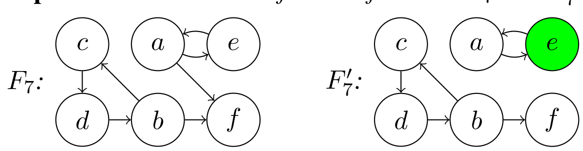


Then the cf, gr, st *and* sst *extensions of F_5 and F_5' are the same but the extensions under* pr *semantics are different.*

Example 6. *Consider the self-attack-free AFs F_6 and F_6'.*

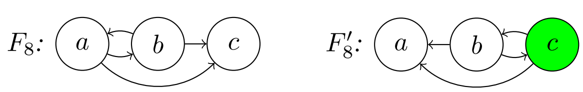
Then the ad, co, gr, pr *and* sst *extensions of F_6 and F_6' are the same but the extensions under* st *semantics are different.*

Example 7. *Consider the self-attack-free AFs F_7 and F_7'.*



Then the ad, co, gr, pr *and* st *extensions of F_7 and F_7' are the same but the extensions under* sst *semantics are different.*

Example 8. *Consider the odd-cycle-free AFs F_8 and F_8'.*

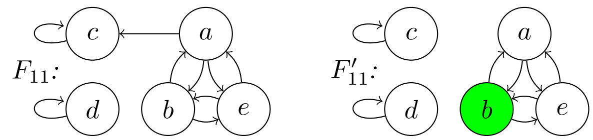


Then the cf, gr *extensions of F_8 and F_8' are the same but the extensions under* pr, st *and* sst *semantics are different.*

Example 9. *Consider the acyclic AFs* $F_9 = (\{a,b\},\{(a,b)\})$ *and* $F_1' = (\{a,b\},\{(b,a)\})$. *Then the* cf *extensions of F_9 and F_9' are the same but the extensions under* ad, co, gr, pr, st *and* sst *semantics are different.*

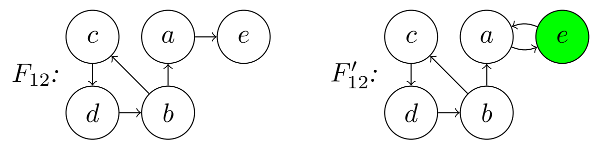
Example 10. *Consider the AFs:* $F_{10} = (\{a,b\},\{(a,a)\})$ *and* $F_{10}' = (\{a,b\},\{(a,a),(b,a)\})$. *Then the* cf, ad, co, gr, pr *and* sst *extensions of F_{10} and F_{10}' are the same but the extensions under* st *semantics are different.*

Example 11. *Consider the AFs F_{11} and F_{11}'.*



Then the cf, ad, co, gr, pr *and* st *extensions of F_{11} and F_{11}' are the same but the extensions under* sst *semantics are different.*

Example 12. *Consider the self-attack-free AFs F_{12} and F_{12}'.*



Then the cf, gr *and* st *extensions of F_{12} and F_{12}' are the same but the extensions under* sst *semantics are different.*

4 Discussion and Future Work

We studied distinguishability in argumentation and discussed its relevance in the context of argumentation framework elicitation. The results we obtained answer the question: which agent types distinguish which semantics? If we ignore cases that follow from Propositions 1, 2 and 3 (reflexivity, monotonicity and transitivity) then we see that the ad semantics is not distinguished by anything, not even in the acyclic case. The co semantics is not distinguished by anything. The gr semantics is distinguished by co. The pr semantics is distinguished by ad and co, in the odd-cycle-free case also by st and sst. Finally, in the cycle-free case, all semantics except cf and ad distinguish each other. The st and sst semantics are distinguished in the absence of self-attacks by $\{$cf, pr$\}$. In the odd-cycle-free case, they are distinguished by pr and by each other. The st and sst semantics in the self-attack-free case are interesting because they demonstrate that the cf and pr semantics together distinguish something that they do not distinguish alone. This is the only case where "the sum of the parts is greater than the whole".

Altogether, our results show that AF elicitation is anything but trivial—asking about one semantics usually does not yield knowledge about other semantics except in the small number of cases that we identified. Thus, we have to find more sophisticated methods to elicit AFs. However, the exceptions we found could serve as a starting point to develop such methods. Therefore, the problem of *how* to elicit an AF in an efficient manner is still subject to future work. For this, we require a procedure to decide which questions to ask. What is interesting here is the relation between an agent's type, the desired equivalence class to elicit, and the minimum number of necessary questions (as a function of the number of arguments). For instance, to determine the co-equivalence class of the agent's AF, it makes no difference whether the agent's type is $\{$co$\}$ or $\{$cf, co$\}$. However, the second type may allow us to elicit the AF more efficiently.

Acknowledgements

The research reported here was supported by the Deutsche Forschungsgemeinschaft (project numbers 375588274 and 423456621).

References

Baroni, P.; and Giacomin, M. 2008. A systematic classification of argumentation frameworks where semantics agree. *Frontiers in Artificial Intelligence and Applications* 172: 37.

Caminada, M. W. A.; Carnielli, W. A.; and Dunne, P. E. 2012. Semi-stable semantics. *J. Log. Comput.* 22(5): 1207–1254. doi:10.1093/logcom/exr033. URL https://doi.org/10.1093/logcom/exr033.

Chen, L.; and Pu, P. 2004. Survey of preference elicitation methods. Technical report. Tech. Rep. IC/200467, Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77(2): 321–358.

Dunne, P. E.; Dvořák, W.; Linsbichler, T.; and Woltran, S. 2015. Characteristics of multiple viewpoints in abstract argumentation. *Artificial Intelligence* 228: 153–178.

Dunne, P. E.; Spanring, C.; Linsbichler, T.; and Woltran, S. 2016. Investigating the Relationship between Argumentation Semantics via Signatures. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 1051–1057.

Kido, H.; and Liao, B. 2019. A Bayesian Approach to Direct and Inverse Abstract Argumentation Problems. *CoRR* abs/1909.04319. URL http://arxiv.org/abs/1909.04319.

Kuhlmann, I. 2021. Towards Eliciting Attacks in Abstract Argumentation Frameworks. *Online Handbook of Argumentation for AI* 2.

Niskanen, A.; Wallner, J.; and Järvisalo, M. 2019. Synthesizing argumentation frameworks from examples. *Journal of Artificial Intelligence Research* 66: 503–554.

Oikarinen, E.; and Woltran, S. 2011. Characterizing strong equivalence for argumentation frameworks. *Artificial intelligence* 175(14-15): 1985–2009.

Pührer, J. 2020. Realizability of three-valued semantics for abstract dialectical frameworks. *Artif. Intell.* 278. doi:10.1016/j.artint.2019.103198. URL https://doi.org/10.1016/j.artint.2019.103198.

Riveret, R.; and Governatori, G. 2016. On learning attacks in probabilistic abstract argumentation. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 653–661.