Strong Explanations in Abstract Argumentation

Markus Ulbricht,[1] Johannes P. Wallner[2]
[1] Department of Computer Science, Leipzig University, Germany
[2] Institute of Software Technology, Graz University of Technology, Austria
mulbricht@informatik.uni-leipzig.de, wallner@ist.tugraz.at

Abstract

Abstract argumentation constitutes both a major research strand and a key approach that provides the core reasoning engine for a multitude of formalisms in computational argumentation in AI. Reasoning in abstract argumentation is carried out by viewing arguments and their relationships as abstract entities, with argumentation frameworks (AFs) being the most commonly used abstract formalism. Argumentation semantics then drive the reasoning by specifying formal criteria on which sets of arguments, called extensions, can be deemed as jointly acceptable. Such extensions provide a basic way of explaining argumentative acceptance. Inspired by recent research, we present a more general class of explanations: in this paper we propose and study so-called strong explanations for explaining argumentative acceptance in AFs. A strong explanation is a set of arguments such that a target set of arguments is acceptable in each subframework containing the explaining set. We formally show that strong explanations form a larger class than extensions, in particular giving the possibility of having smaller explanations. Moreover, assuming basic properties, we show that any explanation strategy, broadly construed, is a strong explanation. We show that the increase in variety of strong explanations comes with a computational trade-off: we provide an in-depth analysis of the associated complexity, showing a jump in the polynomial hierarchy compared to extensions.

1 Introduction

Computational models of argumentation in Artificial Intelligence (AI) (Baroni et al. 2018; Bench-Capon and Dunne 2007) provide formal approaches to reason argumentatively, with a wide variety of application avenues, such as legal reasoning, medical sciences, and e-governmental issues (Atkinson et al. 2017). Reasoning in this way is carried out by instantiation of argument structures from a knowledge base (Bondarenko et al. 1997; Modgil and Prakken 2013; García and Simari 2004; Besnard and Hunter 2008), which represent all that can be argued for. Inconsistencies within knowledge bases are then represented by conflicts among arguments, which are modelled via (directed) attacks between arguments, reflecting a counter argument relation.

For many formal approaches to argumentation in AI, an abstract representation of arguments and their attacks, together referred to as argumentation frameworks (AFs) (Dung 1995), is sufficient in order to provide rational accounts on what can be argued for (Caminada 2018). Known as the area of abstract argumentation, such formalisms provide so-called argumentation semantics (Baroni, Caminada, and Giacomin 2011) on which sets of arguments can be deemed jointly acceptable together. Multiple argumentation semantics were defined, fitting different purposes and range from more inclusive to more cautious modes of reasoning. An important semantics are admissible sets of arguments, which are non-conflicting sets that counter-attack any attack from outside the set, providing a way to argumentatively defend each argument within the set.

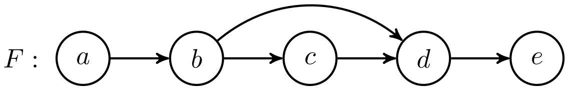

Figure 1: Example AF

Admissible sets, or, more broadly, extensions under a semantics, provide a key feature for argumentation: argumentative explanations in the form of arguments, which can be used to show acceptability of each argument in the set. For instance, acceptance of an argument can be specified as being a member of an admissible set (or an extension of a semantics). This is commonly referred to as credulous acceptance of that argument.

Example 1.1. Assume it is 2020 and some agents discuss whether or not the next conference should be held virtually. Consider the following arguments which are brought forward during the debate: "The conference should be held virtually in order to avoid a 'super spreader' event" (e); "This is not the same experience as a meeting in person" (d); "I would agree with you, but not in 2020" (c); "I would *never* agree with the both of you, because all this flying around destroys our environment" (b); "I think our small community has an overall low impact on climate change" (a).

Here each argument attacks its predecessor, except for b which attacks both d and c. This debate thus induces the AF in Figure 1. Say, we desire to check argumentative (credulous) acceptability of argument e, in favor of a virtual conference. There is one admissible set, $\{a, c, e\}$, that contains

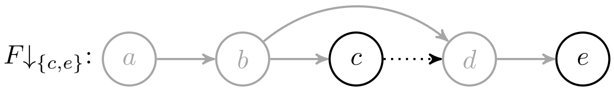

Figure 2: Subframework

e: this set is non-conflicting and defends e against the argument d and counters b by the attack from a.

Importantly, this admissible set $\{a, c, e\}$ is sufficient to show acceptability of e when faced with any possible argument in the AF. That is, by posing arguments a, c, and e, one is equipped to always defend the desired argument e. Interestingly, a closer inspection of the AF F reveals that argument a is not strictly required in being prepared to defend e. Consider the subframework in Figure 2 containing only c and e. If we position ourselves with only these two arguments, we already have sufficient evidence to support e: The only way to counter argument c is b about the environment. Although this is a counterattack to c in a certain sense, it is itself a counterargument to d stating that the conference should take place in person. So in any case, argument d in favor of a meeting in person is defeated. Then presence or absence of the argument a decides whether or not the concerns about climate change are taken seriously in this debate; however (credulously) accepting the issue of holding the conference virtually in 2020 is not affected.

As illustrated in the example, when looking at structural subframeworks representing a current state of the argumentation, admissible sets do not constitute minimal requirements for being prepared to show acceptability of a desired argument under a credulous viewpoint. Put differently, with even less arguments than prescribed by admissibility we can find sufficiently many for our target set to be credulously accepted under admissibility.

