Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning
Main Track

Choices and their Consequences — Explaining Acceptable Sets in Abstract Argumentation Frameworks

Ringo Baumann[1] and **Markus Ulbricht**[1,2]
[1]Leipzig University, Department of Computer Science
[2]TU Wien, Institute of Logic and Computation
{baumann,mulbricht}@informatik.uni-leipzig.de

Abstract

We develop a notion of explanations for acceptance of arguments in an abstract argumentation framework. To this end we show that extensions returned by Dung's standard semantics can be decomposed into i) non-deterministic choices made on even cycles of the given argumentation graph and then ii) deterministic iteration of the so-called characteristic function. Naturally, the choice made in i) can be viewed as an explanation for the corresponding extension and thus the arguments it contains. We proceed to propose desirable criteria a reasonable notion of an explanation should satisfy. We present an exhaustive study of the newly introduced notion w.r.t. these criteria. Finally some interesting decision problems arise from our analysis and we examine their computational complexity, obtaining some surprising tractability results.

1 Introduction

Explainable AI is a highly relevant topic of current research. The ultimate goal is to develop intelligent systems equipped with tools to provide reasons for decisions made and actions taken. Achieving this goal is a key challenge in all areas of AI nowadays as it enables human users to understand artificially intelligent systems. This is inevitable in order to maintain the user's trust in an AI system and hence the system's raison d'être.

These requirements triggered a considerable amount of research, not only for artificial neural networks, but also for several knowledge representation and reasoning formalisms. For example in description logics (Baader, McGuinness, and Nardi 2003) the notions of a *justification* (Horridge et al. 2013) or *pinpointing* (Baader and Peñaloza 2010) the reason for an (undesired) outcome have been introduced and investigated; *explanations* have been studied for Answer Set Programming (Brewka, Eiter, and Truszczynski 2011) in (Dauphin and Satoh 2019) as well as for abstract argumentation frameworks (AFs) (Dung 1995) in e.g. (Saribatur, Wallner, and Woltran 2020). Some approaches also investigate explanations for (non-monotonic) logics in general (Belle 2017; Brewka and Ulbricht 2019).

The present paper is a contribution to explanations in AFs. The field of formal argumentation has become a vibrant research area in Artificial Intelligence. One of the main booster of this development was the seminal paper by Phan Minh Dung in 1995 on abstract argumentation frameworks (AFs). His work is based on the observation that argument evaluation, i.e. the selection of reasonable sets of arguments constituting a coherent world view, can be done without taking into account the internal structure of arguments. Consequently, arguments can be treated as abstract, atomic entities and it suffices to know about the attack relation among the arguments only.

Defining and utilizing explanations in AFs gained quite some attention recently. Several papers view AFs as tools to explain (Zeng et al. 2018; Cocarascu, Cyras, and Toni 2018; Rago et al. 2020) while others propose notions of explanations for acceptable sets of arguments within an AF. For example, based on novel semantics (Fan and Toni 2015), by delving into subframeworks of a given AF (Saribatur, Wallner, and Woltran 2020; Ulbricht and Wallner 2021), or considering the SCCs (Alfano et al. 2020). In this paper, we extend the investigation of the theoretical point of view. As a matter of fact, many mature Dung-style semantics are complete-based. Complete extensions can be characterized as conflict-free fix points of the so-called characteristic function (Dung 1995). Obviously, such a description can be hardly used to explain a certain outcome to an user. However, there is one notable exception in the family of complete semantics, namely the uniquely defined grounded semantics. This semantics can be easily understood as its unique point of view traces back to unattacked arguments. More precisely, unattacked arguments are accepted. Then, further accepted arguments can be obtained given that they are defended by previous ones and so on. Grounded semantics reflects a very skeptical point of view and its acceptance can also be understood in terms of a human-like dialogue (Caminada and Podlaszewski 2012). Our approach for explaining complete extensions is based on three crucial ingredients:

1. We use the easily understandable **grounded semantics** as base line and completion.
2. We make use of the fact that different complete extensions of a given AF are due to different arguments occurring in **even cycles** of the corresponding graph (Dvořák 2012)
3. We utilize the so-called **reduct** of an AF, a simple yet powerful tool that was recently considered to characterize the behavior of AF semantics (Baumann, Brewka, and Ulbricht 2020a).

The main contributions of this paper can be summarized as follows:

- We motivate (Section 3) and introduce (Section 4) so-called explanation schemes which formalize how to decompose any complete extension into their arguments occurring in even cycles and iteration of the characteristic function.
- We utilize the aforementioned schemes in order to develop a notion of explanations for acceptance of arguments in a given AF. We propose desirable criteria for explanations and analyze the notion based on the schemes w.r.t. them (Section 5). We compare our notion to recently proposed ones from the literature.
- We investigate the computational complexity of naturally arising decision problems and show that there is no trade off in comparison to computing extensions (Section 6).

We would like to point out that although the main aim of the paper is to make the outcome (extensions) more understandable for non-experts, we are well aware that there is still a long way to go between theoretical foundations and user-consumable explanations. However, we believe that the following study is an excellent starting point.

2 Background

We fix a non-finite background set $\mathcal{U}$. An AF (Dung 1995) is a directed graph $F = (A, R)$ where $A \subseteq \mathcal{U}$ represents a set of arguments and $R \subseteq A \times A$ models *attacks* between them. For a given $F = (B, S)$ we let $A(F) = B$ and $R(F) = S$. In this paper we consider finite AFs only and we use $\mathcal{F}$ for the set of all these graphs.

For arguments $a, b \in A$, if $(a, b) \in R$ we say that a *attacks* b as well as a *attacks* (the set) E given that $b \in E \subseteq A$. We frequently use the so-called *range* of a set E defined as $E^{\oplus} = E \cup E^{+}$ where $E^{+} = \{a \in A \mid E \text{ attacks } a\}$. The *$E$-reduct of F* is the AF $F^E = (E^*, R \cap (E^* \times E^*))$ where $E^* = A \setminus E^{\oplus}$. This means, F^E is the subframework of F obtained by removing the range of E.

A set $E \subseteq A$ is conflict-free in F (for short, $E \in cf(F)$) iff for no $a, b \in E$, $(a, b) \in R$. We say a set E *defends* an argument a if any attacker of a is attacked by some argument of E. A *semantics* is a function $\sigma : \mathcal{F} \to 2^{2^{\mathcal{U}}}$ with $F \mapsto \sigma(F) \subseteq 2^A$. This means, given an AF $F = (A, R)$ a semantics returns a set of subsets of A. These subsets are called *σ-extensions*. We say that an argument $a \in A$ is *credulously accepted* if $a \in \bigcup \sigma(F)$. Similarly, a is considered as *skeptically accepted* if $a \in \bigcap \sigma(F)$. In case of uniquely defined semantics, i.e. $|\sigma(F)| = 1$ for any F we may simply speak of *accepted* arguments as both notions coincide.

