Strong Explanations in Abstract Argumentation (Extended Abstract)

Markus Ulbricht[1] and Johannes P. Wallner[2]

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

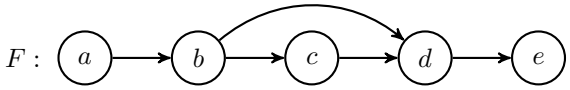
Computational models of argumentation in Artificial Intelligence (AI) [3, 4] provide formal approaches to reason argumentatively, with a wide variety of application avenues, such as legal reasoning, medical sciences, and e-governmental issues [1]. Reasoning in this way is carried out by instantiation of argument structures from a knowledge base [6, 12, 11, 5], 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) [10], is sufficient in order to provide rational accounts on what can be argued for [9]. Known as the area of abstract argumentation, such formalisms provide so-called argumentation semantics [2] 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.

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.

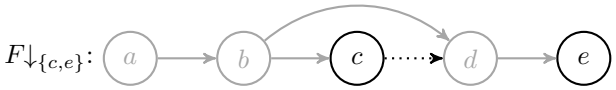
Example 1. Assume it is 2020 and some agents discuss whether or not the next conference should be held virtually. Consider the following arguments which are brought forward during the debate: "The conference should be held virtually in order to avoid a 'super spreader' event" (e); "This is not the same 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 following AF:



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 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 following subframework: 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 [8, 13], initially for strong inconsistency [7], 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 work we study such strong explanations for credulous acceptability under the most common semantics for AFs.

Our work was published in the AAAI'21 proceedings [14]. Main contributions of our work 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.

In this extended abstract we present definitions and properties for strong explanations regarding the mainly the first two items, and refer the reader to the conference paper [14] for more results and details.

1 Background

We recall background on AFs [10] 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 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 [10]: *admissible*, *complete*, *stable*, *preferred*, and *grounded* semantics (abbr. *ad*, *co*, *stb*, *pr*, *gr*).

Definition 2. *Let $F = (A, R)$ be an AF and $E \in cf(F)$.*

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 $co(F)$.*

The notion of a subframework for a given AF F induced by a set $S \subseteq A$ of arguments is defined by $F' = F{\downarrow_S} = (S, R \cap (S \times S))$. 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 $a \in \bigcup \sigma(F)$.

2 Strong Explanations

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.

Definition 3. *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 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 $\mathcal{S}$.

Let us now turn to define our main notion of *strong* σ-explanations. They are inspired by recent related notions [8, 7, 13].

Definition 4. *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. Having established Definition 4 in a formal way, let us now reconsider our motivating example.

Example 2 (Example 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.

Proposition 1. *Let $F = (A, R)$ be an AF, $S \subseteq A$, $X \subseteq A$, and let σ be any semantics $\sigma \in \{ad, co, gr, stb, pr\}$.*

- *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.*

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.

Definition 5. *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 $E \subseteq S \subseteq A$.*

Theorem 1. *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.*

Further, strong explanations form a strictly larger class of explanations. 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). By Example 1 above, also non-admissible sets can be strong explanations, in particular proper subsets of admissible sets, as in the example.

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

Theorem 2. *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.*

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 1. *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.*

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

1. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G.R., Thimm, M., Villata, S.: Towards artificial argumentation. AI Magazine **38**(3), 25–36 (2017)
2. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. The Knowledge Engineering Review **26**, 365–410 (2011)
3. Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.): Handbook of Formal Argumentation. College Publications (2018)
4. Bench-Capon, T.J.M., Dunne, P.E.: Argumentation in artificial intelligence. Artificial Intelligence **171**(10-15), 619–641 (2007)
5. Besnard, P., Hunter, A.: Elements of Argumentation. MIT Press (2008)
6. Bondarenko, A., Dung, P.M., Kowalski, R.A., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. Artificial Intelligence **93**, 63–101 (1997)
7. Brewka, G., Thimm, M., Ulbricht, M.: Strong inconsistency. Artificial Intelligence **267**, 78–117 (2019)
8. Brewka, G., Ulbricht, M.: Strong explanations for nonmonotonic reasoning. In: Description Logic, Theory Combination, and All That, pp. 135–146. Springer (2019)
9. Caminada, M.: Rationality postulates: Applying argumentation theory for nonmonotonic reasoning. In: Baroni, P., Gabbay, D., Giacomin, M., van der Torre, L. (eds.) Handbook of Formal Argumentation, chap. 15. College Publications (2018)
10. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995)
11. García, A.J., Simari, G.R.: Defeasible logic programming: An argumentative approach. Theory and Practice of Logic Programming **4**(1-2), 95–138 (2004)
12. Modgil, S., Prakken, H.: A general account of argumentation with preferences. Artificial Intelligence **195**, 361–397 (2013)
13. Saribatur, Z.G., Wallner, J.P., Woltran, S.: Explaining non-acceptability in abstract argumentation. In: Proc. ECAI. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 881–888 (2020)
14. Ulbricht, M., Wallner, J.P.: Strong explanations in abstract argumentation. In: Proc. AAAI. pp. 6496–6504. AAAI Press (2021)