Recent advances termed strong explanations (Brewka and Ulbricht 2019; Saribatur, Wallner, and Woltran 2020), initially for strong inconsistency (Brewka, Thimm, and Ulbricht 2019), provide us with the key formal ingredient to identify argumentative explanations on AFs as indicated above: a strong explanation is a set of arguments such that a target set of arguments is acceptable in each subframework containing the explaining set. In the example above the subframework induced by $\{c, e\}$ is a strong explanation for e (under admissibility). In this paper we study such strong explanations for credulous acceptability under the most common semantics for AFs. In particular, our main contributions are as follows.

- We show that strong explanations (i) offer provably more variety than extensions under a semantics σ, and (ii) can lead to smaller sets of arguments that can be used to find the target arguments acceptable.
- We show that under basic assumptions, any explanation strategy based on sets of arguments inducing subframeworks is a strong explanation. We further compare explanations based on extensions and strong explanations, and find that subset minimal strong explanations are not necessarily conflict-free, in contrast to σ-extensions.
- We show that relative to extensions, strong explanations have a trade-off in terms of computational complexity: we pinpoint the complexity of several decision tasks for strong explanations, indicating higher complexity than for extensions.

2 Background

We recall background on AFs (Dung 1995) and their semantics.

An AF is a directed graph $F = (A, R)$ where A represents a set of (abstract) arguments and $R \subseteq A \times A$ models *attacks* between them. In this paper we consider finite AFs only. For $a, b \in A$, if $(a, b) \in R$ we say that a *attacks* b as well as a *attacks* (the set) E given that $b \in E \subseteq A$; and $E' \subseteq A$ attacks b if $a \in E'$. We let $E^+ = \{a \in A \mid E \text{ attacks } a\}$ and $E^- = \{a \in A \mid a \text{ attacks } E\}$.

Definition 2.1. Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is conflict-free in F, denoted by $E \in cf(F)$, iff for no $a, b \in E$ we have $(a, b) \in R$. We say a set E defends an argument a (in F) if any attacker of a is attacked by an argument $b \in E$.

In this paper we consider the classical semantics defined by Dung (1995): *admissible*, *complete*, *stable*, *preferred*, and *grounded* semantics (abbr. ad, co, stb, pr, gr). Each semantics returns a set of sets of acceptable positions which are defined as follows (cf. Baroni, Caminada, and Giacomin (2018) for a recent overview).

Definition 2.2. Let $F = (A, R)$ be an AF and $E \in cf(A)$.

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

We refer to an extension under a semantics $\sigma \in \{ad, co, pr, stb, gr\}$ also as σ-extension.

The notion of a subframework for a given AF F induced by a set $S \subseteq A$ of arguments is defined by $F' = F{\downarrow}_S = (A \cap S, R \cap (S \times S))$. That is, F' contains all arguments in S and all incident attacks from arguments in S.

A main reasoning task on AFs is then given by credulous acceptance of an argument under a semantics σ. For an AF F and a semantics σ we say an argument $a \in A$ is *credulously accepted* if $a \in \bigcup \sigma(F)$.

3 Explanation Strategies

A main approach to explanations regarding acceptance of arguments are σ-extensions. Towards a general viewpoint, we define general explanation strategies that are argument based, i.e., focus on sets of arguments as being an explanation (as σ-extensions do). A general argument-explanation strategy can then be defined as a set of sets of arguments. We focus on explanation of credulous acceptance, with a slight generalization of credulous acceptance: for a given set X, we aim to explain when X is part of one σ-extension.

Definition 3.1. Let $F = (A, R)$ be an AF and $X \subseteq A$. An *argument-explanation strategy* for X in F is a set $\mathcal{S} \subseteq 2^A$. A set $S \in \mathcal{S}$ is called an argument-based explanation (according to $\mathcal{S}$).

Two very basic requirements for explanation strategies are that they are, what we call, σ-basic and satisfy σ-existence, when explaining X under semantics σ.

σ-basic $S \in \mathcal{S}$ implies $X \subseteq E$ for some $E \in \sigma(F{\downarrow}_S)$.

σ-existence If $X \subseteq E$ for some $E \in \sigma(F)$ then $\mathcal{S} \neq \emptyset$.

That is, σ-basic states that if $S \in \mathcal{S}$ for an explanation strategy $\mathcal{S}$, then there must be a σ-extension containing X (at least) in the subframework induced by the explanation S. An explanation strategy satisfies σ-existence if there is at least one explanation whenever each argument in X is part of one σ-extension.

Another basic property is monotonicity.

Monotonicity If $S \in \mathcal{S}$, then $S' \in \mathcal{S}$ for any S' with $S \subseteq S' \subseteq A$.

That is, an explanation strategy $\mathcal{S}$ satisfies Monotonicity if for each explanation S we find each superset S' of S in $\mathcal{S}$.

We next view extensions as an argument-explanation strategy, and subsequently provide the main notion studied in this paper: strong explanations.

Extensions as Explanations

Extensions under a semantics σ are the (nowadays classical) approach to explaining why an argument (or a set of arguments) is credulously acceptable. We next phrase σ-extensions straightforwardly as explanation (strategies) for a set X. Due to the usefulness, we explicate in the subsequent definition $\subseteq$-minimal σ-extensions.

Definition 3.2. Let $F = (A, R)$ be an AF, $X \subseteq A$ and σ any semantics. A set $S \subseteq A$ is called a (minimal) *extension-based σ-explanation* for X if (it is minimal s.t.) $X \subseteq S$ and $S \in \sigma(F)$.

It follows that if there is an extension-based σ-explanation for X for a given AF F, then every argument in X is credulously accepted under σ, and the set of all extension-based σ-explanations forms an explanation strategy. By definition, extension-based σ-explanations are σ-basic and satisfy σ-existence. We summarize satisfaction of properties of explanation strategies in Table 1.

