HAL open science

Explaining Semantics and Extension Membership in Abstract Argumentation

Philippe Besnard, Sylvie Doutre, Théo Duchatelle, Marie-Christine Lagasquie-Schiex

▶ **To cite this version:**

Philippe Besnard, Sylvie Doutre, Théo Duchatelle, Marie-Christine Lagasquie-Schiex. Explaining Semantics and Extension Membership in Abstract Argumentation. Intelligent Systems with Applications, 2022, 16, pp.200118. 10.1016/j.iswa.2022.200118 . hal-03771080

HAL Id: hal-03771080

https://ut3-toulouseinp.hal.science/hal-03771080v1

Submitted on 20 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Intelligent Systems with Applications 16 (2022) 200118

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Explaining Semantics and Extension Membership in Abstract Argumentation

Philippe Besnard [a], Sylvie Doutre [b], Théo Duchatelle [c,*], Marie-Christine Lagasquie-Schiex [c]

[a] IRIT, CNRS, Toulouse, France
[b] IRIT, Université Toulouse 1, France
[c] IRIT, Université Toulouse 3, France

ARTICLE INFO

Keywords:
Abstract argumentation
Semantics
Explanabilit

ABSTRACT

This paper explores the computation of explanations in the specific context of abstract argumentation. The explanations that we define are designed to be visual, in the sense that they take the form of subgraphs of the argumentation graph. Moreover, these explanations rely on the modular aspects of abstract argumentation semantics and can consequently be either aggregated or decomposed. We investigate graph properties of these explanations, and their adequacy to desirable explanatory criteria.

1. Introduction

When it comes to explanations of decisions made using an Artificial Intelligence system, Abstract Argumentation, introduced in Dung (1995), is increasingly studied as a formal tool to provide them. In Abstract Argumentation, the main object of study is an argumentation framework, which is a directed graph whose nodes are abstract arguments (abstract in the sense that their internal structure is left unspecified) and whose binary relation is a conflict relation. Given an argumentation framework, the objective is to find the arguments that can collectively be deemed receivable according to some criteria. Such sets of arguments are called *extensions* and the criteria that are considered desirable give rise to *semantics* for selecting extensions. The recent survey by Cyras et al. (2021) indicates that Argumentation can be used to generate explanations in various domains (machine learning notably) and that explanations for the argumentative process itself are also necessary.

In this respect, the main questions which have been addressed so far concern the global acceptability status (credulous or skeptical) of an argument or of a set of arguments; for instance one can find the question: ”Why does a given argument belong to each extension of a given semantics?” (skeptical acceptability). In addition, the approach that is most often used consists in identifying sets of arguments which act as explanations (Baumann and Ulbricht (2021); Borg and Bex (2020b, 2021c); Fan and Toni (2015a); Liao and van der Torre (2020); Ulbricht and Wallner (2021)). One may however argue that, since the argumentative process of Abstract Argumentation already provides ways for selecting arguments, explaining this process by more selection of arguments (although different ones) may not be of much help. Furthermore, beyond the question of the global acceptability of an argument or a set of arguments, many other questions on the outcomes of argumentation or on the process of argumentation itself can be asked, for instance: ”Why is a given set of arguments an extension under a given semantics?”.

Our aim is to take into account several types of questions in the context of Abstract Argumentation: those related to the semantics extensions, and those related to the acceptance of arguments, acceptance in terms of membership in a given extension. Contrastive and non-contrastive questions will be addressed in this later case. Moreover, we will propose answers which will take the form of relevant subgraphs, as in Niskanen and Järvisalo (2020); Racharak and Tojo (2021); Saribatur et al. (2020): our approach is a visual one, which has been shown to be helpful for humans to comply with reasoning principles (Vesic et al. (2022)), and which not only highlights arguments, but also subsets of attacks.

The paper is organised as follows: Section 2 recalls background notions relative to abstract argumentation and graph theory. Section 3 recalls some existing works related to our topic (explanations in abstract argumentation for abstract argumentation). Then Section 4 presents our motivations giving the context of our work and our first assumptions. In Section 5, the definition of our explanations for semantics in Abstract Argumentation is given, and their properties are investigated. Then, for the case of explanations about acceptance, some additional motivations and hypotheses are given in Section 6 and the formal definitions of these explanations are presented in Section 7. Section 8 summarises the results

* Corresponding author.

https://doi.org/10.1016/j.iswa.2022.200118
Received 20 April 2022; Received in revised form 21 July 2022; Accepted 25 August 2022
Available online 31 August 2022
2667-3053/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

of all the previous sections. We compare our approach to the existing ones in Section 9. In Section 10, we provide a discussion on the quality of our explanations. Section 11 concludes and presents future works.

2. Preliminary Notions

In this section, we give the background notions that will be of use in this paper. They include some definitions about argumentation frameworks and the description of some important operations over graphs.

2.1. Abstract Argumentation

We begin by recalling basic notions on Abstract Argumentation. The object handled in this formalism is called an argumentation framework.

Definition 1. (*Argumentation framework (Dung, 1995)*) A *Dung's argumentation framework* is an ordered pair (A, R) such that $R \subseteq A \times A$.

Each element $a \in A$ is called an *argument* and aRb means that a *attacks* b. For $S \subseteq A$, we say that S attacks $a \in A$ iff bRa for some $b \in S$. Any argumentation framework can be represented as a directed graph.

The main asset of Dung's approach is the definition of semantics using some basic properties in order to define sets of acceptable arguments, as follows.

Definition 2. Let $\mathscr{AF} = (A, R)$. An argument $a \in A$ is *acceptable* wrt $S \subseteq A$ iff for all $b \in A$, if bRa then cRb for some $c \in S$ (a is defended by S).

The *characteristic function* of $\mathscr{AF}$ is $F_{\mathscr{AF}} : 2^A \rightarrow 2^A$ such that $F_{\mathscr{AF}}(S) = \{a \in A | a \text{ is acceptable wrt } S\}$ for any $S \subseteq A$.

The semantics originally defined in Dung (1995) are as follows.

Definition 3. Given $\mathscr{AF} = (A, R)$, a subset S of A is said to be:

- a *conflict-free set* iff there are no a and b in S such that a attacks b,
- an *admissible set* iff S is conflict-free and for any $a \in S$, a is acceptable wrt S,
- a *complete extension* iff S is admissible and for any $a \in A$, if a is acceptable wrt S then $a \in S$,
- a *preferred extension* iff S is maximal (in the sense of set-inclusion)[1] admissible set,
- a *grounded extention* iff S is the least fixpoint for $F_{\mathscr{AF}}$,
- a *stable extension* iff S is conflict-free and S attacks any $a \in A \backslash S$.

Some properties have been proven in Dung (1995) establishing a link between the different semantics. For instance:

Proposition 1. *Given $\mathscr{AF} = (A, R)$:*

- *There exists at least one preferred extension.*
- *Every preferred extension is complete, but not vice-versa.*
- *Every stable extension is preferred, but not vice-versa.*
- *The grounded extension is the least (with respect to set-inclusion) complete extension.*

Table 1 illustrates these semantics for the AF given in Figure 1.

2.2. Graph Theory

This section recalls some graph-theoretic notions[2] concerning particular subgraphs and nodes, as well as the successor and predecessor functions.

Definition 4. Let $G = (V, E)$ and $G' = (V', E')$ be two graphs.

- G' is a *subgraph* of G iff $V' \subseteq V$ and $E' \subseteq E$.

[1] We write $\subseteq$-*maximal*.

[2] See Bondy and Murty (2008) for more details about these notions.

Table 1
Acceptable sets of the AF of Fig. 1 under the different semantics.

	Admissible	Complete	Preferred	Grounded	Stable
$\varnothing$	✓	✓		✓	
$\{h\}$	✓	✓			
$\{i\}$	✓	✓			
$\{a, d\}$	✓	✓			
$\{b, c\}$	✓				
$\{a, d, h\}$	✓				
$\{a, d, i\}$	✓				
$\{b, c, h\}$	✓				
$\{b, c, i\}$	✓				
$\{b, c, e\}$	✓	✓			
$\{a, d, h, f\}$	✓	✓	✓		✓
$\{a, d, i, f\}$	✓	✓	✓		✓
$\{b, c, e, h\}$	✓	✓	✓		✓
$\{b, c, e, i\}$	✓	✓	✓		✓

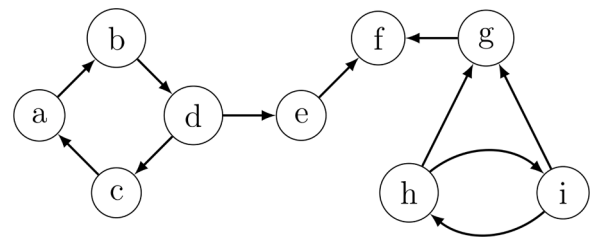


Fig. 1. Example of an argumentation framework (AF) from Borg and Bex (2021a).

- G' is an *induced subgraph of G by V'* if G' is a subgraph of G and for all $a, b \in V', (a, b) \in E'$ iff $(a, b) \in E$. G' is denoted as $G[V']_V$.
- G' is a *spanning subgraph of G by E'* if G' is a subgraph of G and $V' = V$. G' is denoted as $G[E']_E$.

A subgraph G' of G is included in G. In an induced subgraph G' of G by a set of *vertices* S, some vertices of G can be missing but all the edges concerning the kept vertices are present. In a spanning subgraph G' of G by a set of *edges* S, all the vertices of G are present but some edges of G can be missing.

Induced and spanning subgraphs are examples of ways to compute a graph from another single graph. Another interesting operation producing a new graph from other ones is the union that represents the aggregation of the information contained in the two graphs:

Definition 5. (*Graph union*) Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs. We define the *union* of G_1 and G_2 by $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$.

Let us consider also a particular kind of graphs, bipartite graphs, whose set of vertices can be split in two disjoint sets and in which every arc connects a vertex of one part to a vertex of the other part:

Definition 6. (*Bipartite Graph*) Let $G = (V, E)$ be a graph. G is *bipartite* (with *parts* T and U) iff there exists $T, U \subseteq V$ such that $T \cup U = V$ and $T \cap U = \varnothing$ (T and U are a partition of V) and for every $(a, b) \in E$, either $a \in T$ and $b \in U$, or $a \in U$ and $b \in T$. G will be denoted with (T, U, E) and U is the *complement part of* T (and vice-versa).

The next notions are about the *successor* and the *predecessor* functions.

Definition 7. (*Successor and Predecessor functions*) Let $G = (V, E)$ be a graph. The *successor* function of G is the function $E^+ : V \mapsto 2^V$ such that $E^+(v) = \{u \mid (v, u) \in E\}$ and the *predecessor* function of G is the function $E^- : V \mapsto 2^V$ such that $E^-(v) = \{u \mid (u, v) \in E\}$. Let S be a set of vertices, $E^+(S) = \bigcup_{v \in S} E^+(v)$ and $E^-(S) = \bigcup_{v \in S} E^-(v)$.

Let $n \geq 0$. The *n-step successor* (resp. *predecessor*) *function* of G is

$E^{+n}(v) = \overbrace{E^{+} \circ \cdots \circ E^{+}}^{\text{n times}}(v)$ (resp. $E^{-n}(v) = \overbrace{E^{-} \circ \cdots \circ E^{-}}^{\text{n times}}(v)$). By convention, we have $E^{+0}(v) = E^{-0}(v) = v$.[3]

Considering an AF, the successor (resp. predecessor) function represents the arguments that are attacked by (resp. the attackers of) some argument(s). An AF being usually denoted by (A, R), the successor and predecessor functions are thus denoted R^{+} and R^{-} in this context.

Finally, let us consider some vertices having a particular status in a graph.

Definition 8. (*Source, Sink, Isolated vertex*) Let $G = (V, E)$ be a graph and v be a vertex of G. v is said to be a *source* iff $E^{-}(v) = \varnothing$ and it is said to be a *sink* iff $E^{+}(v) = \varnothing$. v is said to be *isolated* iff it is both a source and a sink.

Example 1. Fig. 2 represents the induced subgraph of Fig. 1 by $\{a, b, d, e\}$. Fig. 3 represents the spanning subgraph of the same graph by $\{(a, b), (d, c), (d, e), (h, g), (h, i)\}$. Fig. 2 is a bipartite graph with parts $\{a, d\}$ and $\{b, e\}$. In Fig. 3, f is an isolated vertex, a, d and h are sources and b, c, e, g and i are sinks. Let (A, R) denote the graph of Fig. 2. We have $R^{+}(b) = \{d\}$ and so $R^{+2}(b) = \{e\}$. Similarly, $R^{-}(d) = \{b\}$ and so $R^{-2}(d) = \{a\}$.

The elements (vertices and edges) belonging to the subgraphs are in black. Those in light gray are in the original graph but not in the subgraphs

3. Related Works

Before giving our approach, we present several existing works related to the computation of explanations for Abstract Argumentation. Our presentation will follow the "taxonomy" of the types of explanation proposed in the survey Cyras et al. (2021): explanations can be defined as either subgraphs, or changes, or extensions (sets of arguments).[4]

Let us consider first the category of *subgraphs*. A first example of work defining explanations as *subgraphs* is Saribatur et al. (2020). It was categorised in the second category (change) in Cyras et al. (2021), a choice which can be discussed since Saribatur et al. (2020) seek to explain the credulous non acceptance of some argument, not by changing its status, but by finding a strongly rejecting subframework. A strongly rejecting subframework is an induced subgraph of an argumentation framework that does not credulously accept an argument, and nor do its supergraphs (that are still induced subgraphs of the original AF). As such, strongly rejecting subframeworks capture the core argumentative reasons for why an argument is not credulously accepted under a certain semantics.

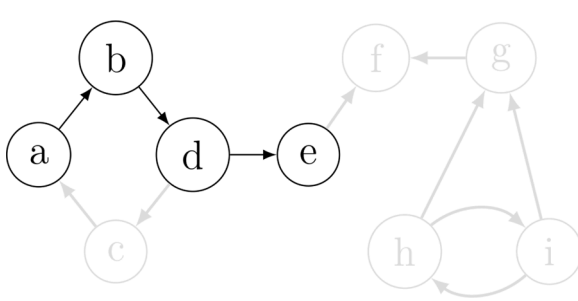


Fig. 2. Induced subgraph of Fig. 1 by $\{a, b, d, e\}$.