In this paper we consider so-called *admissible*, *complete*, *preferred*, *grounded* and *stable* semantics (abbr. ad, co, pr, gr, stb). All mentioned semantics were already introduced by Dung in 1995 (Dung 1995). For the present paper it will be convenient to utilize the so-called characteristic function Γ_F to define the semantics: Given an AF $F = (A, R)$ and $E \subseteq A$, we have $\Gamma_F(E) = \{a \in A \mid E \text{ defends } a\}$.

Definition 2.1. Let $F = (A, R)$ be an AF and $E \in cf(F)$.

1. $E \in ad(F)$ iff $E \subseteq \Gamma_F(E)$,
2. $E \in co(F)$ iff $E = \Gamma_F(E)$,
3. $E \in pr(F)$ iff E is $\subseteq$-maximal in $co(F)$,
4. $E \in gr(F)$ iff $E = \bigcup_{i \in \mathbb{N}} \Gamma^i{}_F(\emptyset)$,
5. $E \in stb(F)$ iff $E \in cf(F)$ and E attacks any $a \in A \setminus E$.

Since $|gr(F)| = 1$ for any AF F, we will sometimes abuse notion and identify $\{G\}$ with G if $G \in gr(F)$. This way, we may write expressions like "$E \cup gr(F)$" whenever there is no risk of confusion.

Let $F = (A, R)$ be an AF. A sequence $a_1, a_2, \ldots, a_n$ of arguments with $a_i \in A$, $(a_i, a_{i+1}) \in R$ for all $i < n$, and $a_i \neq a_j$ for $i \neq j$ is called a *path* in F. If $a_1, \ldots, a_n$ is a path and $(a_n, a_1) \in R$, then the sequence $a_1, \ldots, a_n, a_1$ is called a *cycle*. If n is even, then the cycle is called an even cycle. By $Ev(F)$ we denote the set of all arguments in F occurring in an even cycle.

We recall the modularization property and a characterization of Dung's classical semantics given in (Baumann, Brewka, and Ulbricht 2020a, Propositions 3.2 and 3.4). Both results will be convenient throughout the present paper.

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

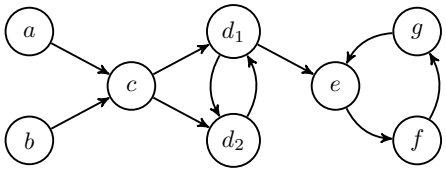
1. *$E \in stb(F)$ iff $F^E = (\emptyset, \emptyset)$,*
2. *$E \in ad(F)$ iff no attacker of E occurs in F^E,*
3. *$E \in pr(F)$ iff no attacker of E occurs in F^E and $\bigcup ad\left(F^E\right) = \emptyset$, and*
4. *$E \in co(F)$ iff no attacker of E occurs in F^E and no argument in F^E is unattacked.*

Proposition 2.3 (Modularization Property)**.** *Given an AF $F = (A, R)$ and $\sigma \in \{ad, co, stb, pr, gr\}$. If $E \in \sigma(F)$ and $E' \in \sigma\left(F^E\right)$, then $E \cup E' \in \sigma(F)$.*

3 Motivation

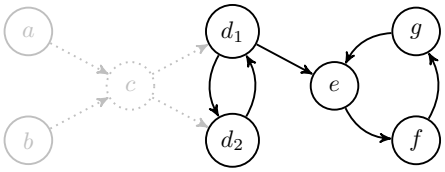
As already stated in the introductory part our approach for explaining complete extensions is based on three essential ingredients, namely grounded semantics, choices on even cycles as well as the application of the recently introduced reduct. Consider the following example.

Example 3.1. Assume F is given as follows:

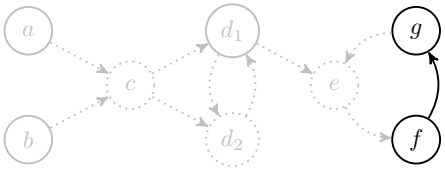


Let us consider complete semantics. According to the fix point characterization as presented in Definition 2.1 we obtain $co(F) = \{\{a, b\}, \{a, b, d_1, f\}, \{a, b, d_2\}\}$. Hence, the argument f is credulously accepted. So, how to explain this acceptance?

First, the most skeptical approach is to accept unattacked arguments (and its consequences), i.e. to compute the grounded extension. Here, we obtain $E_0 = \{a, b\}$ as grounded extension. The reduct F^{E_0} formalizes the situation we face after computing E_0, namely the situation after accepting the grounded extension.



Now, in the second step, we are faced with a choice. We have to decide between choosing d_1 or d_2. As the goal is to explain acceptability of f, let us decide for $E_1 = \{d_1\}$. Accepting E_1 yields the following remaining subframework $\left(F^{E_0}\right)^{E_1}$:



Finally, after the choice and its consequences, we collect all arguments accepted by the most sceptical semantics and we are done. In this case, we obtain $E_2 = \{f\}$ as the grounded extension of $\left(F^{E_0}\right)^{E_1}$. Combining all three steps yields the witnessing set $E = E_0 \cup E_1 \cup E_2 = \{a, b, d_1, f\}$ justifying credulous acceptance of f.

In the previous example, the actual choice that we made was taking d_1 over d_2 yielding the complete extension given as $E = \{a, b, d_1, f\}$ as opposed to $E' = \{a, b, d_2\}$. The aim of this paper is to show that this choice should be interpreted as *explanation* for the acceptance of f.

4 Theoretical Foundations

In the next section, we lay the required theoretical foundations and make some more elaborate observations which are interesting on their own. More specifically, we give a formal definition of so-called explanation schemes. We investigate fundamental properties, in particular we show that an extension $E \in co(F)$ which is not the grounded one needs to contain arguments in $Ev(F)$. Our main result is then that any complete extension can be decomposed in such an explanation scheme, i.e. each $E \in co(F)$ stems from choosing appropriate even cycle arguments and iterating the characteristic function Γ.

4.1 Explanation Schemes

Definition 4.1. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. The triple (E_0, E_1, E_2) is an *explanation scheme* whenever

- $E_0 \in gr(F)$,
- $E_1 \subseteq Ev\left(F^{E_0}\right)$ with $E_1 \in cf(F)$,
- $E_2 \in gr((F^{E_0})^{E_1})$.

If $E_0 \cup E_1 \cup E_2$ defends E_1, then (E_0, E_1, E_2) is a *successful explanation scheme*. If $X \subseteq E_0 \cup E_1 \cup E_2$, then (E_0, E_1, E_2) is a (successful) explanation scheme *for* X.

Note that any explanation scheme is uniquely determined given E_1 as E_0 is the unique grounded extension of F and E_2 the unique grounded extension of $(F^{E_0})^{E_1}$. We will hence sometimes speak of the scheme *induced by* E_1 instead of writing the tuple (E_0, E_1, E_2). Moreover, given E_1 we will sometimes write $E_2(E_1)$ to indicate the functional dependency.

Example 4.2. In our motivating Example 3.1 we found a successful explanation scheme for $X = \{f\}$ which is given as i) $E_0 = \{a, b\}$ being the grounded extension of F, ii) $E_1 = \{d_1\}$ occurs in an even cycle of the reduct F^{E_0}, and iii) $E_2 = \{f\}$ is the grounded extension of $\left(F^{E_0}\right)^{E_1}$.