Extension-based σ-explanations do not satisfy Monotonicity: for all semantics considered in this paper we can find an AF F such that if $E \in \sigma(F)$ then there is an $E' \supsetneq E$ with E' not being a σ-extension. Nevertheless, σ-extensions are robust in a different sense: if E is a σ-extension in an AF F, then E remains being part of a σ-extension in any subframework F' that includes at least E.

Definition 3.3. A semantics σ is called *robust* if for each AF $F = (A, R)$ it holds that $E \in \sigma(F)$ implies that there is an $E' \in \sigma(F{\downarrow}_S)$ with $E \subseteq E'$ for each S with $E \subseteq S \subseteq A$.

Several main semantics of AFs are robust.

Proposition 3.4. *It holds that admissible, complete, grounded, stable, and preferred semantics are robust.*

If we strengthen Definition 3.3 by requiring that in each subframework $F{\downarrow}_S$, we find E *exactly* as a σ-extension, then satisfaction is different.

Definition 3.5. A semantics σ is called strongly robust if for each AF $F = (A, R)$ it holds that $E \in \sigma(F)$ implies that $E \in \sigma(F{\downarrow}_S)$ for each S with $E \subseteq S \subseteq A$.

Proposition 3.6. *It holds that admissible and stable semantics are strongly robust.*

On the other hand, complete, grounded, and preferred are not strongly robust.

Example 3.7. Let $F = (A, R)$ be an AF with $A = \{a, b, c\}$ and $R = \{(c, b), (c, c)\}$, i.e., we have three arguments and an attack from c to b with c a self-attacking argument. It holds that $\{a\}$ is the unique preferred (grounded) extension of F. However, for $S = \{a, b\}$ we find that $\{a, b\}$ is preferred (grounded, complete) in $F{\downarrow}_S$ (since the attack from c onto b is removed).

Upward-closed Extensions

We want to mention that there is a natural way to artificially make extensions as explanations monotonic: one might simply accept any superset S' of an extension $S \in \sigma(F)$ with $X \subseteq S$ as an explanation as well. Formally, this yields:

Definition 3.8. Let $F = (A, R)$ be an AF, $X \subseteq A$ and σ any semantics. A set $S' \subseteq A$ is called an *upward-closed extension-based σ-explanation* for X if there is some $S \in \sigma(F)$ with $X \subseteq S \subseteq S'$.

Although this approach ensures monotonicity by definition (and both σ-basic as well as σ-existence can be seen with reasonable effort), it is clear that from an intuitive point of view, upward-closed extension-based σ-explanations do not provide novel information compared to the extension-based σ-explanation we introduced before. We will thus continue our investigation with a more informative approach.

Strong Explanations

Let us now turn to define our main notion of *strong* σ-explanations. They are inspired by recent related notions (Brewka and Ulbricht 2019; Brewka, Thimm, and Ulbricht 2019; Saribatur, Wallner, and Woltran 2020).

Definition 3.9. Let $F = (A, R)$ be an AF, $X \subseteq A$ a set of arguments and σ any semantics. A set $S \subseteq A$ is called a (minimal) *strong σ-explanation* for X if (it is minimal s.t.) for each AF $F' = F{\downarrow}_{A'}$ with $S \subseteq A' \subseteq A$, there is $E' \in \sigma(F')$ with $X \subseteq E'$.

Speaking in terms of the concepts we considered throughout the present paper so far, the definition of strong σ-explanations is inspired by the σ-basic property and additionally requires monotonicity.

Let us consider the following basic examples of strong explanations.

Example 3.10. Let F be the AF from Figure 3. Assume $X = \{c\}$. Then $S = \{a_1, c\}$ is a strong ad-explanation for X in F. That is, given a_1 and c, no matter with arguments we include, X will always occur in at least one admissible extension.

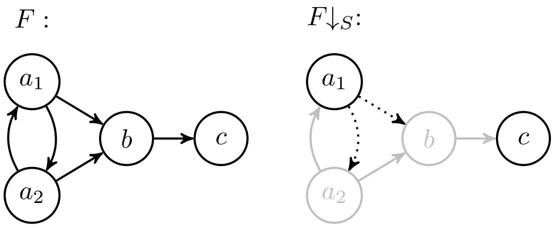

Figure 3: Strong explanation

Having established Definition 3.9 in a formal way, let us now reconsider our motivating example.

Example 3.11 (Example 1.1 ctd.)**.** We formally show that $\{c, e\}$ is a minimal strong ad-explanation for e: it is easy to see that c is required because otherwise the subframework consisting of the arguments d and e would not contain e as a credulously accepted argument. However, the subframework induced by $\{b, c, e\}$ possesses $\{b, e\}$ as an admissible extension and in the whole AF we get $\{a, c, e\}$ as admissible extension. In summary, for each A' satisfying $\{c, e\} \subseteq A' \subseteq A$, there is an admissible extension $E' \in ad(F{\downarrow_{A'}})$ with $X \subseteq E'$.

Let us now collect some basic properties of strong explanations which will be useful throughout the paper:

Proposition 3.12. *Let $F = (A, R)$ be an AF, $S \subseteq A$, $X \subseteq A$, and $\sigma \in \{ad, co, gr, stb, pr\}$ a semantics.*

- *There is a strong σ-explanation S for X iff there is some E with $X \subseteq E \in \sigma(F)$.*
- *If S is a strong σ-explanation for X, then $X \subseteq S$.*
- *S is a strong ad-explanation for X iff S is a strong co-explanation for X iff S is a strong pr-explanation for X.*

Another necessary condition for S to be a strong σ-explanation is defeating all attackers of X. Formally:

Proposition 3.13. *Let $F = (A, R)$ be an AF, $X \subseteq A$, and $\sigma \in \{ad, co, gr, stb, pr\}$. If S is a strong σ-explanation for X, then $X^- \subseteq S^+$.*

Extensions and strong explanations are related in that each σ-extension containing a set X is a strong σ-explanation for X, in case σ is robust.