[3] Note that $E^{+1}(v) = E^{+}(v)$ and $E^{-1}(v) = E^{-}(v)$

[4] Note that another category is also evoked in Cyras et al. (2021): the dialogue-games. This type of explanation could be seen as close to defense-trees (a sub-category of subgraphs) but with several specific features (game protocols and winning conditions). So we do not present these works here.

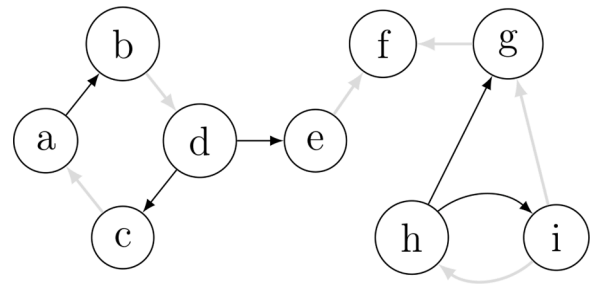


Fig. 3. Spanning subgraph of Figure 1 by $\{(a, b), (d, c), (d, e), (h, g), (h, i)\}$.

Niskanen and Järvisalo (2020) also study subgraphs to obtain explanations for the credulous non acceptance of some argument for a given semantics (except the grounded semantics). The differences here are that the authors consider both induced and spanning subgraphs for their explanations, and the subgraphs are not the explanations themselves, but rather used to characterize explanations. More precisely, they call a set of arguments (resp. of attacks) an explanation if the induced subgraph (resp. spanning subgraph) computed using this set does not credulously accept the queried argument, and nor does any of its supergraph.

Ulbricht and Wallner (2021) propose strong explanations for credulous acceptance of a set of arguments under a given semantics. A strong explanation is a set of arguments such that for every subgraph induced by a superset of the explanation, there exists an extension of the considered semantics that includes the set to explain. Thus, strong explanations for credulous acceptance can be seen as a core set of arguments needed for an argument to be part of at least one extension under the desired semantics.

A specific kind of graph that is also used in explaining argumentative results is *defence trees*. Defence trees are trees where nodes are arguments and each successor of a node is an attacker of that node. As such, they can be used to prove whether an argument is defended or not. Some works, like Racharak and Tojo (2021), use defence trees as explanations for argumentative results. Racharak and Tojo (2021) argue that a defence tree is a dialogical explanation for an argument since it can be used to show that it is defended. Other works, like (Fan and Toni, 2015a), use them to compute their notion of explanations. In these works, defence trees are used to explain the credulous acceptance of some argument under admissibility.

We now turn to the second category, which concerns *changes*. It consists in identifying what elements to remove from the AF in order to modify a given result. This is the method used in Fan and Toni (2015b), in which the authors explain why an argument is not credulously accepted under admissibility. Their explanations consist of sets of arguments or attacks to remove from the AF in order to make the considered argument credulously accepted under admissibility in the resulting subgraph. Such sets were also studied in Ulbricht and Baumann (2019) (in which they were called "*diagnosis*") although the authors restrained themselves to the case where a given semantics does not yield any extension. Diagnoses are also parts of the study of Niskanen and Järvisalo (2020), which provides a way of computing them (as well as explanations as subgraphs) using logical formulas and providing complexity results.

The third category of approach consists of taking *sets of arguments* as explanations. This is probably the most widely used approach to this problem. In most of the works using this method, the point of view is to consider that explanation equates to justification. Hence the restriction to sets of arguments as explanations, since given a set of arguments, the original AF can be used to justify it by the mean of the attack relation.

In Fan and Toni (2015a), the authors define an explanation semantics, called related admissibility, which provides all the reasons why an argument belongs to an admissible set. The idea is to get rid of all the arguments that are not relevant for the acceptance of the considered

argument, that is, those that are not connected to it via the attack relation.

In Borg and Bex (2020b, 2021c), the authors propose a basic framework to compute explanations as sets of arguments for the credulous/skeptical acceptance or non-acceptance of an argument. It is a framework for explanations since it can be parameterised in order to modify the way explanations are computed. In their work, the authors focus on some human biases used to select explanations such as simplicity (taken as minimality), sufficiency and necessity. In subsequent works (Borg and Bex (2021a,b)) the authors extend their framework to adapt it to Structured Argumentation (adding another parameter to control the form of the explanation) and to compute contrastive explanations (the intersection of why an argument (the fact) is accepted and why a set of arguments (the foils) are not).

Some other works define their explanations from the observation that in the computation of an extension, some parts are non-deterministic choices, while others, deterministic, result from the first ones. For instance, in Liao and van der Torre (2020), the authors base their approach on the observation that each Strongly Connected Component (SCC) of an argumentation framework can be seen as making a choice for accepting conflict-free sets of arguments. From these choices results the rest of the accepted arguments. Thus, in a set of arguments, each argument can be explained by the set of arguments that were chosen in a given SCC.

Similarly, in Baumann and Ulbricht (2021), the authors observe that complete and admissible semantics are computed firstly by the computation of the grounded (resp. strongly admissible) extension, then making choices in even cycles, and finally computing the grounded (resp. strongly admissible) extension again. As such, they define the arguments chosen in the even cycles as the explanations for some complete or admissible extension.

4. Motivation and Hypotheses

4.1. Motivation

In order to motivate the explanation problems that we consider and the answers that we propose, let us consider a real-case example from the Dutch National Police from Borg and Bex (2020a).[5]

Example 2. A citizen has ordered a product through an online shop, paid for it, and received a package. However, it is the wrong product, it seems suspicious as if it might be a replica, rather than a real product. Still, the citizen wants to file a complaint of internet trade fraud. While the citizen provides the information from the described scenario, the system constructs further arguments from this, based on the Dutch law. Arguments are obtained (their conclusions are emphasized):

A_1 It is not because the wrong product was received, that it is a case of fraud; then we may consider that it is *not a case of fraud*.

A_2 It is not because the wrong product was received, that the counterparty has not delivered; the *counterparty has delivered*.

A_3 A suspicious product is usually fake, which supports the fact that the *product is fake*.

A_4 The reasons which lead to the conclusion that the product is fake, and the fact that when a product is fake, then usually the counterparty did not deliver, lead to the conclusion that the *counterparty did not deliver*.

A_5 An investigation shows that there is no problem with the product: the *product is not fake*.

A_6 The fact that the complainant paid and was delivered, combined to the assumption that the product is fake and to the other reasons which lead to the conclusion that the counterparty did not deliver, shows that it is likely to be a *case of fraud*.

This scenario, the arguments and their attack relationships, can be represented by the argumentation framework depicted on Fig. 4.

Which conclusion can be drawn from this situation? A_1 concludes that it is not a case of fraud, whereas A_6 concludes that it is a case of fraud. Both arguments are contained in coherent acceptable sets under the complete semantics: $\{A_1\}$, $\{A_1, A_3, A_4\}$ and $\{A_1, A_2, A_5\}$ are complete extensions which contain A_1; $\{A_6, A_3, A_4\}$ is a complete extension which contains A_6. In order to better understand the situation, explanations can be sought regarding the acceptability of these arguments and of the sets which contain them, based on their interactions.

One may wonder why both A_1 and A_6 cannot be accepted together in a complete extension. An explanation should show that the set $\{A_1, A_6\}$ is not a complete extension because it is not conflict-free. This is what our approach will do, by presenting a subgraph which shows both arguments and their attacks. Fig. 5 shows this resulting explanation subgraph.

One may wonder why $\{A_6, A_3, A_4\}$ is a complete extension, in order to understand how this set which contains A_6 is collectively acceptable. An explanation should show three elements: that the set is conflict-free, that it defends all its elements, and that it satisfies the reinstatement principle (any argument defended by the set must belong to the set). Conflict-freeness can be shown by a subgraph which highlights that there are no attacks between the arguments of the set (see Fig. 6); defence can be shown by a subgraph which highlights the attacks which target arguments of the set, and how the set attacks them back; reinstatement can be shown by a subgraph which highlights all the arguments defended by the set, and by showing that they all belong to the set.

These two questions ("Why a set of arguments is (resp. is not) acceptable under a given semantics?") have not really been addressed in the related works.

A reason for this lack of interest may be that, definitions of acceptability semantics being formally given, providing explanations for why a set is acceptable under a given semantics is not necessary; the definition itself can be considered as an explanation.

However, when trying, in a pedagogical perspective, to make an elementary user understand the acceptability semantics, or, whatever be the level of the user, when the considered semantics is composed of several principles, or when facing a large argumentation framework, or when testing, in its development phase, an argumentation solver, explaining how a set is acceptable or not under a given semantics is of interest and of importance. In each of these cases, offering subgraphs which focus on the explanation of the component principles may help understanding how acceptability is constructed.

This choice of this form of explanation, subgraph-based, can be further motivated. We consider that sets of arguments or of attacks are, in some sense, not enough to explain the argumentative process. Indeed, a set of arguments does not give any indication as to how these arguments are related. In order to know how the arguments are connected with each other, one needs to use the graph they are extracted from. Without having this graph at disposal, n arguments may interact with

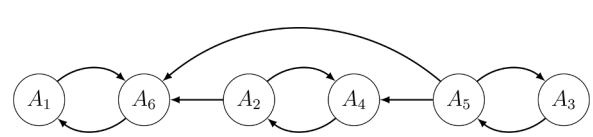


Fig. 4. Delivery Example from Borg and Bex (2020a).

[5] Example which is a slightly adapted and reduced version of the original one of Borg and Bex (2020a).

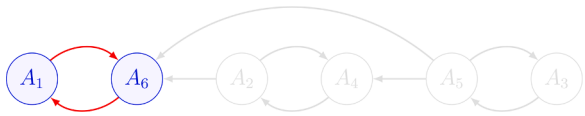


Fig. 5. A subgraph explaining why $\{A_1, A_6\}$ is not conflict-free in Example 2.

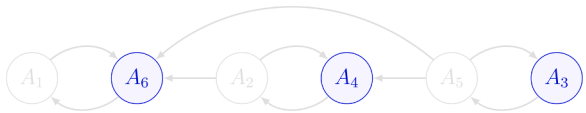


Fig. 6. A subgraph explaining why $\{A_3, A_4, A_6\}$ is conflict-free in Example 2.

each other in 2^{n^2} possible situations. Thus, we may not determine easily which one of these configurations is the correct one. On the contrary, sets of attacks give an indication as to how the arguments are related since the set contains exactly a part of the relation on the arguments. However, when using sets of arcs, one might miss some arguments that were important in the decision without being connected to the arguments implied by the arcs of the set.

For this, we consider subgraphs (i.e. both a set of arguments and a set of attacks) to constitute better explanations. They allow for a potentially full introspection of the original model, but they can also be used to highlight specifically targeted areas of the AF that were relevant in the decision making process.

There is also another reason to use subgraphs: when the need for explanation is in reaction to a decision, one might argue that explaining the decision with a similar result, that is to say selecting arguments (even if in a different way), may not be the best approach in this case. The use of subgraphs here allows to rely on the visual modality of the user, which thus contrasts with what was required to understand the initial decision.

Note that, on the above Example 2, some other questions naturally arise: "Why is A_6 (or A_1) acceptable?" in the sense of credulous acceptability[6] notably, or "Why is A_6 accepted in $\{A_6, A_3, A_4\}$?", "Why is A_1 not accepted in $\{A_6, A_3, A_4\}$?", or even "Why is A_6 accepted in $\{A_6, A_3, A_4\}$ and not A_1?" (contrastive question). Our motivations concerning these questions will be addressed in Section 6.

This motivating example leads to further hypotheses on the context in which the questions are asked. The next section presents them.

4.2. Hypotheses

Regarding the context of this work, a first assumption:

A user asks for an explanation after they have been presented the result of a Formal Argumentation process (typically the selection of arguments via a semantics) by a program that we will refer to as the system. (H1)

If we consider Example 2, it is when presented with the argumentation framework of Fig. 4 and with a complete extension which contains for instance A_1, that a user may ask for an explanation.

This behaviour is also illustrated on Fig. 7 with a different example: we consider a user interested in a specific stable extension presented by the system (*the context of the question* being the result that is presented to the user, the semantics under which it was computed and the argumentation framework that was used to compute it).[7]

In addition to the first assumption:

The user is able to understand argumentation frameworks. (H2).

Indeed, the purpose of explanations would then be to select the relevant parts of this context in order to facilitate the user's inspection process. The hypothesis that we make is that, when presented with a graph interpreted as an argumentation framework, the user is able to understand that the nodes represent arguments and that the arcs represent conflicts between the arguments. The main reason we feel confident about this assumption is that if the user is not able to understand argumentation frameworks, it should not be complicated to describe how to "read" one.

The next hypothesis we make is directly related to the previous one:

The user knows Abstract Argumentation semantics. (H3).

Thus, we assume that the user, even if they are not experts in this domain, is already versed in Abstract Argumentation. Indeed, we assume that the user knows that argumentation semantics are based on several basic principles: conflict-freeness, defence, reinstatement, and complement attack.[8] As such, the explanations we define are not yet for ordinary users.[9]

Finally, as noted in Miller (2019), explanation is a social process, one that does not stop at the selection of an explanation. People receiving explanations expect them to obey to a certain number of rules, and evaluate their quality based on their adequacy with these rules. An example of such rules are Grice's maxims of conversation (Grice, 1975). Grice gave a set of simple rules that people tend to follow when engaging in a cooperative conversation. A cooperative conversation is a discussion that happens between two or more agents that all make efforts in contributing to reaching a common goal which may be for instance exchanging information or achieving social bonding. Grice firstly gives one general principle to follow when engaging in a cooperative conversation: *"Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged"*. Grice calls this the *Cooperative Principle*. Grice then gives four categories of maxims to follow in order to adhere to the Cooperative Principle that he calls *Quantity, Quality, Relation* and *Manner*. He gives the following maxims in these categories (directly cited from Grice (1975)) :

1. *Quantity*: (a) Make your contribution as informative as required (for the current purpose of the exchange); (b) Do not make your contribution more informative than is required
2. *Quality*: (a) Do not say what you believe to be false; (b) Do not say that for which you lack adequate evidence
3. *Relation*: (a) Be relevant
4. *Manner*: (a) Avoid obscurity of expression; (b) Avoid ambiguity; (c) Be brief (avoid unnecessary prolixity); (d) Be orderly

Grice also gives what he calls *supermaxims*, namely "Try to make your contribution one that is true" (*Quality* category) and "Be perspicuous" (*Manner* category).