By definition, E_0, E_1 and E_2 are conflict-free for any scheme induced by E_2. The following lemma states that success is a sufficient condition for $E_1 \cup E_2 \cup E_3$ to be conflict-free as well.

Lemma 4.3. *If an explanation scheme (E_0, E_1, E_2) is successful, then $E_0 \cup E_1 \cup E_2$ is conflict-free.*

The following lemma will play a central role when inferring the main results of this paper. We state it explicitly here as it is interesting on its own. It formalizes the intuition that F^E corresponds to the AF which we obtain from F after setting E to true and E^+ to false. More precisely, it states that if X defends some a in F^E, then $E \cup X$ defends a in F.

Lemma 4.4. *For any AF $F = (A, R)$ and $E \subseteq A$ as well as $X \subseteq A\left(F^E\right)$, $\Gamma_{F^E}(X)$ is the set of arguments in $A \setminus E^{\oplus}$ which is defended by $E \cup X$ in F. In particular, for any integer i, $\Gamma^{i+1}_{F^E}(\emptyset)$ is the set of arguments in $A \setminus E^{\oplus}$ which is defended by $E \cup \Gamma^{i}_{F^E}(\emptyset)$ in F.*

Proof. ($\subseteq$) Let $e \in \Gamma_{F^E}(X)$ and let y be an attacker of e. If y occurs in F^E, then X attacks y. Otherwise, since e occurs in F^E, $y \notin E$ and thus $y \in E^+$. In both cases, $E \cup X$ attacks y. Since y was an arbitrary attacker, $E \cup X$ defends e in F.

($\supseteq$) Assume $E \cup X$ defends $e \in A \setminus E^{\oplus}$. Let y be any attacker of e. If E attacks e, then y does not occur in F^E. Otherwise, some $x \in X$ must attack y. Since $X \subseteq A\left(F^E\right)$, x occurs in F^E. Hence X counterattacks y in F^E. In both cases, X defends e in F^E, i.e. $e \in \Gamma_{F^E}(X)$. □

We also require the following monotonicity result, stating that choosing more arguments results in at least as many arguments in the induced scheme.

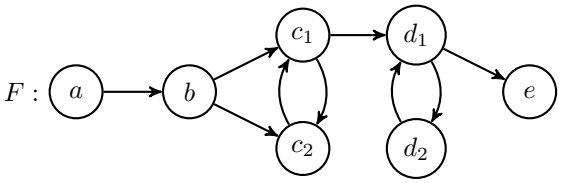
Lemma 4.5. *Let $F = (A, R)$ be an AF. Let E_1 and E_1' induce explanation schemes and let $E_1 \subseteq E_1'$. If the union $E_0 \cup E_1' \cup E_2(E_1')$ is conflict-free, then*

$$E_0 \cup E_1 \cup E_2(E_1) \subseteq E_0 \cup E_1' \cup E_2(E_1').$$

In particular, $E_0 \cup E_1 \cup E_2(E_1)$ is conflict-free as well.

We omit the proof of this lemma due to spaces restrictions. It can be shown by an attentive induction over the number of iterations of Γ and comparing the result on both sides.

Example 4.6. Consider the following AF F.



Consider the explanation schemes induced by $E_1 = \{c_2\}$ and $E_1' = \{c_2, d_2\}$. The first one yields $(\{a\}, \{c_2\}, \emptyset)$ and the second $(\{a\}, \{c_2, d_2\}, \{e\})$. Indeed, we have that $\{a, c_2\} \subseteq \{a, c_2, d_2, e\}$ as predicted by Lemma 4.5.

Our next goal is to prove that any complete extension E of a given AF F can be written as $E_0 \cup E_1 \cup E_2$ for a successful explanation scheme induced by E_1. In this case, we say that E can be *decomposed* into the scheme induced by E_1. To prove this assertion, we need to delve into the structure of complete extensions, especially with regard to even cycles:

Proposition 4.7. *[The role of choices] Let $E \in co(F)$. If E is not the grounded extension of F, then there must be some $a \in E$ with $a \in Ev(F)$.*

Proof. Let $E, E' \in co(F)$ and assume none of them contains an argument occurring in an even cycle. If F' is the AF after turning each argument in an even cycle into a self-attacker, then $E, E' \in co(F')$ as well. However, now it is impossible to defend these arguments (except by attacking them). Thus if F'' is the AF F' after removing all attacks within even cycles, then $E, E' \in co(F'')$, too. Since F'' is even-cycle free, we find $E = E'$ (Dvořák 2012, Proposition 15), i.e. there is only one complete extension containing no such arguments. The only candidate is $gr(F)$. □

Example 4.8. Recall Example 4.6. The complete extensions are $\{a\}$, $\{a, d_2, e\}$ $\{a, c_1, d_2, e\}$, $\{a, c_2\}$, $\{a, c_2, d_1\}$, and $\{a, c_2, d_2, e\}$ where only the grounded extension $\{a\}$ does not contain arguments in $Ev(F) = \{c_1, c_2, d_1, d_2\}$.

We also need the following auxiliary result which shows how to decompose complete extensions w.r.t. the reduct.

Lemma 4.9. *Let $F = (A, R)$ be an AF and let $E \in co(F)$. For any $E' \subseteq E$ there is a set E'' satisfying i) $E' \cup E'' = E$ and ii) $E'' \in co\left(F^{E'}\right)$.*

Proof. The only candidate is $E'' = E \setminus E'$. Clearly, i) is satisfied, so we have to show that E'' is complete in $F^{E'}$. To this end we note that E'' is trivially conflict-free and contains all arguments it defends due to completeness of $E = E' \cup E''$ and the fact that $F^E = (F^{E'})^{E''}$. More precisely, F^E does not contain unattacked arguments and hence, the same applies to $(F^{E'})^{E''}$. We have thus left to show that E'' defends itself in $F^{E'}$. However, $E' \cup E''$ defends E'' and hence, this assertion can be inferred from Lemma 4.4. □

We are now ready to prove the main theorem of this section stating that each complete extension can be found via a successful explanation scheme and vice versa.

Theorem 4.10. *Given an AF $F = (A, R)$ and a set of arguments $E \subseteq A$. We have: $E \in co(F)$ iff E can be decomposed into a successful explanation scheme.*

Proof. ($\Leftarrow$) E_0 and E_2 are defended against all their attackers by definition and due to Lemma 4.4, respectively, and E_1 is defended since the explanation scheme is successful. Hence $E_0 \cup E_1 \cup E_2$ is admissible. Due to the requirement $E_2 \in gr((F^{E_0})^{E_1})$, it must be complete as well.

($\Rightarrow$) Let $E \in co(F)$. Consider the following decomposition:

- E_0 is the grounded extension of F,
- $E_1 = Ev\left(F^{E_0}\right) \cap E$,
- E_2 is the grounded extension of $(F^{E_0})^{E_1}$.

We have to show that $E_0 \cup E_1 \cup E_2 = E$.