Theorem 3.14. *Let $F = (A, R)$ be an AF, $X \subseteq A$ a set of arguments and σ a semantics that is robust. If $E \in \sigma(F)$ s.t. $X \subseteq E$, then E is a strong σ-explanation for X.*

Proof. Let $E \subseteq S \subseteq A$ and consider the subframework $F{\downarrow_S}$. Since σ is robust, there is some $E' \in \sigma(F{\downarrow_S})$ with $E \subseteq E'$. By $X \subseteq E$, we find $X \subseteq E'$ as well. □

Further, strong explanations form a strictly larger class of explanations. We now show several results in this direction. First, strong σ-explanations satisfy Monotonicity, directly by definition. Thus, any superset of a σ-extension containing a set X of arguments is also a strong σ-explanation, but not necessarily also a σ-extension (e.g., if attacks occur in a superset of a σ-extension).

More broadly, *any* argument-based explanation strategy that satisfies Monotonicity and is σ-basic is a strong σ-explanation.

Theorem 3.15. *Let $F = (A, R)$ be an AF and $X \subseteq A$, and $\mathcal{S}$ be an argument-explanation strategy for X in F. If $\mathcal{S}$ is σ-basic and satisfies Monotonicity, then $S \in \mathcal{S}$ is a strong σ-explanation for X.*

Proof. Assume that $\mathcal{S}$ satisfies Monotonicity and is σ-basic, and let $S \in \mathcal{S}$. Let $F{\downarrow_{A'}}$ be a subframework induced by $S \subseteq A' \subseteq A$. By Monotonicity, we have $A' \in \mathcal{S}$. By being σ-basic, we can infer that there is an $E \in \sigma(F{\downarrow_{A'}})$ with $X \subseteq E$. Thus, S is a strong σ-explanation for X. □

We want to emphasize that strong σ-explanations are σ-basic and satisfy Monotonicity themselves; thus we found two rather mild properties which already suffice to characterize them:

Corollary 3.16. *Let $F = (A, R)$ be an AF and $X \subseteq A$, and $\mathcal{S}$ be an argument-explanation strategy for X in F. Then $\mathcal{S}$ is the greatest set in 2^A satisfying σ-basic and Monotonicity iff it is the set of all strong σ-explanations for X.*

Proof. ($\Rightarrow$) By Theorem 3.15, each $S \in \mathcal{S}$ is a strong σ-explanation. Now assume there is some strong σ-explanation S which does not occur in $\mathcal{S}$. Since strong σ-explanations are σ-basic and satisfy Monotonicity, $S \in \mathcal{S}$ must hold; a contradiction.

($\Leftarrow$) Clearly, the set of all strong σ-explanations for X is σ-basic and satisfies Monotonicity. Now assume there is $S \in \mathcal{S}$, but S is no strong σ-explanation. This contradicts Theorem 3.15. Hence the set of all strong σ-explanations is the greatest with the two mentioned properties. □

In addition to inclusion of σ-extensions and their (proper) supersets also proper subsets of a σ-extension can form a strong σ-explanation. Hence strong σ-explanations can provide smaller explanations than given by σ-extensions.

Moreover, a strong σ-explanation for a set X does not necessarily conform to the requirements imposed by the semantics σ. Formally, consider the following property of argument-based explanation strategies.

(Min-)CF If S is (minimal) in $\mathcal{S}$, then $S \in cf(F)$.

That is, according to an argument-based explanation strategy $\mathcal{S}$, if S is an explanation (for a set X), then the preceding property requires that S is conflict-free in the underlying AF F. Weakening the requirement, an argument-based explanation strategy $\mathcal{S}$ satisfies Min-CF, if subset minimal explanations $S \in \mathcal{S}$ are conflict-free. While σ-extensions satisfy (Min-)CF by definition for the main semantics, strong σ-explanations are more varied: they do not satisfy Min-CF (and, thus, also not CF).

Example 3.17. Consider the AF F from Figure 4 and any semantics. Let us verify that $S = \{f, g, d, b\}$ is a minimal explanation for $X = \{b\}$. Consider $F{\downarrow_{A'}}$ with $c \in A'$. Then e attacks b and requires the counterattack provided by g; $E = \{c, g, b\}$ is an extension containing b. If $c \notin A'$, then f must attack a and thus, d must attack e; in this case,

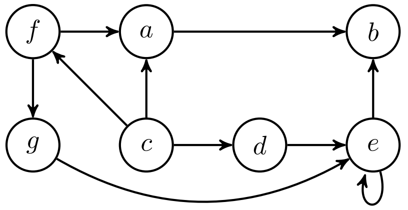

Figure 4: Minimal strong explanation not conflict-free

$E = \{d, f, b\}$ is the desired extension. From the cases discussed here it can be inferred that $S \notin cf(F)$ is minimal.

In general terms, supposing both Monotonicity and conflict-freeness for explanations does not lead to useful argument-based explanation strategies. We show that only trivialized explanation strategies can satisfy both properties.

Proposition 3.18. *Let $F = (A, R)$ be an AF and $X \subseteq A$, and $\mathcal{S}$ be an argument-explanation strategy for X in F. If $\mathcal{S}$ satisfies Monotonicity and CF, then $\mathcal{S} = \emptyset$ or $R = \emptyset$.*

Proof. Assume $\mathcal{S}$ satisfies both Monotonicity and CF. If $\mathcal{S} \neq \emptyset$, then let $S \in \mathcal{S}$. By Monotonicity we get $A \in \mathcal{S}$, and by CF, we can infer that $A \in cf(F)$, implying $R = \emptyset$. □

Another basic requirement which is inherent for admissible-based extensions, but not satisfied by strong explanations is Defense (Example 3.17 is a counterexample):

Defense If $S \in \mathcal{S}$, then S defends itself in F.