We argue that seeking and providing explanation in a question-answer setting (such as the one we place ourselves in) certainly falls into the category of cooperative conversations. As such, we will make efforts to define explanations that adhere to these maxims, and use them as a way to evaluate our explanations. And so the last assumption we make is:

[6] An argument is credulously accepted under a given semantics if it belongs to at least one extension under this semantics.

[7] Of course the user is free to ask about a set and/or a semantics that could be different from those given in the context of the question.

[8] Conflict-freeness: an attacked argument and its attacker cannot belong to the same resulting extension. Defence: an attacked argument must be defended against its attacker by a counter-argument in the same resulting extension. Reinstatement: any defended argument must belong to the resulting extension. Complement attack: any argument that is not in the resulting extension must be attacked by an element of this extension.

[9] This is however the objective we want to reach. Hence, we intend to drop this hypothesis in future work, so that we are able to generate explanations for any user and not only those that already know Abstract Argumentation.

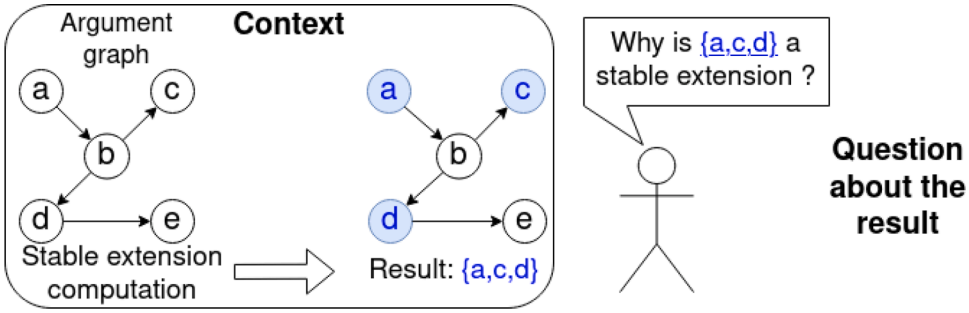


Fig. 7. The first assumption: a user reacts to a given result produced by the system.

Grice's maxims are correct and should thus be followed when engaging on the explanation of the result of the system. (H4).

5. Providing Explanations for Semantics Extensions

In this section, we focus on how to provide answers to a certain class of questions. These questions, introduced in Section 4.1 as "Why a set of arguments is (resp. is not) acceptable under a given semantics?", are reformulated here as: "Why is S [not] an X extension ?" with X ranging over conflict-free, admissible, complete and stable, and S being any subset of arguments in the argumentation framework of the question's context.

Example 3. Consider the argumentation framework depicted on Fig. 1. Imagine that a user asks the question "Why is { a,b,c } a complete extension ?". In this case, $\{a, b, c\}$ is not a complete extension, thus we would need to provide the elements that show what makes it impossible for $\{a, b, c\}$ to be a complete extension. Alternatively, imagine that a user asks the question "Why is { a,d } not a complete extension ?". In this case, $\{a, d\}$ is a complete extension, hence we should provide the elements that show the reasons for this set to be such an extension.

The next step is to discuss what these elements are. To define them, we adopt the modular view saying that the properties we consider (being an extension of a semantics) are a conjunction of different conditions. In other terms, to verify that a set of arguments is an extension of a given semantics, we must verify that this set respects a certain number of conditions. Alternatively, to verify that a set of arguments is not an extension of a given semantics, we must verify that this set does not respect one condition out of a certain number of them. This is undoubtedly a very basic viewpoint, but it is supported by the fact that semantics are precisely defined this way. Hence, to explain why a set of arguments is an extension of a certain semantics, we provide the relevant parts of the argumentation framework that allow to check that all the conditions corresponding to this semantics are satisfied by this set. Consequently, an explanation here is made of two components:

- A part (subgraph) of the argumentation framework denoted by G_X (X being an expression denoting the condition or the semantics we talk about)
- A checking procedure (that answers YES or NO)

When providing explanations, we will focus on having the checking procedures as simple and intuitive as possible. Thus, we will describe them informally in each case. In addition, if we wish to show an explanation for several conditions at once, we may show the aggregation of the reasons for every condition to hold (so the union of the corresponding subgraphs).

In each following subsection, the same methodology will be used:

- Discuss what the explanation is about and identify what is required through an introductory example always based on the running example described in Fig. 1;
- Formally define the explanation (eventually in several steps).

Note that, on the figures representing our explanations, we will use the following convention:

Legend of the colors used in the explanation subgraphs:

- The arguments of the set which is given as input in the question will be in blue.
- Arguments or attacks that may cause a checking procedure to fail will appear in **thick red**.
- Other arguments and attacks of the explanation subgraph will appear in **black**.
- The rest of the original graph on which the subgraph is built will be in light gray.

Some interesting properties will be presented either in each subsection (if they are specific to a given semantics or condition), or regrouped in Section 5.5. Note that the proofs of these results can be found in Doutre et al. (2022).

Note also the notations we use in this section are the following ones:

- $\mathscr{AF} = (A, R)$ denotes the argumentation framework of the context
- σ denotes the semantics used in the context
- S denotes the result given in the context and corresponds to an extension of $\mathscr{AF}$ under σ
- S' denotes the set the user asks about (S' can be equal to S or not)

5.1. Explanation for Conflict-freeness

Recall that a set of arguments is conflict-free if and only if there are no arcs between its members. Hence, if we are to show why a set of arguments is conflict-free, we must show a part of the graph that highlights the absence of arcs within this set. We begin with an example in order to provide an intuition of how to define the explanation.

Example 4. Consider the argumentation framework of Fig. 1 and the questions "Why is {a,d,h,f} a conflict-free extension ?" and "Why is {a,d, e} a conflict-free extension ?". Figs. 8 and 9 show the answers for the first and second question respectively (remember the legend of the colors presented earlier in the section). The idea is to make sure that all the arcs that are present within the given set are shown. Hence, if there is at least one, we can conclude that the set is not conflict-free and if there is none, we can conclude it is.

Following the examples, we realise that we must compute a subgraph of the argumentation framework that shows any arc between the arguments of the set the question is about, if there are some. Preferably, this subgraph should be as small as possible to get rid of any irrelevant information.

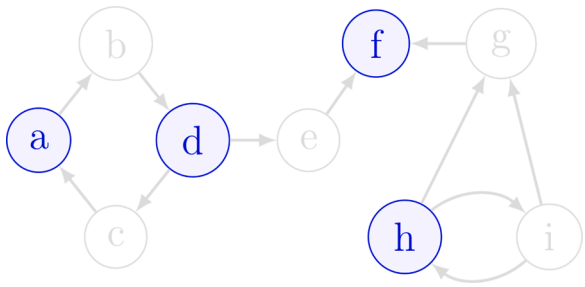


Fig. 8. Explanation on why $\{a, d, h, f\}$ is conflict-free in Fig. 1. All arcs between a, d, h and f are represented and there is none.

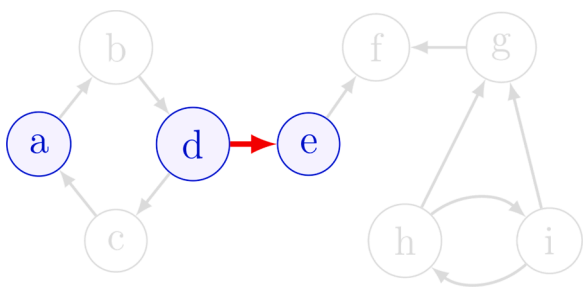


Fig. 9. Explanation on why $\{a, d, e\}$ is not conflict-free in Figure 1. All arcs between a, d and e are represented, and there is one between d and e.

Definition 9. (*Explanation for conflict-freeness*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant subgraph to answer the question "Why is S' a conflict-free extension ?" is[10]

$$G_{CF}(S') = \mathscr{AF}[S']_V$$

The checking procedure for G_{CF} is to verify that the set of edges in the resulting subgraph is empty. Fig. 10 illustrates the built answer giving the subgraph and the checking procedure.

By defining the explanation as an induced subgraph, we make sure to have a reduced size as well as showing all the arcs that are concerned.

Some properties of G_{CF} are given in Section 5.5.

5.2. Explanation for Admissibility

Recall that a set of arguments is admissible if and only if it is conflict-free and all its arguments are acceptable with respect to it. Since to be conflict-free is a condition to be admissible, part of the explanation for admissibility is the explanation for conflict-freeness. The other part of the explanation involves the acceptability of all members of the set with respect to the set itself, which can be understood as defending the point of view represented by the set of arguments against its attackers. Therefore, if we are to show why a set of arguments only contains arguments that are acceptable with respect to it, we must exhibit a part of the graph highlighting that all the attackers of this set are attacked in return. We begin by illustrating on some examples to give the intuition, and then formally define these explanations.

Example 5. Consider the argumentation framework of Fig. 1. Figs. 11 and 12 show why $\{a, d, h, f\}$ defends all its arguments and why $\{b, e\}$ does not. In Fig. 11, we can see that all the attackers of a, d, h and f are the endpoint of an arc whose origin is either a, d, h or f. In Fig. 12, we see that a attacks b and neither b nor e defends b against this attack.

Following the examples, we realise that we must compute a subgraph of the argumentation framework that includes both the set the question is about and its attackers. The only arcs that are relevant are those from the attackers to the set (to show that they are indeed attackers) and from the set to attackers (to show whether the set defends itself or not). So we must first compute an induced subgraph and then a specific spanning subgraph of this induced subgraph.

Definition 10. (*Explanation for defence*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant subgraph to explain whether or not S' contains only arguments that are acceptable w.r.t. S' is

$$G_{Def}(S') = \left| \begin{array}{ll} \mathscr{A}F\left[S' \cup R^{-1}(S')\right]_V)[\{(a,b) & \in R | a \in R^{-1}(S') \text{ and } b \in S', \\ & \text{or } a \in S' \text{ and } b \in R^{-1}(S')\}]_E \end{array} \right.$$

The checking procedure for G_{Def} is to verify that every argument that does not belong to S' in the resulting subgraph is the endpoint of an edge whose origin is in S'. We can now define the explanation for admissibility. As we previously said, this explanation is in two parts: one for conflict-freeness and one for defence. Note that these two subgraphs may be either aggregated into a single one, using the union operator of Definition 5, or separated in the definition of an explanation for admissibility.[11]

Definition 11. (*Explanation for admissibility*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant set of subgraphs (with their checking procedures) to answer the question "Why is S' an admissible extension ?" is

$$G_{Adm}(S') = \{G_{CF}(S'), G_{Def}(S')\}$$

Fig. 13 illustrates on a very simple example the built answer giving the subgraphs for G_{CF} and G_{Def} and the corresponding checking procedures.

To conclude on the explanation for admissibility, we give the following proposition. It formalises a visual behavior of the explanation for defence depending on conflict-freeness.

Proposition 2. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. If S' is conflict-free, $G_{Def}(S')$ is a bipartite graph such that there exists a partition of its vertices with S' as one of its parts.*

As such, if the explanation for defence is computed on a set that is conflict-free, it is possible to separate its vertices in two groups (S' and its attackers) with the arcs always going from one group to the other. Alternatively, if one computes the explanation for defence on a set and if it does not result in a bipartite graph with the set as one of its parts, we may conclude that the set is not conflict-free.

Some other properties of G_{Def} and G_{Adm} are given in Section 5.5.

5.3. Explanation for Completeness

Recall that a set of arguments is complete if and only if it is admissible and all the arguments that are acceptable with respect to it are members of this set. Hence, part of the explanation for completeness is the explanation for admissibility. The other part of the explanation is about the membership of all acceptable arguments. This can be understood as adopting a point of view in which we take all the arguments that we know we can defend (in a sense, we take as much as we are "forced" to). Thus, if we want to show why a set of arguments contains all the arguments that it can defend, we must show a part of the graph highlighting that the arguments this set could defend are either defended

[10] This is the induced graph by S'. Note that S' is not necessarily S, the result presented by the system. The user can ask about any set S' built from $\mathscr{AF}$.

[11] Each option has its advantage and disadvantage: the aggregation is more concise, but less readable than the separation that allows the user to better identify the individual conditions composing the explanation.

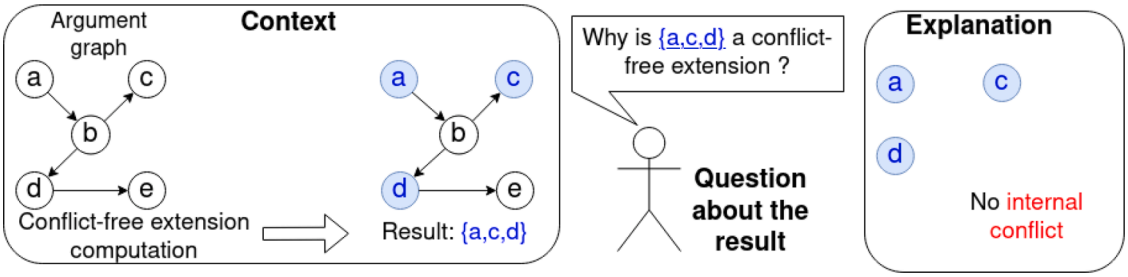


Fig. 10. Explanation G_{CF} about the conflict-freeness (here $S' = S$).

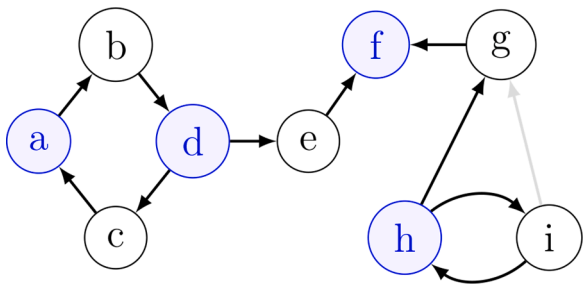


Fig. 11. Explanation G_{Def} on why $\{a, d, h, f\}$ defends all its arguments in Fig. 1. All the attackers of a, d, h and f are attacked in return.