($\subseteq$) $E_0 \subseteq E$ is a classical observation and $E_1 \subseteq E$ is enforced by definition. Now assume we are given $E_0 \cup E_1$ and compute the grounded extension G of $F^* = F^{E_0 \cup E_1}$. Assume "$\subseteq$" is wrong, i.e. there is some $a \in G \setminus E$. Let us assume a is chosen in a way that i) $a \in \Gamma_{F^*}^{i+1}(\emptyset)$ and ii) there is no $a' \in \Gamma_{F^*}^{j}(\emptyset)$ with $a' \notin E$ s.t. $j < i + 1$. By Lemma 4.4, a is defended by $E_0 \cup E_1 \cup \Gamma_{F^*}^{i}(\emptyset)$ in F; by the assumption imposed on i we have $\Gamma_{F^*}^{i}(\emptyset) \subseteq E$ which yields $E_0 \cup E_1 \cup \Gamma_{F^*}^{i}(\emptyset) \subseteq E$. By monotonicity of Γ, E defends a in contradiction to $E \in co(F)$ and $a \notin E$.

($\supseteq$) By Lemma 4.9, there is some complete extension of $F^* = F^{E_0 \cup E_1}$ s.t. the union with E_0 and E_1 yields E. Now if $E_0 \cup E_1 \cup E_2 \subsetneq E$, i.e. $E_2 \subsetneq E \setminus (E_0 \cup E_1)$, then E_2 must be another complete extension of F^* than the grounded one. However, by Proposition 4.7, this contradicts the fact that $E_0 \cup E_1$ contains all even cycle arguments of E. □

Example 4.11. Recall Example 3.1. The reader may verify $co(F) = \{\{a, b\}, \{a, b, d_2\}, \{a, b, d_1, f\}\}$. We already saw in Example 4.2 that $\{a, b, d_1, f\}$ corresponds to the scheme induced by $E_1 = \{d_1\}$. The grounded extension $\{a, b\}$ expectedly can be found via the scheme induced by $E_1' = \emptyset$. Finally, letting $E_1'' = \{d_2\}$ yields a scheme for $\{a, b, d_2\}$.

Example 4.12. Now consider again Example 4.8. We found the complete extensions

$\{a\}$	$\{a, c_1, d_2, e\}$	$\{a, c_2\}$
$\{a, c_2, d_1\}$	$\{a, c_2, d_2, e\}$	$\{a, d_2, e\}$

They decompose into the schemes

$(\{a\}, \emptyset, \emptyset)$	$(\{a\}, \{c_1, d_2\}, \{e\})$	$(\{a\}, \{c_2\}, \emptyset)$
$(\{a\}, \{c_2, d_1\}, \emptyset)$	$(\{a\}, \{c_2, d_2\}, \{e\})$	$(\{a\}, \{d_2\}, \{e\})$

as predicted by the above Theorem 4.10.

Since all complete extensions can be attained via successful explanation schemes, we find that X can be explained iff it is contained in at least one admissible extension.

Corollary 4.13. *There is a successful explanation scheme for X iff $X \subseteq E \in ad(F)$.*

We want to mention that this also entails that there is an explanation scheme for each stable extension of F.

Corollary 4.14. *If $X \subseteq E \in stb(F)$, then there is a successful explanation scheme for X.*

4.2 Further Variants of Explanation Schemes

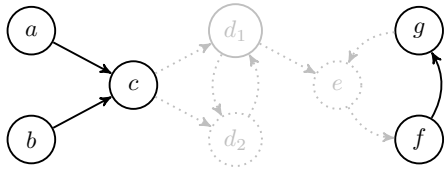
Let us now continue with another notion of explanation schemes. Since E_0 is the grounded extension of F anyway, one might wonder whether the first step can be skipped, moving to choosing E_1 immediately. This yields the following kind of schemes.

Definition 4.15. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. The tuple (E_1, E_2) is an *abbreviated explanation scheme* whenever

- $E_1 \subseteq Ev(F)$ with $E_1 \in cf(F)$,
- $E_2 \in gr\left(F^{E_1}\right)$.

If $E_1 \cup E_2$ defends E_1, then it is *successful*. If $X \subseteq E_1 \cup E_2$, it is a (successful) abbreviated explanation scheme *for* X.

Example 4.16. Let us reconsider Example 3.1 and recall $E_1 = \{d_1\}$:



Computing $gr\left(F^{E_1}\right)$ yields $E_2 = \{a, b, f\}$, i.e. (E_1, E_2) is a successful abbreviated scheme for $X = \{f\}$.

The following proposition formalizes that any explanation scheme can be simulated by a canonical abbreviated one:

Proposition 4.17. *Consider a given explanation scheme (E_0, E_1, E_2) and the abbreviated one (E_1, E_2'), then we have that $E_0 \cup E_1 \cup E_2 = E_1 \cup E_2'$.*

Vice versa, abbreviated explanation schemes may ignore or even attack the grounded extension of an AF.

Example 4.18. Let $F = (A, R)$ where $A = \{a, b, c\}$ and $R = \{(a, b), (a, c), (b, c), (c, b)\}$. Clearly, $gr(F) = \{\{a\}\}$. However, an abbreviated scheme may set $E_1 = \{b\}$ resulting in an unsuccessful one since $gr(F^{E_1}) = \{\{a\}\}$ attacks b.

The observation we just made holds in general.

Lemma 4.19. *If for the scheme (E_1, E_2), $E_1 \cup E_2$ is in conflict with the grounded extension G of F, then the scheme is unsuccessful.*

However, if our abbreviated scheme is successful, then we can show a counterpart to Proposition 4.17, that is, we find an explanation scheme inducing the same set of arguments.

Proposition 4.20. *If (E_1, E_2) is a successful abbreviated scheme, then there is a corresponding scheme (E_0, E_1', E_2') s.t. $E_0 \cup E_1' \cup E_2' = E_1 \cup E_2$.*

We thus infer the following main theorem.

Theorem 4.21. *A set $E \subseteq A$ is in $co(F)$ iff E can be decomposed into an abbreviated successful explanation scheme.*

Example 4.22. Recall Examples 4.8 and 4.12. We found the complete extensions

$$\{a\} \quad \{a, c_1, d_2, e\} \quad \{a, c_2\}$$
$$\{a, c_2, d_1\} \quad \{a, c_2, d_2, e\} \quad \{a, d_2, e\}.$$

They decompose into the abbreviated schemes

$$(\emptyset, \{a\}) \quad (\{c_1, d_2\}, \{a, e\}) \quad (\{c_2\}, \{a\})$$
$$(\{c_2, d_1\}, \{a\}) \quad (\{c_2, d_2\}, \{a, e\}) \quad (\{d_2\}, \{a, e\}).$$

One may wonder whether checking the extension in hindsight is really necessary. In principle, this is not the case, given that E_1 is admissible (modularization property). We thus consider the following version.