The last formal property we are going to discuss within this section is called *independence*. It formalizes that an explanation S should not rely on any argument which is not contained in S, that is, should provide sufficient evidence for X independent of the remainder of the AF.

Independence If S is an explanation in F and $a \notin S$, then S is an explanation in $F \setminus \{a\}$.

As one can easily infer, Independence is satisfied by extension-based σ-explanations, whenever σ is strongly robust: Simply consider the definition of strongly robust semantics, let $S = A \setminus \{a\}$ and use that there must be $E \in \sigma(F{\downarrow_S})$:

Lemma 3.19. *For any strongly robust σ, each extension-based σ-explanation for any set X satisfies Independence.*

Moreover, strong explanations satisfy Independence by definition and we thus end up with the following:

Proposition 3.20. *Let $F = (A, R)$ be an AF, $X \subseteq A$, $\sigma \in \{ad, co, gr, stb, pr\}$, and $\tau \in \{ad, stb\}$. Then each*

1. *strong σ-explanation S for X satisfies Independence,*
2. *extension-based τ-explanation S for X satisfies Independence.*

A counterexample for the second item in the above proposition for $\tau \in \{co, gr, pr\}$ can be obtained as follows:

Example 3.21. Consider again the AF from Example 3.7. If we let $X = \{a\}$, then of course $S = \{a\}$ itself is an extension based σ-explanation for X. However, in $F{\downarrow_{\{a,b\}}}$, S is not a complete extension anymore.

strategy	σ-extensions	strong σ-expl.
σ-basic	✓	✓
σ-existence	✓	✓
Monotonicity	X	✓
Min-CF	✓	X
Defense	✓	X
Independence	$\{ad, stb\}$	✓

Table 1: Summary of properties of explanation strategies for $\sigma \in \{ad, co, gr, stb, pr\}$.

A summary of the compliance of our explanation strategies with the desirable properties we developed is reported in Table 1.

4 Computational Complexity

In this section we investigate the complexity of reasoning via strong σ-explanations under all semantics considered in this paper. To keep this section within reasonable space, we focus on the following natural computational (decision) tasks.

Ver-Expl$_\sigma$
Input: (F, S, X) where $F = (A, R)$ and $S, X \subseteq A$
Output: TRUE iff S is a strong σ-explanation for X in F

Ver-Min-Expl$_\sigma$
Input: (F, S, X) where $F = (A, R)$ and $S, X \subseteq A$
Output: TRUE iff S is a minimal strong σ-expl. for X in F

So given an AF $F = (A, R)$, a semantics σ, and $X \subseteq A$ as well as $S \subseteq A$, the tasks are to decide whether (i) S is a strong σ-explanation for X in F, and (ii) whether in addition S is subset minimal. Similarly as in the case of σ-extensions, both tasks give crucial insights into computational properties of strong σ-explanations.

Before delving into our results, we first remark that existence results, i.e., tasks that ask whether a strong σ-explanation exists, boil down to complexity of credulous acceptance in AFs under semantics σ (with the minor difference that instead of querying a single argument we ask for a set of arguments to be credulously accepted in a single σ-extension). It was already established that deciding credulous acceptance of an argument is decidable in polynomial time for grounded semantics and NP-complete for admissible, complete, preferred, and stable semantics. It is straightforward to see that the same complexity bounds hold for asking whether there is a σ-extension containing all arguments in a queried argument set X. Moreover, the decision tasks of verifying whether a given set E is a σ-extension are, likewise, established. We recall the corresponding results in Table 2, see also Dvořák and Dunne (2018). Verifying whether a set of arguments E (i) contains a set X, and (ii) is a subset minimal σ-extension containing X (i.e., there is no σ-extension E' containing X and $E' \subsetneq E$) is decidable in polynomial time for grounded and stable semantics: the grounded extension unique and if a set is stable all other stable extensions are incomparable. The problem is coNP-complete for preferred and complete semantics: If a set is preferred (and contains X), then each other preferred extension is incomparable; if a set is complete (and contains X), there might be a smaller complete extension which also contains X. The latter claim requires a formal reduction to establish hardness, which we do not give here due to space restrictions. For admissible semantics, the problem is in P, as we show next.

Proposition 4.1. *Deciding whether a given set of arguments is subset minimally admissible such that a given set of arguments is contained is in P.*

Proof. For given sets $E, X \subseteq A$ and AF $F = (A, R)$, for each $a \in E$ perform the following

1. $E' := E \setminus \{a\}$
2. $E' := E' \setminus \{a \in E' \mid a$ not defended by E' in $F\}$
3. if $E' = \emptyset$ or $E' \in ad(F)$ terminate, otherwise go to 2.

The above algorithm terminates if A is finite. If a returned E' is admissible and $X \subseteq E'$, then E is not minimal. Otherwise, if for each $a \in E$ either an empty set or an admissible E' not containing X is returned, we claim that E is subset minimal under the stated conditions. Suppose E is not subset minimal, but there is an $E' \in ad(F)$ with $E' \subsetneq E$ and $X \subseteq E'$. There exists an $a \in E \setminus E'$. Let $def_F(S) = \{a \in S \mid s$ defended by S in $F\}$. Iterating $def_F(E \setminus \{a\})$ results in an admissible set that contains E': it holds that $E' = def_F(E')$ (is a fixed-point), for any $T \supseteq E'$ we have $E' \subseteq def_F(T)$ (since defense is monotonic), and applying $def_F(T)$ either yields a fixed-point (then the result is admissible and a superset of E'), or a proper subset T' of T is returned for which it holds that $E' \subseteq T'$. Thus, iteratively applying $def_F(E \setminus \{a\})$ results in a superset (not necessarily proper) of E' that is admissible. □

Let us now turn to strong explanations. As already pointed out in the introduction, in comparison to σ-extensions we have to accept a higher computational complexity in general. As the following result shows, already for grounded extensions the verification problem is coNP-complete.