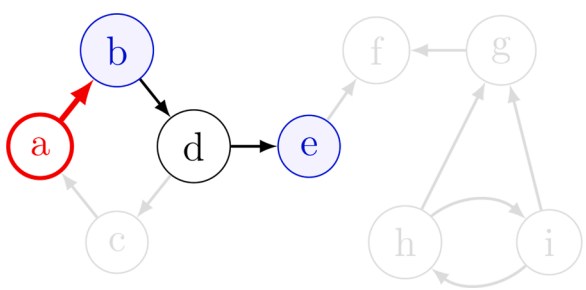


Fig. 12. Explanation G_{Def} on why $\{b, e\}$ does not only contain arguments that are acceptable w.r.t. $\{b, e\}$ in Fig. 1. b is attacked by a and a is not attacked by b or e in return.

(and so, part of it) or not defended (and so, not part of it). We begin by illustrating on some examples to give the intuition, and then formally define these explanations.

Example 6. Consider the argumentation framework of Fig. 1. Figs. 14 and 15 show why $\{a, d, h, f\}$ accepts all the arguments it defends and why $\{b, c\}$ does not. In Fig. 14, we can see that all the arguments that can be reached in two steps via the attack relation from $\{a, d, h, f\}$ are in fact a, d, h and f themselves. If we consider the attackers of these arguments, we observe that they are all attacked by a, d, h or f. Since these arguments are all included in the set, we can conclude that $\{a, d, h, f\}$ accepts all the arguments it defends. In Fig. 15, we see that e is defended by b, but e does not belong to the set $\{b, c\}$. Hence, $\{b, c\}$ does not accept all the arguments it defends.

Following the examples, we realise that we must compute a subgraph

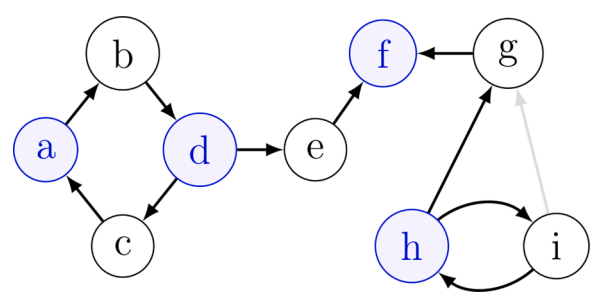


Fig. 14. Explanation G_{Reins2} on why $\{a, d, h, f\}$ accepts all the arguments it defends in Fig. 1. The arguments $\{a, d, h, f\}$ could defend are those that can be reached in two steps via the relation from either a, d, h or f. They are in fact a, d, h and f themselves, since they are all defended and all belong to the set.

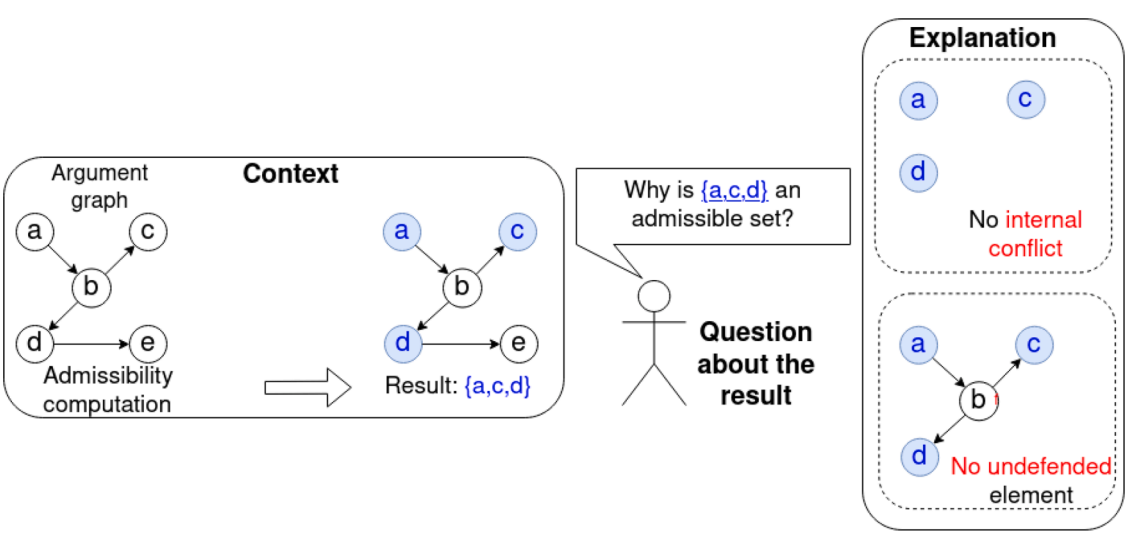


Fig. 13. Explanation G_{Adm} about admissibility (here $S' = S$).

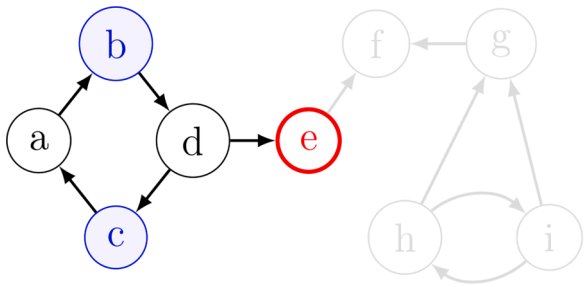


Fig. 15. Explanation G_{Reins2} on why $\{b, c\}$ does not accept all the arguments it defends in Fig. 1. e is defended by b but does not belong to the set.

of the argumentation framework induced by the set of arguments the question is about, and the arguments this set could defend along with the attackers of these arguments. Moreover, the only arcs that are relevant are those from the attackers to the arguments the set could defend (to show that they are indeed attackers) and from the set to these attackers (to show whether the set indeed defends these arguments or not). So the corresponding answer will be a specific spanning subgraph of the induced subgraph.

Moreover, we also must take into account the unattacked arguments that must be always accepted. These elements could be missing in the subgraph evoked previously. So another subgraph is needed. This implies that the reinstatement principle can be viewed as two sub-principles:

Definition 12. (*Explanation for reinstatement*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The two relevant subgraphs to explain whether or not S' contains all arguments that are acceptable w.r.t. S' are

$$G_{Reins1}(S') = \mathscr{AF}[\{a \in A | R^-(a) = \varnothing\}]_V$$

$$G_{Reins1}(S') = \left| \begin{array}{l} (\mathscr{A}F[S' \cup R^2(S') \cup R^{-1}(R^2(S'))]_V) \\ \quad [\{(a,b) \in R \mid a \in R^{-1}(R^2(S')) \text{ and } b \in R^2(S')\} \\ \quad \cup \{(a,b) \in R \mid a \in S' \text{ and } b \in R^{-1}(R^2(S'))\}]_E \end{array} \right.$$

The checking procedure for G_{Reins1} is to verify that every argument in the resulting subgraph belongs to S'.

The checking procedure for G_{Reins2} is to verify that for every argument in $R^2(S')$ in the resulting subgraph, if it is not in S' then at least one of its attackers is not the endpoint of an edge whose origin is in S'.

Much like with the explanation for admissibility, we now have what we need to define the explanation for completeness. Once again, this explanation is the set of its different components.

Definition 13. (*Explanation for completeness*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant set of subgraphs (with their checking procedures) to answer the question "Why is S' a complete extension ?" is

$$G_{Co}(S') = \{G_{CF}(S'), G_{Def}(S'), G_{Reins1}(S'), G_{Reins2}(S')\}$$

Example 7. Consider the argumentation framework of Fig. 1 and the question "Why is {a,d,h,f} a complete extension ?". Fig. 16 shows the answer for this question.

Some properties of G_{Reins1}, G_{Reins2} and G_{Co} are given in Section 5.5.

5.4. Explanation for Stability

The last semantics we turn to is the stable semantics. Recall that a set of arguments is stable if and only if it is conflict-free and if it attacks all the arguments that do not belong to it. Therefore, part of the explanation for stability is the explanation for conflict-freeness. The other part involves the attack from S to its complement. Consequently, if we want to show why a set S of arguments is stable, we must show a part of the graph either highlighting S and the attacks from S to all the other arguments, or highlighting the set of the other arguments and all the elements of S which attack them. We choose here to characterise the former explanation, but the latter can be characterised and be relevant as well. We begin by illustrating on some examples the intuition, and then formally define these explanations.

Example 8. Consider the argumentation framework of Fig. 1. Fig. 17 shows why $\{a, d, h, f\}$ attacks all the other arguments in the argumentation framework and Fig. 18 shows why $\{b, c, i\}$ does not. In Fig. 17, we can see that there is an arc from a, d, h or f[12] to every other argument in the argumentation framework. In Fig. 18 however, we see that there are some arguments that are attacked by neither b, nor c, nor i (namely, e and f).

Following the examples, we realise that we must compute a subgraph of the argumentation framework that includes all the arguments. However, the only arcs that are relevant are those from the set the question is about to any argument that is not in that set. So we need to compute a specific spanning subgraph.

Definition 14. (*Explanation for complement attack*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant subgraph to explain whether or not S' attacks its complement is

$$G_{CA}(S') = \mathscr{AF}[\{(a,b) \in R \mid a \in S' \text{ and } b \notin S'\}]_E$$

The checking procedure for G_{CA} is to verify that all arguments not belonging to S' are attacked by an argument of S'.

In a similar fashion to the previous semantics, we can now use the explanation for complement attack and the explanation for conflict-freeness, to define the explanation for stability.

Definition 15. (*Explanation for stability*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S' \subseteq A$. The relevant set of subgraphs (with their checking procedures) to answer the question "Why is S' a stable extension ?" is

$$G_{Sta}(S') = \{G_{CF}(S'), G_{CA}(S')\}$$

Example 9. Consider the argumentation framework of Fig. 1 and the question "Why is {a,d,h,f} a stable extension?". Fig. 19 shows the answer for this question.

To conclude on the explanation for stability, we give a similar proposition on the explanation for complement attack as the one that was given on the explanation for defence.

Proposition 3. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. $G_{CA}(S')$ is a bipartite graph such that there exists a partition of its vertices with S' as one of its parts and all vertices in S' are sources in it.*[13]

Like the proposition on the explanation for defence, this proposition formalises a visual behaviour of the explanation for complement attack.

[12] In this example, f does not attack any argument, so its presence in the explanation may seem useless. Nevertheless, in a first step, it seems important to keep in each explanation all the elements of the set of interest to the user, in order to give them the most complete view about the properties of this set. In future works, when we will study the notion of minimal explanations, this constraint should be relaxed.

[13] In $G_{CA}(S')$, since any vertex in S' is a source, by definition of bipartite graphs, any vertex in the complement part of S' is a sink.

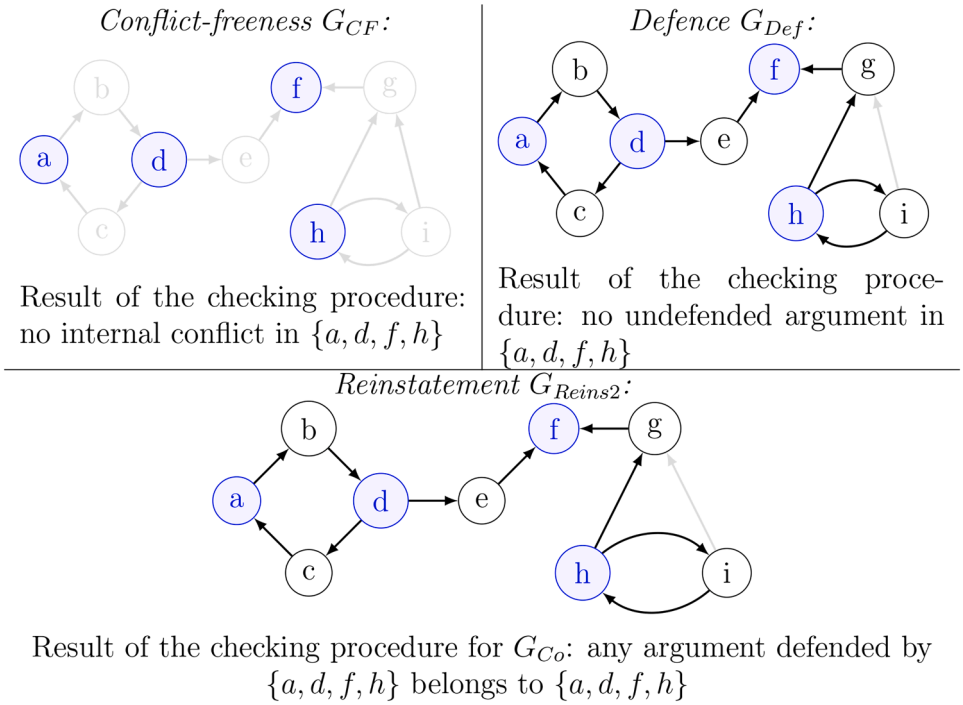


Fig. 16. Explanation G_{Co} on why $\{a, d, h, f\}$ is complete in Fig. 1. It is the sequence of Figs. 8, 11 and 14. Note that, in this example, the subgraph for the defence and the subgraph for the reinstatement are identical but not the checking procedures. Note also that the subgraph corresponding to G_{Reins1} is not given here since it is empty (there is no unattacked argument in the running example).

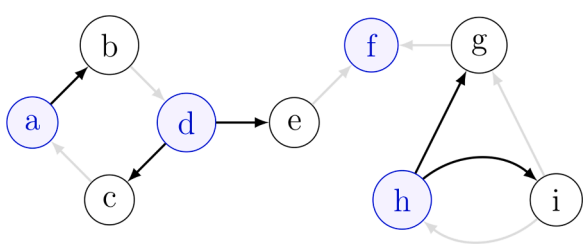


Fig. 17. Explanation G_{CA} on why $\{a, d, h, f\}$ attacks all the other arguments in Fig. 1. For every other argument in the argumentation framework, there is an arc from a, d, h or f to that argument.

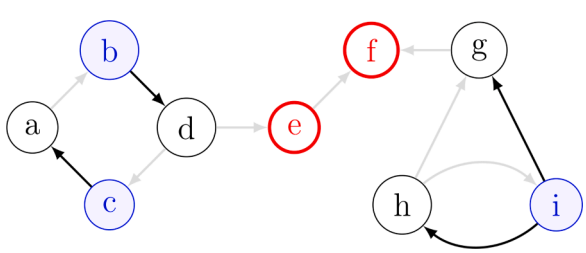


Fig. 18. Explanation G_{CA} on why $\{b, c, i\}$ does not attack all the other arguments in Fig. 1. e and f are not attacked by b, c or i.

This one, however, is unconditional and more precise.

Some other properties of G_{CA} and G_{Sta} are given in the next section.