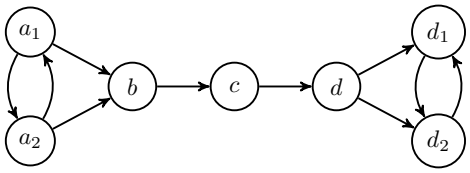
Definition 4.23. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. The triple (E_0, E_1, E_2) is a *secure explanation scheme* whenever

- $E_0 \in gr(F)$,
- $E_1 \subseteq Ev\left(F^{E_0}\right)$ s.t. $E_1 \in ad(F^{E_0})$,
- $E_2 \in gr((F^{E_0})^{E_1})$.

If $X \subseteq E_0 \cup E_1 \cup E_2$, then (E_0, E_1, E_2) is a secure explanation scheme *for* X.

In this case, however, not all complete extensions can be attained.

Example 4.24. Consider the following AF F:



The reader may verify that $\{a_1, d_1\}$ induces a successful scheme for F, namely $(\emptyset, \{a_1, d_1\}, \{c\})$ where we see that c defends d_1 against d. However, since c is required to defend d_1 and does not occur in any even cycle, there is no secure scheme which attains the complete extension $\{a_1, c, d_1\}$.

Within the scope of this paper, this is sufficient evidence to not continue investigating secure schemes, although we are convinced that they are an interesting topic of future work as well.

5 Excursus: Admissible Extensions and Explanation Schemes

The goal of this section is to extend the results we obtained so far to admissible extensions. As it turns out, they can also be computed by selecting even cycle arguments and iterating a suitable operator. First we require a version of Proposition 4.7 adapted to admissible semantics. This will be based on the notion strong admissibility, firstly introduced in (Baroni and Giacomin 2007) and further studied in (Caminada and Dunne 2019; Baumann, Linsbichler, and Woltran 2016). We will define it according to (Baumann, Linsbichler, and Woltran 2016, Definition 7).

Definition 5.1. Let $F = (A, R)$ be an AF. A set $E \subseteq A$ is strongly admissible ($E \in ad^s(F)$) if there are finitely many pairwise disjoint $A_1, \ldots, A_n$ s.t. $E = \bigcup_{1 \leq i \leq n} A_i$ with $A_1 \subseteq \Gamma_F(\emptyset)$ and $\bigcup_{1 \leq i \leq j} A_i$ defends A_{j+1}.

Now we introduce a variant of explanation schemes for admissible sets. Thereby, we formalize a decomposition into strongly admissible extensions and non-deterministic choices in the even cycles in F.

Definition 5.2. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. The tuple (E_1, E_2) is a *strong admissibility-based explanation scheme* whenever

- $E_1 \subseteq Ev(F)$ with $E_1 \in cf(F)$,
- $E_2 \in ad^s\left(F^{E_1}\right)$.

If $E_1 \cup E_2$ defends E_1, then it is successful. If $X \subseteq E_1 \cup E_2$, it is a (successful) strong admissibility-based explanation scheme *for* X.

The following result can be viewed as a strong admissible based counterpart to Proposition 4.7. It states that an admissible extension which does rely on choices made in even cycles must be strongly admissible.

Proposition 5.3. *Let $F = (A, R)$ be an AF. If $E \in ad(F)$ satisfies $E \cap Ev(F) = \emptyset$, then $E \in ad^s(F)$.*

Proof. Assume E is admissible. We let E' be a maximal subset of E with $E' \in ad^s(F)$. By maximality (and due to Definition 5.1), E' does not defend any argument in the difference $E'' = E \setminus E'$. The rest is similar to the proof of (Dvořák 2012, Proposition 15), adjusted to our setting. Let $e_1 \in E''$. Since e_1 is not defended by E', there is some $s_1 \notin (E')^+$ attacking e_1. Since $E \cap Ev(F) = \emptyset$, e_1 cannot counterattack s_1. Now assume for any integer $i > 0$ there is some $s_i \notin (E')^+$ attacking e_i. Then $\{e_1, \dots e_i\}$ cannot attack s_i since this would induce an even cycle. There is thus some $e_{i+1} \in E'' \setminus \{e_1, \dots e_i\}$ attacking s_i. Since E' does not defend e_{i+1}, there is an attacker s_{i+1} of e_{i+1} not attacked by E'. By finiteness, this procedure must eventually fail. We conclude that choosing $e_1 \in E''$ must be impossible, i.e. $E'' = \emptyset$ and hence, $E = E' \in ad^s(F)$. □

As the reader may already predict, the next step is a counterpart to Lemma 4.9.

Lemma 5.4. *Let $F = (A, R)$ be an AF and let $E \in ad(F)$. For any $E' \subseteq E$ there is a set E'' satisfying i) $E' \cup E'' = E$ and ii) $E'' \in ad\left(F^{E'}\right)$.*

Having established these two results, we are now ready to infer the following characterization of admissible sets.

Theorem 5.5. *A set $E \subseteq A$ is in $ad(F)$ iff E can be decomposed into a strong admissibility-based explanation scheme.*

Proof. ($\Rightarrow$) Given $E \in ad(F)$ we let $E_1 = E \cap Ev(F)$ and apply Lemma 5.4 to $E_1 \subseteq E$: It must be the case that $E \setminus E_1 \in ad\left(F^{E_1}\right)$. We have to show that $E \setminus E_1$ is even strongly admissible in F^{E_1}. However, if not, then E_1 does not contain all even cycle arguments of E due to Proposition 5.3. Hence $(E_1, E \setminus E_1)$ is the desired scheme.

($\Leftarrow$) Let (E_1, E_2) be such a successful scheme. E_2 is defended against all its attackers due to Lemma 4.4 and E_1 is defended since the explanation scheme is successful. Hence $E_1 \cup E_2$ is admissible. □

6 Choices and Explanations

In this section we make use of the theoretical investigation we performed so far and consider a novel notion of explanations for a set X of arguments which builds upon our notion of explanation schemes. In a nutshell, we formalize that the non-determinism in our extension is the part that should be explained since iterating the grounded extensions does not yield unexpected results. We also compare the notion we obtain with two recently introduced ones from (Alfano et al. 2020) as well as (Ulbricht and Wallner 2021). The latter paper also proposed some desirable criteria for explanations in order to compare them on an abstract level. We extend this list and perform a comparison of the three notions, revealing that they all provide different points of view.

Let us now turn to the notion of an explanation. Our goal is to formalize the intuition that an explanation for $X \subseteq A$ is a set S of arguments justifying acceptance of the set X in the given AF $F = (A, R)$. Since there might exist multiple explanations for X, we consider sets $\mathcal{S} \subseteq 2^A$ with $S \in \mathcal{S}$ iff S is an explanation.

Definition 6.1. Let $F = (A, R)$ be an AF, σ any semantics and $X \subseteq A$. An *explanation strategy* for X in F w.r.t. σ is a set $\mathcal{S} \subseteq 2^A$. A set $S \in \mathcal{S}$ is called an explanation.

Note that we do not impose any properties on explanations other than being sets of arguments. The reason is that in principle, any subset of 2^A should be a possible strategy. Whether or not it is a *reasonable* one shall be decided by inspecting the desirable criteria we develop below.

The most natural kind of explanations are extensions.