We want to mention that the following result can also be inferred via (Baumeister, Neugebauer, and Rothe 2018, Theorem 14)[1], but we decided to sketch a novel proof here in order to hint at the technique which is required to infer the subsequent complexity results.

Theorem 4.2. *The problem* VER-EXPL_{gr} *is coNP-complete.*

Proof. (Sketch). For hardness, let us assume we are given a formula $\Phi = \exists X \phi(X)$ with $\phi = \{C_1, \ldots, C_r\}$ in CNF over variables in $X = \{x_1, \ldots, x_n\}$. We adapt the well-known standard translation (see e.g. (Dvořák and Dunne 2018, Reduction 3.6)): We let

$$\begin{aligned} A =& \{\varphi\} \cup \{\bar{\varphi}\} \cup \{C_i \mid i = 1, \ldots, r\} \cup \\ & X \cup \{\bar{x} \mid x \in X\} \cup \{\bot_j \mid j = 1, \ldots, n\} \cup \\ & \{d_j \mid j = 1, \ldots, n\} \cup \{\bar{d}_j \mid j = 1, \ldots, n\} \end{aligned}$$

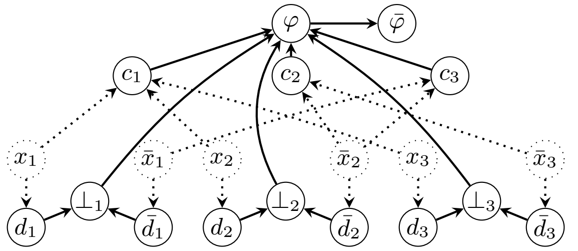

Figure 5: Illustration of the AF F from Theorem 4.2, applied to ϕ with clauses $\{\{x_1, x_2, x_3\}, \{\bar{x}_2, \bar{x}_3\}, \{\bar{x}_1, \bar{x}_2\}\}$.

and the set R of attacks is given via

$$\begin{aligned} R =& \{(\varphi, \bar{\varphi})\} \cup \{(C_i, \varphi) \mid i = 1, \ldots, r\} \cup \\ & \{(x, C_i) \mid x \in C_i,\ i = 1, \ldots, r\} \cup \\ & \{(\bar{x}, C_i) \mid \neg x \in C_i,\ i = 1, \ldots, r\} \cup \\ & \{(x_j, d_j), (\bar{x}_j, \bar{d}_j) \mid j = 1, \ldots, n\} \cup \\ & \{(d_j, \bot_j), (\bar{d}_j, \bot_j) \mid j = 1, \ldots, n\} \cup \\ & \{(\bot_j, \varphi) \mid j = 1, \ldots, n\}. \end{aligned}$$

An example of this construction is depicted in Figure 5. Note that the usual mutual attacks between the X-variables are omitted. Let $F = (A, R)$. We have: $S := A \setminus (X \cup \bar{X})$ is a gr-explanation for $\{\bar{\varphi}\}$ iff $\Phi = \exists X : \phi(X)$ evaluates to false.

($\Rightarrow$) Suppose the contrary, i. e. Φ evaluates to true. Consider a satisfying assignment $\omega : X \to \{0, 1\}$ and let X_ω be the corresponding set of X-variables, i. e. $x_i \in X_\omega$ iff $\omega(x_i) = 1$ and $\bar{x}_i \in X_\omega$ iff $\omega(x_i) = 0$. It is easy to see that by construction, φ is contained in the grounded extension of $F\downarrow_{S \cup X_\omega}$. That is, all C_i arguments are attacked (because the assignment satisfies ϕ) and all $\bot_j$ are attacked (because the assignment is well-defined). Thus $\bar{\varphi}$ is rejected in this subframework and we hence see that S is no gr-explanation.

($\Leftarrow$) Suppose the contrary, i. e. S is no gr-explanation for $\{\bar{\varphi}\}$. Then there must be a set $X_\omega \subseteq X \cup \bar{X}$ s.t. the grounded extension of $F\downarrow_{S \cup X_\omega}$ does not contain $\bar{\varphi}$. In this case, φ is contained. This means no $\bot_j$ occurs in the grounded extension which in turn means X_ω corresponds to a well-defined (partial) assignment $\omega : X \to \{0, 1\}$. Since φ is defended by X_ω, ω must even be a satisfying assignment. □

Utilizing either an analogous adaptation of the standard construction (Dvořák and Dunne 2018, Reduction 3.6) or (Baumeister, Neugebauer, and Rothe 2018, Theorem 16) we establish the complexity for the remaining semantics.

Theorem 4.3. *The problem* VER-EXPL_σ *is* Π_2^p*-complete for* $\sigma \in \{ad, co, stb, pr\}$.

In order to tackle the verification problem which also asks for minimality of the explanation at hand in an elegant way, we exploit the expressive power of explanations as follows: Consider an AF $F = (A, R)$ and assume $X \subseteq A$. The following gadget $\mathcal{G}_F(S, X) = (A_\mathcal{G}, R_\mathcal{G})$ makes sure that any σ-explanation for X in $F \cup \mathcal{G}_F(S, X)$ necessarily contains $S := \{s_1, \ldots, s_n\}$: We let

$$A_\mathcal{G} = \{g_i, b_i, v_i, c_i \mid i = 1, \ldots, n\}$$

[1] The set S is a strong σ-explanation for X in F iff X is necessarily credulously accepted in the incomplete AF $(S, A \setminus S, R)$ where the arguments in S are definite and the other ones uncertain.