5.5. Results on Explanations for Semantics Extensions

In this section, we provide some results on the explanations we defined previously.[14] We begin with soundness results. These results establish a link between visual properties of the explanations (described in the checking procedure given in the corresponding definition) and the answer they provide to the question that brought their computation.

We begin with the condition of conflict-freeness.

Theorem 1. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. S' is conflict-free iff C_{CF} is satisfied by S', with C_{CF}: "there are no attacks in $G_{CF}(S')$".*[15]

So, in order to know whether a set S' of arguments is conflict-free or not, one might just compute $G_{CF}(S')$ and see whether attacks are present or not.

We continue with the condition of defence.

Theorem 2. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$ be a conflict-free set of arguments. $S' \subseteq F_{\mathscr{AF}}(S')$ iff C_{Def} is satisfied by S', with C_{Def}: "there are no source vertices among $R^{-1}(S')$ in $G_{Def}(S')$".*[16]

So, in order to know whether a conflict-free set S' of arguments defends all its arguments or not, one might just compute $G_{Def}(S')$ and look at the attackers of S'. If one of them has no arc pointing towards it, S' does not defend all its arguments, otherwise it does.

We go on with the condition of reinstatement. For this condition, the results do not take the form of an equivalence but of two implications.

Theorem 3. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. If C_{Reins1} and C_{Reins2} are satisfied by S' then $F_{\mathscr{AF}}(S') \subseteq S'$, with C_{Reins1}: "all vertices in $G_{Reins1}(S')$ are in S'" and C_{Reins2}: "all vertices in $R^2(S') \backslash S'$ are the endpoint of an arc whose origin is a source vertex in $G_{Reins2}(S')$".*

So, by computing $G_{Reins1}(S')$ and verifying that all its vertices are in S', and by computing $G_{Reins2}(S')$ and verifying that the arguments that S'

[14] Let us recall that all the proofs can be found in Doutre et al. (2022).
[15] C_{CF} corresponds exactly to the checking procedure given in Definition 9.
[16] C_{Def} corresponds exactly to the checking procedure given in Definition 10.

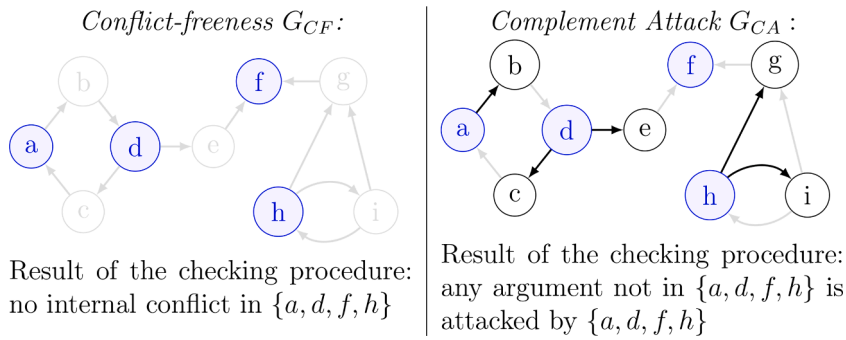


Fig. 19. Explanation G_{Sta} on why $\{a, d, h, f\}$ is stable in Fig. 1.

defends but are not in S' are all targeted by a source vertex, one verifies that S' contains all the arguments that are acceptable w.r.t. S'.

Theorem 4. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. If $F_{\mathscr{AF}}(S') \subseteq S'$ then C_{Reins1} and C'_{Reins2} are satisfied by S', with C'_{Reins2}: "all vertices in $R^2(S') \backslash S'$ are the endpoint of an arc whose origin is a source vertex or is in $R^2(S')$, in $G_{Reins2}(S')$".*

So, if we compute $G_{Reins1}(S')$ on a set S' of arguments which we know contains all the arguments that are acceptable w.r.t. it, we know that all the arguments in $G_{Reins1}(S')$ will be contained in S'. Likewise, if we compute $G_{Reins2}(S')$ on a similar set, we know that all the arguments that S' defends but which are not in S' will be targeted by a source vertex or a vertex of $R^2(S')$.

From theorems 3 and 4 follows the next corollary, which shows an equivalence result:

Corollary 1. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$ such that $R^2(S')$ is conflict-free. $F_{\mathscr{AF}}(S') \subseteq S'$ iff C_{Reins1} and C_{Reins2} are satisfied by S'.*[17]

We finish the soundness results with the condition of complement attack.

Theorem 5. *Let $\mathscr{AF} = (A, R)$ and $S' \subseteq A$. $A \backslash S' \subseteq R^+(S')$ iff C_{CA} is satisfied by S', with C_{CA}: "there are no isolated vertices in the complement part of S' in $G_{CA}(S')$".*[18]

So, in order to know if a set S' of arguments attacks its complement or not, one might just compute $G_{CA}(S')$ and look at the arguments not in S'. If one of them is isolated, S' does not attack its complement, otherwise it does.

The following proposition gives a unicity result for the explanations. In particular, it shows that the computation of the explanation is deterministic, in the sense that only one explanation is computed for one given set.

Proposition 4. *Let $\mathscr{AF} = (A, R)$, $S', S'' \subseteq A$ and $\sigma \in \{CF, Def, Reins_1, Reins_2, CA\}$. If $S' = S''$, then $G_\sigma(S') = G_\sigma(S'')$.*

Finally, we provide a result showing that our explanations are empty (i.e. equal to the empty graph) only in very particular situations. For some of them, it only happens when computed on an empty set S'. For $G_{Reins1}(S')$, it happens when the initial argumentation framework does not contain unattacked arguments. For $G_{CA}(S')$, it only happens when the initial argumentation framework is itself the empty graph.

Proposition 5. *Let $\mathscr{AF} = (A, R)$, $S' \subseteq A$. Let $\sigma \in \{CF, Def, Reins_2\}$: $G_\sigma(S') = (\varnothing, \varnothing)$ iff $S' = \varnothing$. $G_{Reins1}(S') = (\varnothing, \varnothing)$ iff $\{a | R^-(a) = \varnothing\} = \varnothing$. $G_{CA}(S') = (\varnothing, \varnothing)$ iff $\mathscr{AF} = (\varnothing, \varnothing)$.*

6. Motivation and Hypotheses: Complement

Section 4 motivated the issues which have been addressed in the first part of this paper: questions and explanations regarding the acceptability of extensions.

It is now the acceptance of arguments, their membership in given extensions provided by the system, which will be addressed in this second part of the paper. We start with motivating the questions and the explanations that will be provided, and additional hypotheses which should be considered.

As introduced at the end of Section 4.1, other questions naturally arise when facing the real-case situation of delivery and potential fraud described in Example 2 and Fig. 4.

Remember that A_1 concluded that it is not a case of fraud, whereas A_6 concluded that it is. One may wonder "Why is A_6 (or A_1) acceptable?" in the sense of credulous acceptance, that is, in terms of an acceptable extension which contains the argument.

Many approaches have tackled this question as shown by the related works (Section 3). The explanations to this question show the attackers of the considered arguments, and show how they can be defended against these attackers.

In this paper, we take a different point of view on these questions, different in two respects. First, we consider the question in the case where an acceptable set under the semantics is given as a context (recall Hypothesis (H1)), that is, we consider questions such as "Why is A_6 accepted in $\{A_6, A_3, A_4\}$?". Second, the answers that we provide highlight the *necessity* of the presence of the argument in the context set, by showing the consequence of the absence of the argument from the set, on the acceptability of the context set. In other words, it is the question "Why is $\{A_3, A_4\}$ (not) a complete extension?" which is addressed when the above question is asked. $\{A_3, A_4\}$ actually is a complete extension: it is conflict-free, it defends all its elements, and, if A_6 is defended against A_2 and A_5 by A_4 and A_3, it is A_6 which defends itself against A_1; hence all the arguments which are defended by $\{A_3, A_4\}$ belong to the set. The presence of A_6 is not necessary for the set to be acceptable; this is the answer that our approach will give. The usual point of view on such a question would be to show how A_6 is defended (by the set under consideration) against all its attackers, that is, to show a *sufficient* reason for the acceptability of A_6.

If our necessity-based approach may seem not fully satisfying when it comes to explain why A_6 is accepted in $\{A_6, A_3, A_4\}$, it is when the question is "Why is A_3 accepted in $\{A_6, A_3, A_4\}$?": $\{A_6, A_4\}$ is not a complete extension; A_6 is attacked by A_5, and it is not defended. An explanation subgraph will highlight the attack from A_5 to A_6, and the fact that A_5 is not attacked back by the set. The presence of A_3 is necessary in the considered set as it defends A_6 against A_5. Showing this necessary role of the argument is not common in the existing approaches; this is an originality of this work.

Let us consider now a question about the non-acceptance of an

[17] C_{Reins1} and C_{Reins2} correspond exactly to the checking procedures given in Definition 12.

[18] C_{CA} corresponds exactly to the checking procedure given in Definition 14.

argument. The question "Why is A_1 not accepted in $\{A_6, A_3, A_4\}$?" will amount to wonder, in our approach, whether the set augmented with A_1 is or is not a complete extension; the question "Why is $\{A_6, A_3, A_4, A_1\}$ not a complete extension?" will be addressed. As we have already shown, A_1 and A_6 attack each other, making any set containing the two of them not conflict-free, and then not a complete extension. The question regarding the non-acceptance of an argument in an extension will thus consist in checking if the considered argument prevents the set from being acceptable. If in this example, it is the case, it may happen that it is not. For instance, consider the question "Why is $\{A_3, A_4\}$ not accepted in $\{A_1\}$?". $\{A_1, A_3, A_4\}$ is a complete extension, and the answer to the question will show how it is so. $\{A_3, A_4\}$ is not accepted in $\{A_1\}$, but these arguments may be considered altogether and form an acceptable set.

As for the question on acceptance, by focusing on the "*toxicity*" of an argument or a set of arguments in an extension, the answer here may not seem fully satisfying. A complementary answer to this question, that we do not provide in this paper, may consist in showing how $\{A_1\}$ is an acceptable set by itself.

In addition to these questions on acceptance and non-acceptance of an argument or of a set of arguments in a given extension, *contrastive* questions may be considered as well. "Why is A_6 accepted in $\{A_6, A_3, A_4\}$ and not A_1?" is an example of such a question. As before, these questions will focus on the necessity of the presence of the arguments or on their toxicity. Hence, the above question will come down to replacing A_6 by A_1 in the considered set, and to checking whether this set is a complete extension. $\{A_1, A_3, A_4\}$ is a complete extension. The answer to the question that our approach provides, is that A_6 may well be replaced by A_1 in the set, and the set would still be acceptable. A_6 is not necessary in the set, and when removed, A_1 is not toxic. A complementary answer, that we do not consider in this paper, may consist in giving a sufficient condition on why A_6 is accepted in the initial set, and a sufficient condition of why A_1 is not accepted in the resulting set.

The approach that we choose can be further motivated. In general, it is common to consider that a question of the form *"Why P ?"* is in fact a question of the form *"Why P and not Q ?"*, with Q left implicit. P and Q are often referred to as the *fact* and the *foil* respectively (see Miller (2019)). Thus, the key to this kind of question is to be able to identify the implicit foil.

Now, in our case, the minimal question is "Why is S P ?" with S a subset of arguments and P a property on S. The natural foil to this kind of question would then be that S does not enjoy property P. In other words, the question "Why is S P ?" is in fact the question "Why is S P and not $\overline{P}$?" with $\overline{P}$ representing the absence of property P. So we can consider here that the fact is "S being P" and the foil is "S being $\overline{P}$" (i.e. "S not being P"). So, the approach we use to answer questions relies on the following hypothesis:

To explain "S being P" is to show that "S not being P" is not possible. (H5).

Put it differently, to explain "S being P" is to show its *necessity*. This is already the approach we used in Section 5 on questions related to why a given set of arguments is (or not) an extension of some semantics. Independently from what the user perceives to be true, it holds that either "S being P" or "S being $\overline{P}$" is true, but not both. The choice we made was to show the element supporting the truth, and thus that its contrary is not possible. In the case of questions related to why an argument or set of arguments is (resp. is not) part of an extension of some semantics, we adopt the point of view of showing that the argument or set of arguments cannot not be part (resp. be part) of that extension. That is, we directly compute the foil suggested by the question and show it to the user so that they may come themselves to the conclusion that the argument or set of arguments must (resp. must not) be part of that extension.

In what follows, we will call *positive questions* the questions related to the membership/acceptance of an argument or set of arguments in an extension (Why is X accepted?, with X being either an argument x or a set of arguments S). We will call *negative questions* the questions related to the non-membership/acceptance (Why is X not accepted?).

Contrastive questions will be further described in Section 7.2. A hypothesis which should however be mentioned from now on:

If a contrast is made in the question, the contrast is only made on the arguments. (H6).

This assumption is specific to the questions regarding the acceptance of arguments or sets of arguments, and does not apply to the entire document. It only has the effect of focusing on a certain number of contrastive questions and not on all that are possible in this specific context. "Why is A_6 accepted in $\{A_6, A_3, A_4\}$ and not a stable extension ?" is an example of question that we do not consider in this setting.

7. Providing Explanations for Extension Membership

Following the motivation described in the previous section, we will present non-contrastive questions and their answers, before studying contrastive questions. Note that the notations used in this section are those introduced before in Section 5.1. For the legend of the figures, we use almost the same legend as the one presented in Section 5.1, except for the blue elements that are not the elements the user asks about but the elements corresponding to the question which is really addressed (see Section 6).

7.1. Non-contrastive Questions

In this section, we precisely define the answers to non-contrastive questions on acceptance. Taking into consideration the variation on elements of interest in the questions (an argument or a set or arguments), there are two possible positive non-contrastive questions and two possible negative non-contrastive questions. The case of a single argument equivalent to the case of a singleton set containing this argument.

The questions are answered using the principles illustrated above. We begin with the positive questions:

Definition 16. (*Explanation for positive non-contrastive question on acceptance of a set of arguments*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question "Why is S' accepted ?" are given by the explanation for σ on $S \backslash S'$.

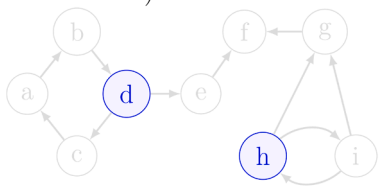
Following Section 6, the question "Why is S' accepted ?" is in fact the question "Why is S \ S' not a σ-extension ?", so in the subgraph which will be presented as an explanation, only the arguments of $S \backslash S'$ will be in blue (the explanation amounting to show why this set is or is not an extension under σ).