Example 6.2. For an AF $F = (A, R)$, a semantics σ, and $X \subseteq A$ the set $\mathcal{S} = \{E \in \sigma(F) \mid X \subseteq E\}$ is an explanation strategy for X w.r.t. σ. Then, each $E \in \sigma(F)$ satisfying $X \subseteq E$ is an explanation.

Motivated by the observation that computing the grounded extension is deterministic, an explanation scheme induces an explanation as follows.

Definition 6.3. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. If (E_0, E_1, E_2) is a successful explanation scheme for X, then E_1 is called an *scheme-based explanation for X*.

Note that our notion of an explanation is independent of the considered semantics.

Example 6.4. Recall Example 3.1. As we have now formally established, $\{d_1\}$ is a minimal scheme-based explanation for $\{f\}$ in the given AF F.

In order to broaden our investigation, let us also consider two other recently introduced definitions of explanations from the literature. First, we define strong explanations (Ulbricht and Wallner 2021, Definition 3.9). The idea here is as follows: Given a knowledge base $\mathcal{K}$ of a monotonic logic and some ϕ entailed by $\mathcal{K}$, it is common to view a minimal subset of $\mathcal{K}$ entailing ϕ as the reason for ϕ being entailed. This approach does not make sense for non-monotonic logics though and hence the following adjustment is proposed in this recent paper:

Definition 6.5. 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 some $E' \in \sigma(F')$ with $X \subseteq E'$.

That is, instead of considering only one subframework of F s.t. X occurs in an extension, the condition is refined in order to impose a monotonic behavior.

Our approach is similar in its spirit to (Alfano et al. 2020, Definition 4). For this, recall that a strongly connected component (SCC) of an AF is a set $S \subseteq A$ s.t. in the subframework $F{\downarrow_S}$, for each two arguments $a, b \in S$ there is some (possibly empty) path from a to b in the directed graph $F{\downarrow_S}$; an SCC S is called initial if the arguments in S are not attacked by any argument occurring in some other SSC. The idea in (Alfano et al. 2020) is to recursively consider the

grounded extension of a the current AF and then perform some choice in an initial SCC in order to continue the iteration. The authors simply speak of "explanations", we will call them recursion-based explanations here in order to avoid the generic term in presence of other notions.

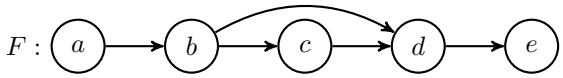
Definition 6.6. Let $F = (A, R)$ be an AF and let σ be any semantics. Let G be the grounded extension of F. For some $a \in A$, let $a^- = \{b \in A \mid (b, a) \in R\}$. We call a sequence $S = (a_1, \ldots, a_n)$ a *recursion-based extension explanation* for E w.r.t. F if either $n = 0$ and $E = G$ or

- a_1 belongs to some initial SCC of F^G,
- $(a_2, \ldots, a_n)$ is an explanation for $E \setminus G$ w.r.t the AF F^G without incoming attacks of a_1, i.e. $\left(F^G\right)\downarrow_{A \setminus a_1^-}$.

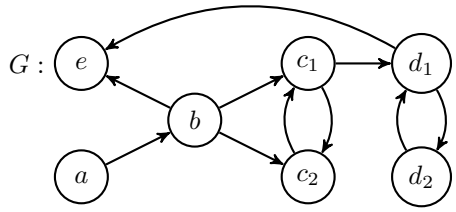
Given $X \subseteq A$, the set $\{a_1, \ldots, a_n\}$ is a *recursion-based explanation* for X if (there is some order s.t.) the sequence $(a_1, \ldots, a_n)$ is a recursion-based extension explanation for some E with $X \subseteq E$.

Let us perform a quick comparison of the three notions. As the following example shows, they are incomparable in general.

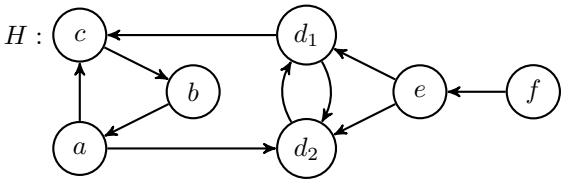
Example 6.7. Consider the following AF taken from (Ulbricht and Wallner 2021):



With reasonable effort we verify that $S = \{a, c, e\}$ is a strong explanation for $X = \{e\}$, but the notions of scheme-based and recursion-based explanations agree on the empty explanation since the given AF is acyclic. Now consider a modification of the AF from Example 4.6, called G here:



We have the scheme-based explanation $E_1 = \{c_1, d_2\}$ for $\{e\}$, but (c_1, d_2) is no recursion-based explanation since d_2 does not occur in an SCC of the reduct $G^{\{c_1\}}$. By consideration of the subframework induced by the arguments in G excluding a we see that this is no strong explanation, either. Finally, let H be the following AF:



We have that $S = (b, d_1)$ is a recursion-based explanation for $X = \{b\}$ since it induces the extension $E = \{b, d_1\}$. However, since $b \notin Ev(F)$, $\{b, d_1\}$ is no scheme-based explanation. Considering the subframework without f shows that $\{b, d_1\}$ is no strong explanation, either.

In the following, we want to compare the properties of our explanation notions on an abstract level. To this end we consider criteria proposed in (Ulbricht and Wallner 2021) and we develop some further general desirable properties an explanation should respect.

Let us start with the ones that have already been considered.

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

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

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

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

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

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

The σ-existence property simply ensures that there is an explanation for X iff X is accepted w.r.t. σ. The second property, σ-basic, formalizes the intuition that the reduct F^S can be interpreted as setting the arguments in S to true and inspecting the remaining AF. In this case, an explanation S can only be considered successful if X is then rendered acceptable. Monotonicity formalizes the idea that whenever S suffices to explain X, then so should every superset of S. This somewhat contradicts the idea behind CF and Defense, stating that an explanation should be an acceptable point of view on its own. Finally, independence requires that an explanation does not rely on other arguments than the explanation itself.

Extending basic to the other common reasoning mode, we also consider a strengthened version where skeptical acceptance is required. We exclude $\sigma = ad$ here.

σ-skeptical $S \in \mathcal{S}$ implies $X \subseteq E \cup S$ for each extension $E \in \sigma(F^S)$.

As a weaker version of σ-skeptical we consider the following property where skeptical acceptance of X is assumed.

Skeptical Acceptance If X is skeptically accepted w.r.t. σ and S is an explanation for X, then $X \subseteq S \cup E$ for each $E \in \sigma(F^S)$.

The last property we are going to consider formalizes a compatibility requirement for notions of explanations; as long as they are not conflicting, they can be merged to create novel explanations.

Compatibility If S and S' are explanations for X and X', respectively, with $S \cup S' \in cf(F)$, then $S \cup S'$ is an explanation for $X \cup X'$.