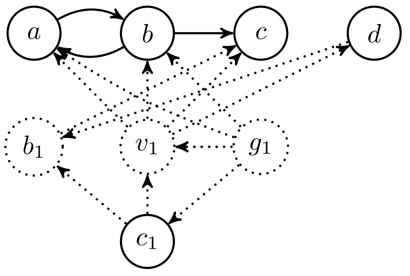

Figure 6: Illustration of the AF $F \cup \mathcal{G}_F(\{d\}, \{c\})$ for $F = (\{a, b, c, d\}, \{(a, b), (b, a), (b, c)\})$.

strategy	σ-extensions	strong σ-expl.
verification gr	in P	coNP-c
verification ad, co, stb	in P	Π_2^p-c
verification pr	coNP-c	Π_2^p-c
ver-min gr	in P	D_1^p-c
ver-min ad and stb	in P	D_2^p-c
ver-min co and pr	coNP-c	D_2^p-c

Table 2: Complexity of explanation strategies.

$$\begin{aligned}R_{\mathcal{G}} = &\{(g_i, c_i), (g_i, v_i), (c_i, b_i), (c_i, v_i) \mid i = 1, \dots, n\} \cup \\ &\{(v_i, a) \mid i = 1, \dots, n,\ a \in A\} \cup \\ &\{(g_i, a) \mid i = 1, \dots, n,\ a \in A \setminus (X \cup \{s_i\})\} \cup \\ &\{(b_i, x) \mid x \in X\} \cup \\ &\{(s_i, b_i) \mid i = 1, \dots, n\} \cup \\ &\{(g_i, g_j), (g_i, b_j), (g_i, v_j) \mid 1 \le j < i, 2 \le i \le n\}\end{aligned}$$

Formally, we obtain the following:

Lemma 4.4. *Let $F = (A, R)$ be an AF and let S be a σ-explanation for X. Let $S' \subseteq S$. Let $S' = \{s_1, \dots, s_n\}$ with $(s_i, s_i) \notin R$ for any i and S' does not attack X. Then $S \cup \{c_i \mid i = 1, \dots, n\}$ is a strong σ-explanation for X in $F \cup \mathcal{G}_F(S', X)$ and for each $s \in S'$, $S \cup \{c_i \mid i = 1, \dots, n\} \setminus \{s\}$ fails to explain X. Moreover, $S \cup \{c_i \mid i = 1, \dots, n\} \setminus \{c_i\}$ fails to explain X for each $i \in \{1, \dots, n\}$.*

Equipped with the gadget $\mathcal{G}$ we are now ready to control whether or not a certain explanation is minimal. In the following, we augment the previous constructions and make sure that all but one additional argument, say $\top$, are necessary in the considered explanations, whereas $\top$ is required iff a certain condition is met. This yields more expressive power and thus D_1^p and D_2^p-hardness, respectively.

Theorem 4.5. *The problem* VER-MIN-EXPL$_{gr}$ *is D_1^p-complete.*

Theorem 4.6. *The problem* VER-MIN-EXPL$_{\sigma}$ *is D_2^p-complete for $\sigma \in \{ad, co, stb, pr\}$.*

A summary of these results can be found in Table 2.

Finally, let us demonstrate how to utilize Theorem 3.14 to compute strong explanations in polynomial time in certain cases. Since gr is robust and hence grounded explanations can be computed in polynomial time (Theorem 3.14), the following can act as a starting point for $\sigma \in \{ad, co, pr\}$.

Proposition 4.7. *Let $F = (A, R)$ be an AF, $X \subseteq A$ a set of arguments and $\sigma \in \{ad, co, pr\}$. If S is a gr-explanation for X, then it is a strong σ-explanation for X.*

Via Proposition 3.12 (item 3), this yields the following tractability result.

Corollary 4.8. *If X is contained in the grounded extension G of F, a σ-explanation of size $|G|$ can be computed in polynomial time for $\sigma \in \{gr, ad, co, pr\}$.*

5 Related Work

Studies of explanations and (computational) argumentation naturally come together; we discuss research related to strong explanations. Fan and Toni (2015b) consider (minimal) removal of sets of arguments or attacks such that the induced subframework credulously accepts an argument to show non-acceptance of arguments. Alfano et al. (2020) study a notion of explanations defined as a sequence of choices (arguments) by considering the strongly connected components of an AF. Explanations within (abstract) argumentation were connected to the notion of dispute trees (Cyras et al. 2018; Fan and Toni 2015a). Admissible extensions, under certain minimality criteria, were studied by Caminada, Dvořák, and Vesic (2016), and correspondences to game-theoretic notions were shown. Abduction (Sakama 2018) reflecting modifications to an AF such that an argument can be labeled in, out or undecided (under labeling-based semantics (Caminada and Gabbay 2009)), and diagnoses (Baumann and Ulbricht 2019) were proposed for AFs. The notion of (belief) revision was applied to explanations in argumentation theory (Rotstein et al. 2008; Falappa, Kern-Isberner, and Simari 2002). So-called critical sets (Booth et al. 2014) were proposed, which are subsets of the arguments in an AF s.t. each (complete) labeling that assigns the same labels to the arguments in the subset also label the remaining arguments the same. Moreover, as discussed before, reasoning in incomplete AFs (Baumeister, Neugebauer, and Rothe 2018) is connected to strong explanations. More broadly, explanations were studied using formal machinery from argumentation (Seselja and Straßer 2013; Rago, Cocarascu, and Toni 2018; Cyras et al. 2019). In contrast, we study strong σ-explanations, which are inspired by recent research on strong explanations in nonmonotonic reasoning (Brewka and Ulbricht 2019), strong inconsistency (Brewka, Thimm, and Ulbricht 2019), and strongly rejecting subframeworks (Saribatur, Wallner, and Woltran 2020; Niskanen and Järvisalo 2020). Our notion differs from these works by focusing on positive acceptance and directly operating on AFs.