Example 10. Consider the argumentation framework of Fig. 1 and the admissible extension $\{a, d, h\}$. Suppose the user asks the question "Why is a accepted ?". Fig. 20 shows the corresponding answer: $\{d, h\}$ without a is not admissible (it is conflict-free but it does not respect the defence property) and thus a is necessary in the extension presented by the system.[19]

We turn now to negative questions. The methodology is very similar to the positive questions, the difference being that instead of removing arguments from the extension, we add them.

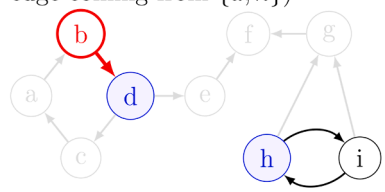
[19] Note that, for some element the user is interested in, it could happen that the removal of this element does not produce any effect. This point has been introduced in Section 6 and it will be further discussed in Section 10.

Step 1, explanation G_{CF}: conflict-freeness of $\{d, h\}$ (checking procedure: no arc)



Result of the checking procedure: OK
So the absence of a is not a problem for the conflict-freeness of $\{d, h\}$

Step 2, explanation G_{Def}: defence for $\{d, h\}$ (checking procedure: every argument $\notin \{d, h\}$ is the endpoint of an edge coming from $\{d, h\}$)



Result of the checking procedure: NOK
So the absence of a is a problem for the defence of $\{d, h\}$

Fig. 20. Explanation on why a is accepted in the admissible extension $\{a, d, h\}$: explanation G_{Adm} on why $\{d, h\}$ is not admissible.

Definition 17. (*Explanation for negative non-contrastive question on acceptance of a set of arguments*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question "Why is S′ not accepted ?" are given by the explanation for σ on $S \cup S'$.

Following Section 6, the question "Why is S′ not accepted ?" is in fact the question "Why is S ∪ S′ not a σ-extension ?", so in the subgraph which will be presented as an explanation, the arguments of $S \cup S'$ will be in blue (the explanation amounting to show why $S \cup S'$ is or is not an extension under σ).

Example 11. Consider the argumentation framework of Fig. 1 and the complete extension $\{h\}$. Suppose the user asks the question "Why is a not accepted ?". Fig. 21 shows the answer for this question.

7.2. Contrastive Questions

In this section we turn to precisely define the answers to contrastive questions on acceptance. We begin by extending the distinction made between questions by the presence/absence of a negation on the property to contrastive questions, using the property of the contrastive statement as well as the property of the reference statement. This yields to four types of questions: positive-positive, positive-negative, negative-positive and negative-negative. Taking into consideration the variation on elements of interest in the questions, there are two possible questions for each type. Please note that since we use "and not" as the connection between the reference statement and the contrast statement, we must reverse the type of the property expressed in the contrast statement.

Example 12. The question "Why is a not accepted and not b accepted ?" is a negative-negative question, although the contrast statement's property is expressed as positive. Note also the case of an implicit information: the question "Why is a accepted and not b ?" must be understand as "Why is a accepted and not b accepted ?" that is a positive-negative question.

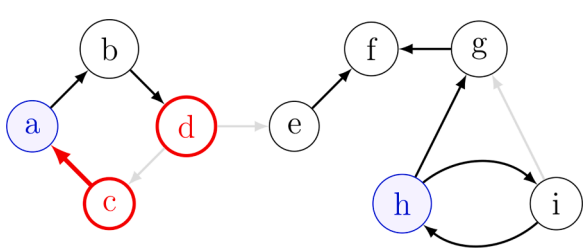


Fig. 21. Explanation on why a is not accepted in the complete extension $\{h\}$: explanation G_{Co} on why $\{a, h\}$ is not complete. Considering $\{a, h\}$, the conflict-freeness is satisfied (no arc between a and h), but the defence property is not (a cannot be defended by h or by itself), nor is the reinstatement property (a defends d, which is not in the considered set). So the checking procedures corresponding to the step of defence and to the step of reinstatement fail.

To provide answers to these questions, we treat the contrast statement the same way as the reference statement. Note that a contrastive question is *not* a sequence of two questions; there is a specificity given by the contrast that influences the building of the answer. So the resulting graph can be obtained using a combination of the previous definitions given for the reference statement and for the contrast statement. We will only cover the case in which the element of interest is a set of arguments. The case in which it is one argument is dealt with in the same way as in the previous section.

Definition 18. (*Explanation for positive-positive contrastive question on acceptance of a set of argument*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ be an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question "Why is S′ accepted and not S″ not accepted ?" are given by the explanation for σ on $(S \backslash S') \backslash S''$.[20]

Following the remarks given in Section 6, the question "Why is S′ accepted and not S″ not accepted ?" is in fact the question "Why is (S \ S′) \ S″ not a σ-extension ?". So, in the subgraph which will be presented as an explanation, only the arguments of $(S \backslash S') \backslash S''$ will be in blue (the explanation amounting to show why such a set is or is not an extension under σ).

Example 13. Consider the argumentation framework of Fig. 1 and the stable extension $\{b, c, e, h\}$. Suppose the user asks the question "Why is {b} accepted and not {c} not accepted ?". Fig. 22 shows the answer for this question.

Definition 19. (*Explanation for positive-negative contrastive question on acceptance of a set of argument*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ be an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question "Why is S′ accepted and not S″ ?" are given by the explanation for σ on $(S \backslash S') \cup S''$.[21]

Following Section 6, the question "Why is S′ accepted and not S″ ?" is in fact the question "Why is (S \ S′) ∪ S″ not a σ-extension ?". So, in the

[20] So it is the combination of Def. 16 over the reference statement and Def. 16 over the contrast statement.

[21] So it is the combination of Def. 16 over the reference statement and Def. 17 over the contrast statement.

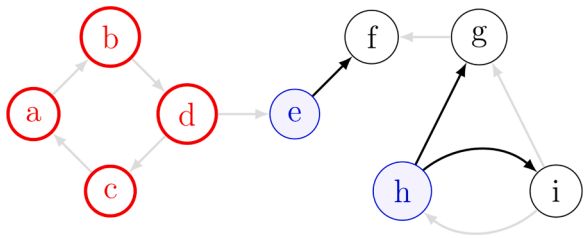


Fig. 22. Explanation on why $\{b\}$ and $\{c\}$ are accepted in the stable extension $\{b,c,e,h\}$: explanation G_{Sta} on why $\{e,h\}$ is not stable. Without b and c, e and h alone fail to attack every other argument in the argumentation framework. Hence, b and c are necessary in this stable extension. So it is the checking procedure corresponding to the step of complement attack that fails.

subgraph which will be presented as an explanation, the arguments of $(S\backslash S^{'}) \cup S''$ will be in blue (the explanation amounting to show why such a set is or is not an extension under σ).

Example 14. Consider the argumentation framework of Fig. 1 and the conflict-free extension $\{a,d,h\}$. Suppose the user asks the question "Why is {d,h} accepted and not {b,c} ?". Fig. 23 shows the answer for this question.

Definition 20. (*Explanation for negative-positive contrastive question on acceptance of a set of argument*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ be an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question "Why is S$^{'}$ not accepted and not S$''$?" are given by the explanation for σ on $(S \cup S^{'})\backslash S''$.[22]

Following Section 6, the question "Why is S$^{'}$ not accepted and not S$''$?" is in fact the question "Why is (S $\cup$ S$^{'}$) $\setminus$ S$''$ not a σ-extension ?". So, in the subgraph which will be presented as an explanation, the arguments of $(S \cup S^{'})\backslash S''$ will be in blue.

Example 15. Consider the argumentation framework of Fig. 1 and the stable extension $\{a,d,i,f\}$. Suppose the user asks the question "Why is $\{b, c\}$ not accepted and not $\{a, d\}$?". Fig. 24 shows the answer for this question.

Definition 21. (*Explanation for negative-negative contrastive question on acceptance of a set of argument*) Let $\mathscr{AF} = (A, R)$ be an argumentation framework and $S \subseteq A$ be an extension of $\mathscr{AF}$ for semantics σ. The relevant subgraphs and checking procedures to answer the question

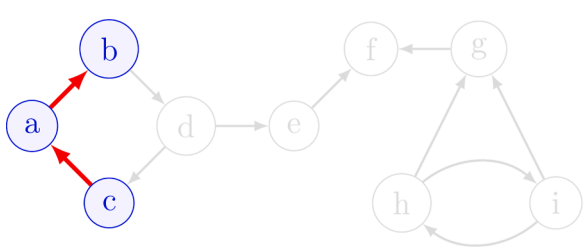


Fig. 23. Explanation of why $\{d,h\}$ is accepted and not $\{b,c\}$ in the conflict-free extension $\{a,d,h\}$: explanation G_{CF} on why $\{a,b,c\}$ is not conflict-free. We see that removing d and h while adding b and c gives rise to internal conflicts in the extension. So it is the checking procedure corresponding to the step of conflict-freeness that fails.

[22] So it is the combination of Def. 17 over the reference statement and Def. 16 over the contrast statement.

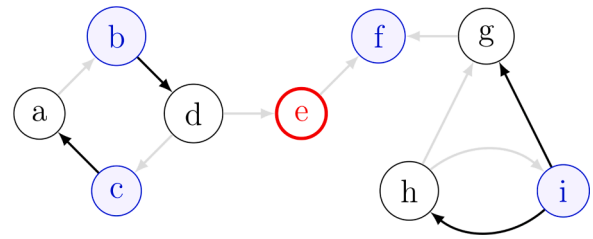


Fig. 24. Explanation on why $\{b,c\}$ is not accepted and not $\{a,d\}$ in the stable extension $\{a,d,i,f\}$: explanation G_{Sta} on why $\{i,f,b,c\}$ is not stable. If we add b and c to the extension while removing a and d we see that the extension fails to meet the conditions for being stable (e is not attacked by any argument of the extension). So it is the checking procedure corresponding to the step of complement attack that fails.

"Why is S$^{'}$ not accepted and not S$''$ accepted ?" are given by the explanation for σ on $(S \cup S^{'}) \cup S''$.[23]

Following Section 6, the question "Why is S$^{'}$ not accepted and not S$''$ accepted ?" is in fact the question "Why is (S $\cup$ S$^{'}$) $\cup$ S$''$ not a σ-extension ?". So, in the subgraph which will be presented as an explanation, the arguments of $(S \cup S^{'}) \cup S''$ will be in blue.

Example 16. Consider the argumentation framework of Fig. 1 and the complete extension $\{b, c, e\}$. Suppose that the user asks the question "Why is {h,f} not accepted and not {d} accepted ?". Fig. 25 shows the answer for this question.

8. Summary on Answers

We give in this section a summary of the results of our approach.

First of all, our assumptions (the two last ones concerning only a category of questions, those about acceptance):[24]

(H1): *A user asks for an explanation after they have been presented the result of a Formal Argumentation process (typically the selection of arguments via a semantics) by a program that we will refer to as* the system.
(H2): *The user is able to understand argumentation frameworks.*
(H3): *The user knows Abstract Argumentation semantics.*
(H4): *Grice's maxims are correct and should thus be followed when engaging on the explanation of the result of the system.*
(H5): *To explain "S being P" is to show that "S not being P" is not possible.*

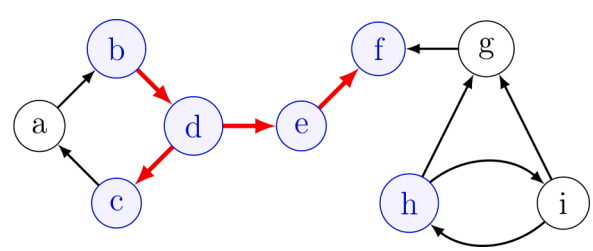


Fig. 25. Explanation on why $\{h,f\}$ and $\{d\}$ are not accepted in the complete extension $\{b,c,e\}$: explanation G_{Co} on why $\{b,c,e,h,f,d\}$ is not complete. If h, f and d are added to the extension, internal conflicts appear. So it is the checking procedure corresponding to the step of conflict-freeness that fails.

[23] So it is the combination of Def. 17 over the reference statement and Def. 17 over the contrast statement.

[24] Note that the assumption (H5) could be also used in the questions about semantics extensions. Nevertheless, due to the way our answers are built, it is useless.

(H6): *If a contrast is made in the question, the contrast is only made on the arguments.*

Then, let $\mathscr{AF} = (A, R)$ be the argumentation framework that is the main component of the system, the questions we are interested in:

Questions about semantics extensions: let S be the set of arguments produced by the system for a given semantics σ (σ can be either the conflict-freeness, or the admissibility, or the completeness, or the stability), and let consider that the user asks about S' that denotes a set of arguments,

Q1"Why is S' [not] an extension for semantics σ?" (S' can be or not equal to S)

Note that Question **Q1** will be instantiated wrt σ. Moreover this question uses some subquestions related to the principles behind the semantics (defence, reinstatement and complement attack).

Questions about extension membership: let S be the set of arguments produced by the system for a given semantics σ and let consider that the user asks about S', S'' that denote sets of arguments (eventually singletons),

Q2 (question +): Why is S' accepted?
Q3 (question -): Why is S' not accepted?
Q4 (question $+\ +$): Why is S' accepted and not S'' not accepted?
Q5 (question $+\ -$): Why is S' accepted and not S'' accepted?
Q6 (question $-\ +$): Why is S' not accepted and not S'' not accepted?
Q7 (question $-\ -$): Why is S' not accepted and not S'' accepted?

The answers corresponding to Question **Q1** are given in Tables 2 and 3.

The answers corresponding to questions **Q2** to **Q7** are given in Table 4.

It is worth noting that some contrastive questions are equivalent to non contrastive ones:

- given that $(S \backslash S') \backslash S'' = S \backslash (S' \cup S'')$, **Q4**: Why is S' accepted and not S'' not accepted? is equivalent to **Q2**: Why is $S' \cup S''$ accepted?. So a positive-positive question can be expressed as a positive question.
- given that $(S \cup S') \cup S'' = S \cup (S' \cup S'')$, **Q7**: Why is S' not accepted and not S'' accepted? is equivalent to **Q3**: Why is $(S' \cup S'')$ not accepted?. So a negative-negative question can be expressed as a negative question.