The following theorem summarizes which properties are satisfied and which not. The high level observation is that all proposed notions provide different points of view. We want to mention that dis-satisfaction of some of these properties does not always mean that the corresponding notion behaves poorly. For example scheme- and recursion-based explanations do not satisfy independence since removing an argument might disrupt some even cycle or SCC, respectively. Moreover, the three notions are not designed to defend themselves, but to facilitate a certain target set.

stategy	exten-sion	strong expl.	scheme expl.	recurs. expl.
σ-existence	✓	✓	✓	✓
σ-basic	✓	✓	✓	✓
Monotonicity	×	✓	×	×
CF	✓	×	✓	✓
Defense	✓	×	×	×
Independence	$\{ad, stb\}$	✓	×	×
σ-skeptical	✓	$\{gr\}$	✓	✓
Skept. Acc.	✓	$\{gr\}$	✓	✓
Compatibility	✓	✓	✓	×

Table 1: Summary of properties of explanation strategies for $\sigma \in \{ad, co, gr, stb, pr\}$. Results highlighted in gray are taken from (Ulbricht and Wallner 2021)

Theorem 6.8. *Satisfaction of the properties discussed above is as depicted in Table 1.*

Sketch of Proof. (Dis-)satisfaction of the properties is mostly due to rather simple considerations or counterexamples regarding the explanation notions. Note that scheme-based and recursion-based explanations do not satisfy Independence since removing an argument might disrupt an even cycle or SCC, respectively. For the same reason, recursion-based explanation do not satisfy Compatibility. Besides compatibility, the most interesting case is σ-skeptical. Regarding scheme-based and recursion-based explanations we note that F^S is s.t. $X \subseteq E \cup S$ for the grounded extension E of F^S. Hence the claim holds for any semantics under consideration; moreover, this also shows satisfaction of Skeptical Acceptance. Extensions satisfy σ-skeptical since $X \subseteq S$ must already hold. □

We want to state satisfaction of Compatibility by scheme-based explanations as a result on its own. The reason is that this observation will be very important later on and we hence want to state it explicitly here. The proof of this assertion can be found in the supplementary material.

Theorem 6.9. *If E_1 and E_1' are successful explanations for X and X', respectively, with $E_1 \cup E_1' \in cf(F)$, then $E_1 \cup E_1'$ is successful explanation for $X \cup X'$.*

7 Computational Complexity

In this section we examine the computational complexity of natural decision problems induced by our notion of scheme-based explanations. From a computational point of view, we are interested in whether we are given a minimal subset s.t. a certain extension is induced.

Definition 7.1. Let $F = (A, R)$ be an AF. Let $X \subseteq A$. If E_1 is minimal s.t. (E_0, E_1, E_2) is a successful explanation scheme for X, then E_1 is called a minimal *scheme-based explanation for X*.

On the contrary, looking for preferred and stable extensions in particular, we might consider maximal sets E_1 s.t. $E_1 \subseteq Ev(F)$ is conflict-free.

Definition 7.2. An explanation scheme (E_0, E_1, E_2) is *maximal* if there is no E_1' with $E_1 \subsetneq E_1'$ s.t. E_1' induces an explanation scheme.

Having formally established these notions, we consider the following problems: i) verifying that a given set is a (minimal) explanation for X, ii) checking whether there is some explanation for X, and motivated by the fact that schemes may or may not be successful we also discuss the complexity of deciding whether iii) all explanation schemes are successful and iv) all maximal schemes are.

VER-EXPL
Input: (F, E_1, X) where $F = (A, R)$ and $E_1, X \subseteq A$
Output: TRUE iff E_1 is an explanation for X in F

VER-MIN-EXPL
Input: (F, E_1, X) where $F = (A, R)$ and $E_1, X \subseteq A$
Output: TRUE iff E_1 is a minimal explanation for X in F

EXIST-EXPL
Input: (F, X) where $F = (A, R)$ is an AF and $X \subseteq A$
Output: TRUE iff there is an explanation E_1 for X

ALL-SAFE
Input: An AF F
Output: TRUE iff all schemes are successful

MAX-SAFE
Input: An AF F
Output: TRUE iff all maximal schemes are successful

First we want to mention that due to Theorem 4.21, deciding whether there exists a scheme-based explanation for X is equivalent to deciding whether there is a complete extension containing X. Hence:

Theorem 7.3. *The problem* EXIST-EXPL *is* NP*-complete.*

Verifying an explanation only requires computing the reduct and grounded extension of a given AF. This is of course a tractable task.

Theorem 7.4. *The problem* VER-EXPL *can be solved in* P.

Additionally verifying minimality requires an algorithm which successively removes arguments to search for a smaller successful scheme for a given set X. This can still be done in P.

Theorem 7.5. *The problem* VER-MIN-EXPL *can be solved in* P.

We refer the reader to the supplementary material for a formal proof of this assertion. Although checking whether all explanation schemes are successful seems to be a demanding task, we can utilize Theorem 6.9 to significantly reduce our search space. The following lemma formalizes this observation.

Lemma 7.6. *If $F = (A, R)$ is an AF and there is any unsuccessful explanation scheme (E_0, E_1, E_2) with conflict-free E_1, then there is an unsuccessful one of the form $(E_0, \{e_1\}, E_2(\{e_1\}))$, i.e. it is induced by a singleton.*

Proof. Due to Theorem 6.9 applied to $X = X' = \emptyset$, if all possible schemes of the form $(E_0, \{e_1\}, E_2(\{e_1\}))$ are successful, then by induction, all schemes are. □

Hence the following tractability result can be shown.

Theorem 7.7. *The problem* ALL-SAFE *can be solved in polynomial time.*

Proof. By Lemma 7.6 it suffices to iterate over the linearly many $e_1 \in Ev(F^{E_0})$ after computing the grounded extension E_0 of F. For each e_1 another iteration of computing the grounded extension has to be performed, all of which can be done in P. □

However, checking only the maximal ones is intractable since we can not utilize the shortcut from Lemma 7.6 anymore. To prove this, we adapt the standard translation (Dvorák and Dunne 2018, Reduction 3.6) in a straightforward fashion. As an aside, we want to mention that within the standard translation, an explanation for φ corresponds exactly to a satisfying assignment of the given formula. Thus, our notion of explanations yields a very natural set of arguments here.

Theorem 7.8. *The problem* MAX-SAFE *is* coNP*-complete.*

Proof. For hardness, let us assume we are given a formula $\Phi = \exists X \phi(X)$ with $\phi = \{C_1, \dots, C_r\}$ in CNF over variables in $X = \{x_1, \dots, x_n\}$. We adapt the well-known standard translation (see e.g. (Dvorák and Dunne 2018, Reduction 3.6)): We let (cf. Figure 1)

$$A = \{\varphi\} \cup \{\top\} \cup \{\top'\} \cup \{c \mid c \in \phi\} \cup \{x, \bar{x} \mid x \in X\}$$

and the set R of attacks is given via

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

We claim that there is a maximal scheme which is unsuccessful iff Φ is true.

($\Rightarrow$) Assume Φ is true. Let $\omega : X \to \{0, 1\}$ be a satisfying assignment for ϕ. Let

$$X_\omega = \{x_i \mid \omega(x_i) = 1\} \cup \{\bar{x}_i \mid \omega(x_i) = 0\}$$

be the corresponding canonical set of X arguments and consider the set $E_1 = X_\omega \cup \{\top\}$. Trivially, all of them occur in an even cycle. Let us now compute $gr(F^{E_1})$. Due to ω being a satisfying assignment, standard considerations yield $\varphi \in \Gamma_{F^{E_1}}(\emptyset)$. Hence we found an unsuccessful scheme due to φ attacking $\top$.