6 Conclusions

In this paper we proposed strong explanations as a wide class of explanations for positive (credulous) acceptance for AFs. We showed that, under mild assumptions, argument-based explanation strategies are strong explanations, and we proved the computational complexity of reasoning under strong explanations, indicating a complexity trade-off between strong explanations and extensions.

Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI "ScaDS.AI Dresden/Leipzig", and by the Austrian Science Fund (FWF): 30168-N31.

References

Alfano, G.; Calautti, M.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2020. Explainable Acceptance in Probabilistic Abstract Argumentation: Complexity and Approximation. In *Proc. KR*, 33–43. IJCAI Organization.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards Artificial Argumentation. *AI Magazine* 38(3): 25–36.

Baroni, P.; Caminada, M.; and Giacomin, M. 2011. An introduction to argumentation semantics. *The Knowledge Engineering Review* 26: 365–410.

Baroni, P.; Caminada, M.; and Giacomin, M. 2018. Abstract Argumentation Frameworks and Their Semantics. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baumann, R.; and Ulbricht, M. 2019. If Nothing Is Accepted–Repairing Argumentation Frameworks. *Journal of Artificial Intelligence Research* 66: 1099–1145.

Baumeister, D.; Neugebauer, D.; and Rothe, J. 2018. Credulous and Skeptical Acceptance in Incomplete Argumentation Frameworks. In *Proc. COMMA*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, 181–192. IOS Press.

Bench-Capon, T. J. M.; and Dunne, P. E. 2007. Argumentation in artificial intelligence. *Artificial Intelligence* 171(10-15): 619–641.

Besnard, P.; and Hunter, A. 2008. *Elements of Argumentation*. MIT Press.

Bondarenko, A.; Dung, P. M.; Kowalski, R. A.; and Toni, F. 1997. An Abstract, Argumentation-Theoretic Approach to Default Reasoning. *Artificial Intelligence* 93: 63–101.

Booth, R.; Caminada, M.; Dunne, P. E.; Podlaszewski, M.; and Rahwan, I. 2014. Complexity Properties of Critical Sets of Arguments. In *Proc. COMMA*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, 173–184. IOS Press.

Brewka, G.; Thimm, M.; and Ulbricht, M. 2019. Strong inconsistency. *Artificial Intelligence* 267: 78–117.

Brewka, G.; and Ulbricht, M. 2019. Strong explanations for nonmonotonic reasoning. In *Description Logic, Theory Combination, and All That*, 135–146. Springer.

Caminada, M. 2018. Rationality Postulates: Applying Argumentation Theory for Non-monotonic Reasoning. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*, chapter 15. College Publications.

Caminada, M. W. A.; Dvořák, W.; and Vesic, S. 2016. Preferred semantics as socratic discussion. *Journal of Logic and Computation* 26(4): 1257–1292.

Caminada, M. W. A.; and Gabbay, D. M. 2009. A Logical Account of Formal Argumentation. *Studia Logica* 93(2-3): 109–145.

Cyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2018. Assumption-Based Argumentation: Disputes, Explanations, Preferences. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*, chapter 7, 365–408. College Publications.

Cyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for Explainable Scheduling. In *Proc. AAAI*, 2752–2759. AAAI Press.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2): 321–357.

Dvořák, W.; and Dunne, P. E. 2018. Computational Problems in Formal Argumentation and their Complexity. In Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds., *Handbook of Formal Argumentation*. College Publications.

Falappa, M. A.; Kern-Isberner, G.; and Simari, G. R. 2002. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence* 141(1/2): 1–28.

Fan, X.; and Toni, F. 2015a. On Computing Explanations in Argumentation. In *Proc. AAAI*, 1496–1502. AAAI Press.

Fan, X.; and Toni, F. 2015b. On Explanations for Non-Acceptable Arguments. In *Proc. TAFA, Revised Selected Papers*, volume 9524 of *Lecture Notes in Computer Science*, 112–127. Springer.

García, A. J.; and Simari, G. R. 2004. Defeasible Logic Programming: An Argumentative Approach. *Theory and Practice of Logic Programming* 4(1-2): 95–138.

Modgil, S.; and Prakken, H. 2013. A general account of argumentation with preferences. *Artificial Intelligence* 195: 361–397.

Niskanen, A.; and Järvisalo, M. 2020. Smallest Explanations and Diagnoses of Rejection in Abstract Argumentation. In *Proc. KR*, 667–671.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. In *Proc. IJCAI*, 1949–1955. ijcai.org.

Rotstein, N. D.; Moguillansky, M. O.; Falappa, M. A.; García, A. J.; and Simari, G. R. 2008. Argument Theory Change: Revision Upon Warrant. In *Proc. COMMA*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, 336–347. IOS Press.

Sakama, C. 2018. Abduction in argumentation frameworks. *Journal of Applied Non-Classical Logics* 28(2-3): 218–239.

Saribatur, Z. G.; Wallner, J. P.; and Woltran, S. 2020. Explaining Non-Acceptability in Abstract Argumentation. In *Proc. ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 881–888.

Seselja, D.; and Straßer, C. 2013. Abstract argumentation and explanation applied to scientific debates. *Synthese* 190(12): 2195–2217.