From these equivalences, it follows that S' and S'' can be swapped in

Table 2
The answers given for some informal questions about the underlying principles used in semantics (the subgraph is the one that is built to answer the question and the checking procedure is always applied on this subgraph).

Sub-question about defence (Explanation G_{Def}, see Def. 10): $\mathbf{Q_{def}}$ "Does S' contain only arguments that are acceptable w.r.t. S'?"	
Subgraph:	$(\mathcal{AF}[S \cup R^{-1}(S')]_V)$ $[\{(a,b) \in R \mid a \in R^{-1}(S')$ and $b \in S',$ or $a \in S'$ and $b \in R^{-1}(S')\}]_E$
Checking Proc.:	(C_{Def}, see Theo. 2) every argument that does not belong to S' in the resulting subgraph is the endpoint of an arc whose origin is in S'
Sub-question about reinstatement (Explanations G_{Reins1}, G_{Reins2}, see Def. 12): $\mathbf{Q_{reins}}$ "Does S' contain all arguments that are acceptable w.r.t. S'?"	
Subgraph:	$\mathcal{AF}[\{a \in A \mid R^{-}(a) = \varnothing\}]_V$
Checking Proc.:	(C_{Reins1}, see Theo. 3) every argument in the subgraph must belong to S'
Subgraph:	$(\mathcal{AF}[S \cup R^2(S') \cup R^{-1}(R^2(S'))]_V)$ $[\{(a,b) \in R \mid a \in R^{-1}(R^2(S'))$ and $b \in R^2(S')\}$ $\cup\ \{(a,b) \in R \mid a \in S'$ and $b \in R^{-1}(R^2(S'))\}]_E$
Checking Proc.:	(C_{Reins2}, see Theo. 3) for every argument in $R^2(S')$, if it is not in S' then at least one of its attackers is not the endpoint of an arc whose origin is in S'
Sub-question about complement attack (Explanation G_{CA}, see Def. 14): $\mathbf{Q_{compAtt}}$ "Does S' attack its complement?"	
Subgraph:	$\mathcal{AF}[\{(a,b) \in R \mid a \in S'$ and $b \notin S'\}]_E$
Checking Proc.:	(C_{CA}, see Theo. 5) all arguments not belonging to S' are attacked by an argument of S'

Table 3
The answers given for Question **Q1**, for each semantics. For some semantics, several subgraphs (and checking procedures) exist in the explanation, one for each underlying principle used in this semantics (S' is an element given by the user in their question).

Q1 for conflict-freeness (Explanation G_{CF}, see Def. 9):	
"Why is S' [not] a conflict-free extension?"	
Subgraph:	$\mathscr{AF}[S']_V$
Checking Proc.:	(C_{CF}, see Theo. 1) The set of edges is empty
Q1 for admissibility (Explanation G_{Adm}, see Def. 11):	
"Why is S' [not] an admissible extension?"	
Subgraph:	The ones for **Q1** about conflict-freeness G_{CF}
Checking Proc.:	
Subgraph:	The ones for the question about defence G_{Def}
Checking Proc.:	
Q1 for completeness (Explanation G_{Co}, see Def. 13):	
"Why is S' [not] a complete extension?"	
Subgraph:	The ones for **Q1** about conflict-freeness G_{CF}
Checking Proc.:	
Subgraph:	The ones for the question about defence G_{Def}
Checking Proc.:	
Subgraph:	The ones for the question about reinstatement G_{Reins1} and G_{Reins2}
Checking Proc.:	
Q1 for stability (Explanation G_{Sta}, see Def. 15):	
"Why is S' [not] a stable extension?"	
Subgraph:	The ones for **Q1** about conflict-freeness G_{CF}
Checking Proc.:	
Subgraph:	The ones for the question about complement attack G_{CA}
Checking Proc.:	

Table 4
The answers given for questions **Q2** to **Q7** wrt a given semantics σ (S is the result presented by the system to the user before they ask for an explanation and S', S" are the elements given by the user in their question).

Q2: Why is S' accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $S \backslash S'$ as an extension for σ?"
Q3: Why is S' not accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $S \cup S'$ as an extension for σ?"
Q4: Why is S' accepted and not S'' not accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $(S \backslash S') \backslash S''$ as an extension for σ?"
Q5: Why is S' accepted and not S'' accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $(S \backslash S') \cup S''$ as an extension for σ?"
Q6: Why is S' not accepted and not S'' not accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $(S \cup S') \backslash S''$ as an extension for σ?"
Q7: Why is S' not accepted and not S'' accepted?	
Subgraph:	The ones for **Q1** applied to σ:
Check. Proc.:	"Why is $(S \cup S') \cup S''$ as an extension for σ?"

Q4 and **Q7** without any consequence on the answer which is provided. Fact and foil are treated equivalently in these negative-negative and positive-positive questions.

This can lead to a little simplification of our approach, considering only two kinds of contrastive questions, the positive-negative (**Q5**) and the negative-positive (**Q6**) questions (the other ones being transformed into non contrastive questions).

Regarding these two remaining questions, a specific case can lead to another simplification: if $S' \cap S'' = \varnothing$, then **Q5** and **Q6** lead to equivalent reformulations with an identical answer. Indeed **Q5**: Why is S' accepted and not S'' accepted? is equivalent to **Q6**: Why is S'' not accepted and not S' not accepted? and their answers are the same since $(S \backslash S') \cup S'' = (S \cup S'') \backslash S'$. Thus, in this case, considering only either **Q5** or **Q6** could be sufficient.

Finally, if the formulation of the contrastive questions that we choose imposed "and not" to introduce the contrastive part, the questions **Q4** and **Q6** may be reformulated in a more natural way: **Q4**: Why is S' accepted and S'' accepted?, **Q6**: Why is S' not accepted and S'' accepted?

9. Comparison with Related Works

Before we focus on the quality of our explanations in Section 10, we compare our work with those that were mentioned in Section 3.

In Saribatur et al. (2020), the authors define strongly rejecting subframeworks as explanations for the credulous non acceptance of some argument. Our approach differs in that we define explanations for "Why a set of arguments is (not) an extension of a given semantics?", or "Why some arguments are (not) part of an extension?". So, we are not interested in the same questions. In addition, there are some semantics not considered in our work (namely the grounded and preferred semantics), our subgraphs are not only induced subgraphs but also spanning subgraphs and our second kind of explanations are contrastive. In Niskanen and Järvisalo (2020) and Ulbricht and Wallner (2021), the authors define explanations for the credulous non acceptance and acceptance of some argument respectively as sets of arguments or attacks. Their definition is *based on* the behavior of the induced (respectively spanning) subgraph resulting from the considered set of arguments (respectively attacks). Our work instead considers the subgraphs to be the explanation itself. Moreover, the subgraphs we define are computed using both the induced subgraph and the spanning subgraph operations, while Niskanen and Järvisalo (2020) consider them separately (Ulbricht and Wallner (2021) only use induced subgraphs). Finally, the problems targeted in these works are not the same as those we target.

There also exist works that use graphs to explain, but not subgraphs. These works are Fan and Toni (2015a); Racharak and Tojo (2021) and they rely on the concept of defence trees. While not being subgraphs technically speaking, one can easily retrieve the subgraph represented by a defence tree using the original AF. Hence one could wonder what are the connections between a subgraph used as explanation and the subgraph implied by a specific defence tree. Alternatively, we could also explore the existence of specific defence trees inside a subgraph used as an explanation. So, there may exist some ties between the two approaches. Apart from the technical difference between the two methods used, a more fundamental one between the works of Fan and Toni (2015a); Racharak and Tojo (2021) and ours is that we do not explain the same problem. Indeed, Fan and Toni (2015a); Racharak and Tojo (2021) are interested in explaining the credulous acceptance of some argument under admissibility.

We turn to the works that consider changes as explanations (Fan and Toni (2015b); Niskanen and Järvisalo (2020); Ulbricht and Baumann (2019)). As noted in Niskanen and Järvisalo (2020), diagnoses can be seen as a kind of dual of the computation of induced and spanning subgraphs. Indeed, each diagnosis infers an induced or spanning subgraph, and conversely, each induced or spanning subgraph is computed using (the complement of) a diagnosis. As such, the links between the two approaches are very strong. One could thus wonder what are the properties of the complement of a set used to compute a certain induced or spanning subgraph, or what can be said about the induced or spanning subgraph computed from the complement of a given diagnosis. Although our view on explanations is closely tied to theirs, the authors of these works seek to explain different problems from those we are addressing.

We continue with the works that use sets of arguments as explanations (Baumann and Ulbricht (2021); Fan and Toni (2015a); Liao and van der Torre (2020) and the works from Borg and Bex). Although the links between subgraph-based methods of explanation and extension-based methods are less direct than with diagnosis-based methods, there are still some that can be studied. Indeed, one could wonder what are the links between a subgraph computed by a subgraph-based method and the subgraph induced by the set computed

by an extension-based method. Or, conversely, what can be said about the set that was used to compute an induced subgraph and the set computed by an extension-based method. Whether the explanation of an extension-based method is included in the explanation of a subgraph-based method can also be asked, the converse as well. Those are questions that could help explore the ties between the two methods, and which should be investigated in future work. At a more fundamental level, one may note that the subgraph induced by the set computed by an extension-based method may contain more attacks than the subgraph that may be defined directly as an explanation. This may reveal that, in extension-based explanations, arguments are considered as the only relevant elements to explain, while in subgraph-based approaches, attacks and thus the structure of the AF are also considered relevant. Borg and Bex, as well as Fan and Toni (2015a) are focused on explaining the credulous and/or skeptical (non-)acceptance of some arguments, which is not the same problem as we consider. Note that Borg and Bex (2021b) provide a notion of contrastive explanations, just like we do with our second kind of explanations. Baumann and Ulbricht (2021); Liao and van der Torre (2020) however are interested in the same problem as our first kind of explanations: explaining an extension of some semantics. Yet, there is no obvious connection between their method and ours.

10. Quality of Explanations

In this section, we wish to provide several insights on the quality of the explanations that we propose in this paper, taking into account several points of view.

To begin with, we clearly stated that we wished to adhere to Grice's maxims as much as possible. Thus, we use these maxims to evaluate our explanations.

First, consider the maxims of **Quantity** that require to say what is necessary, but also to not say what is not necessary. We claim that our explanations contain both the necessary information and unnecessary information. The former results from our choice to answer to questions of the (minimal) form "Why is e p ?", with e an element of interest (here, an argument or set of arguments) and p a property on e, by showing that "e not being p" must not be the case. The latter results from our choice to look for all reasons that could allow to affirm it. To illustrate for the case of explanations on why a set of arguments is an extension under some semantics or not, we use Figs. 8 and 9 on conflict-freeness. It is easy to see that all the necessary information is present in these explanations. On both cases, if one node (or arc) would have been discarded, we could have missed an information leading to a change of conclusion. On the other hand, one can see that in general, the explanations display more information than is necessary. In Fig. 9, the arc from d to e is sufficient to conclude that the set is not conflict-free (use Theorem 1). Thus, the node a could be removed without changing the status of the conclusion. Similar observations can be made on the other explanations on why a set of arguments is an extension under some semantics or not. Concerning explanations on the acceptance of arguments in an extension, we already argued in Section 6 that our explanations contain the necessary information. However, one can see, for instance on Fig. 20, that these explanations may also contain unnecessary information. Indeed, in this example, the main reason for $\{d,h\}$ to not be admissible after the removal of a is that d is no longer defended by a (b is a source vertex in $R^{-1}(\{d,h\})$, see Theorem 2). Hence, one could consider the information shown about h defending itself against i as superfluous.

The other categories of maxims are more straightforward. We begin with the maxims of **Quality**. First of all, the results of Section 5.5 support the soundness of our approach as they establish direct links between argumentative results and visual properties of our explanations. We deem them visual because they rely on formal notions that only deal with the structure of the computed subgraph. As such, these properties provide information about how the subgraphs are organised, or less formally, how they can be "drawn". In addition, it is obvious that, since our explanations are computed using information which is available to both the system and the user, we have all the evidence needed to support them. Moreover, we believe that computing explanations using induced and spanning subgraphs may never lead to *false* explanations, unless the original AF is twisted and modified prior to the explanations' computation, which is never done here. Concerning the maxim of **Relation**, one might consider the unnecessary information that is present in our explanations to be not relevant. On the other hand, we also make sure that all the relevant information is contained in our explanations. Finally, we believe that the category of **Manner** has more to do with a translation from our graphical explanations to a dialogue in natural language, which we are not interested in yet.

The second point of view is the suitability of our approach to the studied object. Indeed, one of the strengths of our explanations is that they are built in a **modular** way, and thus rely on the specific features of the semantics at hand. We motivate our choice by showing why, in our opinion, explanations cannot rely only on a feature that is common amongst all the semantics such as defence. Indeed, *explanations* should make more understandable how results are obtained, how they are *computed*. Thus, one of the problem with *explaining* all semantics in the same way (that is, for instance, only showing defence in every case) might infer the incorrect bias that all semantics are in fact the same, since the explanations show that they are all *computed* in the same way. Moreover, it is worth noting that not all semantics are defined *based* on this notion of defence. For instance, stable extensions are stable because they attack all the arguments that are not part of it. Thus, when asking why a set of arguments is stable, or why some arguments are part of a stable extension, or why an argument is credulously/skeptically accepted under stable semantics, it seems misplaced, or even confusing, to invoke the fact that stable extensions defend all their arguments. The same goes for preferred or grounded semantics, since they add a constraint of maximality/minimality to the selection of arguments. One could even argue that it would not be a relevant answer for admissible extensions, because these extensions might defend arguments that are not part of them. Thus, justifying the presence of an argument in an admissible extension merely by the fact that it is defended by the other arguments might lead to the legitimate reply "Then why is this other argument not in the extension although it is defended as well?". As such, only invoking defence for justifying the selection of an argument only seems relevant in the case of complete semantics, where indeed, defence equates to acceptance.

The third point of view is related to the size of our explanations. It is worth noting that sometimes our explanations consist of the entire original AF (see for instance Fig. 25). The scope of our explanations is fairly restrained in general (in the worst case, we keep arguments that are in a "distance" of 3 from the arguments[25] of the set for which we compute an explanation in the case of completeness). However, even with such a limited depth of search, if the set from which we begin the search contains arguments that are sparse and span all across the AF, one can easily see that the search will tend to cover the entire AF. Thus, in these situations, the explanations will indeed tend to be the entire AF. However, we have also seen through the examples of this paper that when considering less large sets, or sets that do not span across all the AF, the explanations tend to be more local and to be restrained to precise areas of the AF.