($\Leftarrow$) Now assume Φ is false. Clearly, for any maximal cf-scheme (E_0, E_1, E_2), the set E_1 is of the form

$$\text{either} \quad E_1 = X_\omega \cup \{\top\} \quad \text{or} \quad E_1 = X_\omega \cup \{\top'\}$$

with X_ω corresponding to an assignment as above. Since the set $X_\omega \cup \{\top'\}$ defends itself, there is nothing to be shown. So assume we are given $E_1 = X_\omega \cup \{\top\}$. Since Φ is false, ω is no satisfying assignment and hence there is some c_i which is not attacked by X_ω. By maximality of E_1, c_i is even defended by X_ω and thus, $c_i \in \Gamma_{F^{E_1}}(\emptyset)$ and therefore φ is attacked by $E_1 \cup E_2(E_1)$ implying $\top$ is defended, i.e. the scheme is successful. □

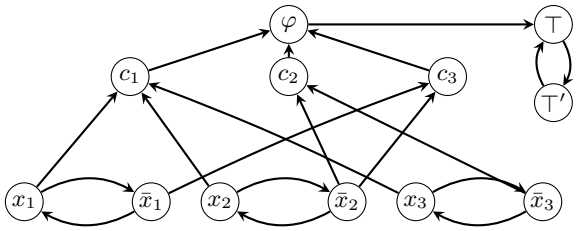


Figure 1: Illustration of the AF F from Theorem 7.8, applied to ϕ with clauses $\{\{x_1, x_2, x_3\}, \{\bar{x}_2, \bar{x}_3\}, \{\bar{x}_1, \bar{x}_2\}\}$.

Summarizing our complexity results, the overall picture is that most of the arising decision problems regarding our explanation schemes are either equivalent to standard reasoning problems for AFs or mostly tractable. The only exception is when insisting on the maximal schemes.

8 Conclusion and Future Work

In this paper we developed a notion of explanations for AFs based on the observation that the non-determinism in Dung's standard semantics can be traced back to the even cycles occurring in an AF. We formalized this observation and compared different notions of so-called explanation schemes to delve into the structure of complete extensions. Furthermore, we showed that a natural notion of explanations induced by these results is well-behaving w.r.t. reasonable criteria we proposed. Finally, we investigated the computational complexity of related decision problems.

While other works on explanations view AFs as tool to explain decisions of AI systems (Zeng et al. 2018; Cocarascu, Cyras, and Toni 2018; Rago et al. 2020), the present paper focused on the inner structure of AFs themselves. The closest to our work are probably (a) the explanation notion in the paper (Alfano et al. 2020) which is similar in spirit to our notion, but relies on a recursive definition and the SCCs of the given AF instead of pinpointing the even cycle arguments and (b) two recent publications on explaining (non-)acceptability (Saribatur, Wallner, and Woltran 2020; Ulbricht and Wallner 2021), where the authors take any subframework of a given AF into consideration, motivated by recent work on inconsistency in AFs (Brewka, Thimm, and Ulbricht 2019; Baumann and Ulbricht 2018). In comparison, our approach focuses solely on the given AF at hand and thus does not insist on certain properties of subframeworks. We utilize the reduct however, which proved to be an efficient tool to examine the interaction between arguments (Baumann, Brewka, and Ulbricht 2020b; 2020a).

The paper induces several future work directions. We believe the most exciting ones are: i) algorithm design and implementation of our approach which allow to cope with acceptability in terms of complete semantics in static as well as dynamic environments (Gaggl et al. 2020), ii) an in-depth comparison to other notions of explanations, and iii) investigate whether this decomposition of extensions works analogously for other related formalisms as well, e.g. Answer Set Programming or Reiter's Default Logic.

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) through project Y698.

References

Alfano, G.; Calautti, M.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2020. Explainable Acceptance in Probabilistic Abstract Argumentation: Complexity and Approximation. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, 33–43.

Baader, F., and Peñaloza, R. 2010. Axiom pinpointing in general tableaux. *Journal of Logic and Computation* 20(1):5–34.

Baader, F.; McGuinness, D. L.; and Nardi, D. 2003. *The description logic handbook: theory, implementation, and applications*. Cambridge university press.

Baroni, P., and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence* 171:675–700.

Baumann, R., and Ulbricht, M. 2018. If nothing is accepted - repairing argumentation frameworks. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR*, 108–117.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2020a. Comparing Weak Admissibility Semantics to their Dung-style Counterparts – Reduct, Modularization, and Strong Equivalence in Abstract Argumentation. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, 79–88.

Baumann, R.; Brewka, G.; and Ulbricht, M. 2020b. Revisiting the foundations of abstract argumentation: Semantics based on weak admissibility and weak defense. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2742–2749. AAAI Press.

Baumann, R.; Linsbichler, T.; and Woltran, S. 2016. Verifiability of argumentation semantics. In *Proceedings of the 6th International Conference of Computational Models of Argument, COMMA*, 83–94.

Belle, V. 2017. Logic meets probability: Towards explainable AI systems for uncertain worlds. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 5116–5120.

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

Brewka, G.; Eiter, T.; and Truszczynski, M. 2011. Answer set programming at a glance. *Communications of the ACM* 54(12):92–103.

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

Caminada, M., and Dunne, P. 2019. Strong admissibility revisited: Theory and applications. *Argument and Computation* 1–24.

Caminada, M., and Podlaszewski, M. 2012. Grounded semantics as persuasion dialogue. In *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, 478–485.

Cocarascu, O.; Cyras, K.; and Toni, F. 2018. Explanatory predictions with artificial neural networks and argumentation. In *Proceedings of the 2nd Workshop on Explainable Artificial Intelligence*.

Dauphin, J., and Satoh, K. 2019. Explainable ASP. In *International Conference on Principles and Practice of Multi-Agent Systems*, 610–617. Springer.

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.

Dvorák, 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.

Dvořák, W. 2012. *Computational Aspects of Abstract Argumentation*. Ph.D. Dissertation, Technische Universität Wien.

Fan, X., and Toni, F. 2015. On computing explanations in argumentation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, 1496–1502.

Gaggl, S. A.; Linsbichler, T.; Maratea, M.; and Woltran, S. 2020. Design and results of the second international competition on computational models of argumentation. *Artificial Intelligence* 279.

Horridge, M.; Bail, S.; Parsia, B.; and Sattler, U. 2013. Toward cognitive support for OWL justifications. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 53:66–79.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a framework for interactive explanations for recommendations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, 805–815.

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.

Ulbricht, M., and Wallner, J. P. 2021. Strong explanations in abstract argumentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zeng, Z.; Fan, X.; Miao, C.; Leung, C.; Jih, C. J.; and Soon, O. Y. 2018. Context-based and explainable decision making with argumentation. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 1114–1122.