Another point of view is given by the way we cover the targeted domain. Our explanations concern all the classical **semantics** defined by Dung, except the grounded and the preferred ones. This is because our definitions of explanations are graphical and rely on the *modular* aspect of semantics. In particular, the grounded and preferred semantics include an aspect of *minimality* and *maximality* respectively. Minimality (resp. maximality) may be shown by answering the question on why the set reduced (resp. augmented) of a non empty subset of arguments is not

[25] So the elements that belong to $R^{-1}(R^2(S))$ for a given set S of arguments.

an extension under the considered semantics, and this, for any such subset. These subsets may be numerous, and the explanation that would hence be given may not be that intelligible. Such a solution is not that satisfactory in this sense. Another solution would consist to let the user ask why a given set they thought would be a minimal (resp. maximal) extension, is not. This would not be a direct explanation of why a set is minimal or maximal, but it may be helpful to a user who would have had in mind a different set: in the case of minimality (resp. maximality), a set which would be a subset (resp. superset) of an actual minimal (resp. maximal) one would not satisfy the properties of the semantics. However, finding a direct graphical explanation of minimality and maximality (and then, of the grounded and preferred semantics) is a challenge for future work.

The next point of view is related to the completeness of our explanations. We have proven that our explanations are sound but sometimes they could be more complete. This what was already suggested in Sections 4.1 and 6. Let us focus on situations like the one presented on Fig. 26. In this case, we supposed that the system delivered $\{a, d, h\}$ as an admissible extension and that the user asked the question "Why is h accepted ?". Following our methodology expressed in our assumption (H5), the system thus attempts to show how it cannot be the case that h is not accepted by showing what would happen if h is indeed not accepted. In this case, this amounts to showing the explanation for admissibility on $\{a, d\}$ (Fig. 26) and looking for a problematic situation like an internal conflict or a defenceless argument. However, such a situation does not occur since $\{a, d\}$ is also an admissible extension in Fig. 1. In consequence, our goal to show that h is necessary in the result presented by the system has failed. In order to solve this kind of problem, two possibilities could be taken into account:

1. We could relax our assumption (H5) and define an additional answer showing that the presence of h is not a problem (this additional answer completing the answer we already built based on the absence of h).
2. We could consider that the original question of the user becomes "If h being accepted is not necessary, then why is h accepted?". This question could naturally be interpreted as a suggestion from the user to the system to change its result. In other terms, the user's question could be interpreted as offering feedback to the system. So a possible improvement of our system could be the adaptation of its results to the preferences of the user by a sequence of such questions. The same can be said about the addition of arguments in the result instead of their removal.

Of course, this focused example on a specific question in a specific context can be extrapolated to a more general case, in which the computation of an answer results in an explanation for an eligible result.

To finish with, we present here some works that have already been done in trying to identify desirable properties for explanations and then we show how our approach could be positioned with respect to these properties. However, since there is no consensus on a clear notion of "explanation", these properties are as many as there are fields of research in AI.

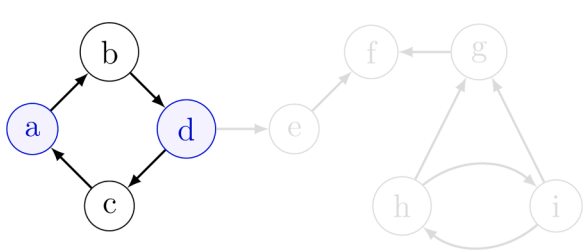


Fig. 26. Explanation G_{Adm} on why h is accepted in the admissible extension $\{a, d, h\}$. The removal of h does not lead to an internal conflict or a defenceless argument. Thus, h is not necessary in order to obtain the admissibility.

In the case of machine learning for instance, Ross et al. (2017) include their notion of local explanation directly into the function their model is trying to optimize. This way they constrain their model to provide only the "best" explanations. This notion of "best" explanation is unclear in our case because we do not identify a range of possible explanations and we lack other approaches on the same problems to make comparisons.

Riedl (2019) advocates to design human-centered AI. That is, AI systems that include a theory of mind of their users based on commonsense knowledge in order to better understand them, but also that can produce "rationales" (plausible a posteriori explanations) in order to help humans to better understand them. We share the same objective identified in Riedl (2019): define explanations that can be understood and are usable by non-expert users. For now however, this is not the case. Our system still relies on Hypothesis (H3) which assumes a degree of expertise in the domain. In addition, it does not include a theory of mind of the users, nor do we measure the plausibility of our explanation.

Rudin (2019) advocates to use interpretable systems instead of trying to explain those that are not. Interpretability is considered as a domain-specific property which results in constraints of form in accordance with some structural knowledge in the models that enjoy this property. Recalling the description of interpretability from Rudin (2019), it is arguable that Abstract Argumentation, which relies on a representation of knowledge as a graph, is an interpretable method. So, one might wonder if explanations might even be needed. Consider this: this understanding of what interpretability is at a global level only implies that interpretability allows introspection (of the model). Yet, this introspection might (and often is) complex and time-consuming. So the role of explanations in this case could become facilitating this introspection process. This is indeed what we do in our method by selecting only the relevant part in the decision-making process.

Amgoud and Ben-Naim (2022) study the explainability of classifiers through a certain number of axioms. These axioms represent concepts such as yielding explanations that are non empty, or that do not rely on empty reasons. They show that some of these axioms are incompatible and give rise to different kinds of explainers. Even though the axioms studied in Amgoud and Ben-Naim (2022) concern classifiers, some of the concepts represented can be extrapolated to other domains, and in particular to our system. For instance, their axiom called *success* states that an explanation should never be empty. Our system partially respects this axiom since the emptyness of our explanations is related to very particular situations (see Proposition 5). Their axiom called *explainability* states that an explanation should not rely on empty reasons. Although this might not be obvious to transfer to our system, it seems reasonable to state that our explanations only rely on empty reasons when they are empty themselves. So we could consider that our system partially respects this axiom.

11. Conclusion and Future Work

To conclude, we have provided explanations for Abstract Argumentation as answers to questions. The questions we consider are related to why a given set of arguments is (or not) an extension under some semantics, and to why a given argument or set of arguments is (or not) part of an extension of some semantics. In the former case, we stopped at non-contrastive questions, while we studied a particular case of contrast in the latter case, namely the contrast with another argument or set of arguments.

The answers we provide for these questions (i.e. explanations) are defined as subgraphs. The choice of subgraphs as explanations naturally yields visual explanations which are more easily understandable, and which can always be used to potentially generate other forms of explanations. The answers to questions related to why a given set of arguments is (or not) an extension under some semantics make use of the

modularity of Abstract Argumentation semantics. That is, we take advantage of how semantics are composed in one another, and of the different principles that govern them in order to build our explanations. They subsequently take the form of several subgraphs illustrating each principle that composes the semantics at hand. These subgraphs, each of them associated with a specific checking procedure, can then be shown sequentially or aggregated together in one explanation.

The answers to questions related to why a given argument or set of arguments is (or not) part of an extension of some semantics are always contrastive answers. More precisely, for these questions, we consider the question's context as the fact, and the question as providing the necessary elements to deduce the foil. We thus proceed to compute said foil as our answers using the content of the questions as well as our previous kind of answers. In the case of contrastive questions, we subsequently naturally interpret the contrast in the question as additional information to compute the foil.

We also discuss several aspects of our explanations, including but not limited to their accordance with Grice's maxims of conversation and some particular cases of answers.

Our present work can be extended in many ways. Our explanations are defined in a question-answer setting. However, for now, they are only presented as answers to fixed questions based on the general understanding of the question's meaning. We would like in future works to emphasize this question-answer link by making the explanation (so the computed subgraph) *result* from the question that is asked. This would require a way to automatically model and generate questions, and then find a way to compute the desired subgraph from the elements present in the question.[26]

Ideally, this process of generating questions and making the explanation result from these questions would allow to consider a large array of questions. The point is to define this process in such a way that computing answers from similar questions do not require much changes. This would allow for navigating more easily in the range of possible questions and define answers to questions that resemble some questions that have already been answered.

For now we only considered the case of conflict-free, admissible, complete and stable semantics, we also intend to study answers to questions related to grounded and preferred semantics. Finally, a more complete formal study of the properties of our explanations, as well as an empirical evaluation of their quality are to be done, following the ideas proposed in Amgoud and Ben-Naim (2022); Riedl (2019); Rudin (2019).

References

Amgoud, L., & Ben-Naim, J. (2022). Axiomatic foundations of explainability. *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI 2022)*. Austria: Vienne. https://hal.laas.fr/hal-03702681.

Baumann, R., & Ulbricht, M. (2021). Choices and their consequences - explaining acceptable sets in abstract argumentation frameworks. In M. Bienvenu, G. Lakemeyer, & E. Erdem (Eds.), *Proceedings of the 18th international conference on principles of knowledge representation and reasoning KR* (pp. 110–119). Online event: IJCAI Organization. https://doi.org/10.24963/kr.2021/11.

Besnard, P., Doutre, S., Duchatelle, T., & Lagasquie-Schiex, M. C. (2022). Question-based explainability in abstract argumentation. Research Report,IRIT/RR–2022–01–FR, IRIT : Institut de Recherche en Informatique de Toulouse, France.

Bondy, J. A., & Murty, U. S. R. (2008). Graph Theory. *Graduate Texts in Mathematics*. Springer. https://doi.org/10.1007/978-1-84628-970-5

Borg, A., & Bex, F. (2020a). Explaining arguments at the dutch national police. *AI approaches to the complexity of legal systems XI-XII* (pp. 183–197). Springer.

Borg, A., & Bex, F. (2020b). Necessary and sufficient explanations in abstract argumentation. *Computing Research Repository (CoRR)*. 2011.02414.

Borg, A., & Bex, F. (2021a). A basic framework for explanations in argumentation. *IEEE Intelligent Systems, 36*, 25–35. https://doi.org/10.1109/mis.2021.3053102

Borg, A., & Bex, F. (2021b). Contrastive explanations for argumentation-based conclusions. Computing Research Repository (CoRR), arXiv preprint: 2107.03265.

Borg, A., & Bex, F. (2021c). Necessary and sufficient explanations for argumentation-based conclusions. In J. Vejnarová, & N. Wilson (Eds.), *Proceedings of the 16th european conference on symbolic and quantitative approaches to reasoning with uncertainty, ECSQARU* (pp. 45–58). Prague, Czech Republic: Springer. https://doi.org/10.1007/978-3-030-86772-0_4.

Cyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative XAI: A survey. In Z. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence IJCAI* (pp. 4392–4399). Online Event / Montreal, Canada: IJCAI Organization. https://doi.org/10.24963/ijcai.2021/600.

Doutre, S., Duchatelle, T., & Lagasquie-Schiex, M. C. (2022). Explainability of extension-based semantics. https://hal-univ-tlse3.archives-ouvertes.fr/hal-03657060. Research Report,IRIT/RR–2022–05–FR, IRIT, France.

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*, 321–357. https://doi.org/10.1016/0004-3702(94)00041-x

Fan, X., & Toni, F. (2015a). On computing explanations in argumentation. In B. Bonet, & S. Koenig (Eds.), *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 1496–1502). Austin, Texas, USA: AAAI Press.

Fan, X., & Toni, F. (2015b). On explanations for non-acceptable arguments. In E. Black, S. Modgil, & N. Oren (Eds.), *Theory and applications of formal argumentation, TAFA - third international workshop* (pp. 112–127). Buenos Aires, Argentina: Springer. https://doi.org/10.1007/978-3-319-28460-6_7.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and Semantics: vol. 3. Speech acts* (pp. 41–58). Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004368811_003.

Liao, B., & van der Torre, L. (2020). Explanation semantics for abstract argumentation. In H. Prakken, S. Bistarelli, F. Santini, & C. Taticchi (Eds.), *Computational models of argument - proceedings of COMMA 2020* (pp. 271–282). Perugia, Italy: IOS Press. https://doi.org/10.3233/FAIA200511.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Niskanen, A., & Järvisalo, M. (2020). Smallest explanations and diagnoses of rejection in abstract argumentation. In D. Calvanese, E. Erdem, & M. Thielscher (Eds.), *Proceedings of the 17th international conference on principles of knowledge representation and reasoning KR* (pp. 667–671). Rhodes, Greece: IJCAI Organization. https://doi.org/10.24963/kr.2020/67.

Racharak, T., & Tojo, S. (2021). On explanation of propositional logic-based argumentation system. In A. P. Rocha, L. Steels, & H. J. van den Herik (Eds.), *Proceedings of the 13th international conference on agents and artificial intelligence ICAART* (pp. 323–332). Online Streaming: SCITEPRESS. https://doi.org/10.5220/0010318103230332.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *CoRR, abs/1901.11184*.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. In C. Sierra (Ed.), *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, melbourne, australia, august 19-25, 2017* (pp. 2662–2670). ijcai.org.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell., 1*, 206–215.

Saribatur, Z. G., Wallner, J. P., & Woltran, S. (2020). Explaining non-acceptability in abstract argumentation. In G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, & J. Lang (Eds.), *Proceedings of the 24th european conference on artificial intelligence ECAI* (pp. 881–888). Santiago de Compostela, Spain: IOS Press. https://doi.org/10.3233/FAIA200179.

Ulbricht, M., & Baumann, R. (2019). If nothing is accepted - repairing argumentation frameworks. *Journal of Artificial Intelligence Reseach, 66*, 1099–1145. https://doi.org/10.1613/jair.1.11791

Ulbricht, M., & Wallner, J. P. (2021). Strong explanations in abstract argumentation. *Proceedings of the thirty-fifth AAAI conference on artificial intelligence* (pp. 6496–6504). Online event: AAAI Press.

Vesic, S., Yun, B., & Teovanovic, P. (2022). Graphical representation enhances human compliance with principles for graded argumentation semantics. In P. Faliszewski, V. Mascardi, C. Pelachaud, & M. E. Taylor (Eds.), *21st international conference on autonomous agents and multiagent systems, AAMAS 2022* (pp. 1319–1327). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[26] A first attempt has been done, considering a formal grammar for defining our questions (see Besnard et al. (2022)) but it requires more work to be exploitable automatically.