Explainability of Extension-Based Semantics

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

Explainability of Extension-Based Semantics

Sylvie Doutre
Théo Duchatelle
Marie-Christine Lagasquie-Schiex

IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3
118 route de Narbonne, 31062 Toulouse, France
{doutre, theo.duchatelle, lagasq}@irit.fr

Tech. Report
IRIT/RR- -2022- -05- -FR

April 2022

Abstract

This paper defines visual explanations to the Verification Problem in argumentation, that is, of why a set of arguments is or is not acceptable under a given semantics. These explanations rely upon the modularity of the acceptability semantics, and they take the form of subgraphs of the original argumentation graph. Graph properties that these subgraphs satisfy depending on whether or not the set is acceptable, are established. Properties of the proposed explanations are addressed, and the potential of the modularity of the approach is highlighted.

Note that this research report is the complete version of a paper submitted to a conference. In this complete version, the reader can find the proofs of the results given in the submitted paper.

Contents

1 Introduction

In the context of Explainable Artificial Intelligence (XAI), Abstract Argumentation is increasingly studied as a formal tool to provide explanations of decisions made using an Artificial Intelligence system. The recent survey by [1] 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.

Regarding the argumentation process, 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. The most common approach consists in identifying some set(s) of arguments which act as explanation(s) (as in [2, 3, 4, 5, 6, 7]). 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 fully helpful. Furthermore, beyond the question of the global acceptability of an argument or of a set of arguments, many other questions on the outcomes of argumentation or on the process of argumentation itself can be asked.

A grammar which defines sets of such questions can be found in [8]. Answers to these questions are presented there in the form of relevant subgraphs, as in [9, 10, 11]. As such, this approach is a visual one, graphs having been shown to be helpful for humans to comply with argumentation reasoning principles [12]. This graph-based approach not only highlights arguments, but also subsets of attacks.

This paper aims at defining explanations to the *Verification Problem* Ver_σ defined as follows: given an Argumentation Framework $\mathscr{A}$, a set of arguments S and an extension-based semantics σ, "Is S acceptable under σ in $\mathscr{A}$?". The answer to this problem is "yes" or "no". We aim at providing a visual explanation of why the answer is so. For this reason, we define the *Explanation Verification Problem* XVer_σ: "Why is S (not) acceptable under σ in $\mathscr{A}$?". Subgraphs answering this question will be formally defined for some acceptability semantics in Dung's framework [13], and properties they satisfy will be established, depending on whether the answer to the corresponding verification problem is "yes" or "no". This methodology follows the line of [14] in that an explanation for a set S satisfying a semantic principle σ is a (set of) subgraph(s) G of $\mathscr{A}$ such that G satisfies a given graph property C. Moreover, the semantics that we consider are based on a modular definition, which allows the explanations to be decomposed.

The paper is organised as follows: Sec. 2 recalls background notions relative to abstract argumentation and graph theory. Sec. 3 defines explanations to the acceptability of a set of arguments under atomic semantic principles, and it investigates properties such explanations should satisfy. Explanations for semantics based on a composition of atomic principles are presented in Sec. 4. Related works are discussed in Sec. 5, and Sec. 6 concludes, presents potentials of the modular approach as avenues for future works.

2 Preliminary Notions

In this section we give some basic definitions on which our work is built. These concern Dung's Abstraction Framework formalism as well as some graph-theoretic notions.

2.1 Abstract Argumentation

We begin by recalling background notions on Abstract Argumentation.

Def. 1 (Argumentation Framework (AF) [13])**.** A *Dung's Argumentation Framework (AF)* is an ordered pair $\mathscr{A} = (A, R)$ such that $R \subseteq A \times A$.

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

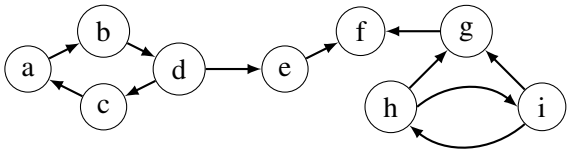

Figure 1: Example of an AF from [15]

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

Def. 2. Given $\mathscr{A} = (A,R)$, $a \in A$ is *acceptable* wrt $S \subseteq A$ iff for all $b \in A$, if *bRa* then *cRb* for some $c \in S$. The *characteristic function* of $\mathscr{A}$ is $F_{\mathscr{A}} : 2^A \to 2^A$ such that $F_{\mathscr{A}}(S) = \{a \in A \mid a \text{ is acceptable wrt } S\}$ for any $S \subseteq A$.[1] Let $S \subseteq A$. S satisfies the principle:

CF	(*conflict-freeness*)	iff there are no a and b in S such that a attacks b
Def	(*defence*)	iff for any $a \in S$, a is acceptable wrt S
Reins	(*reinstatement*)	iff $\forall a \in A$, if a is acceptable wrt S then $a \in S$
CA	(*complement-attack*)	iff each argument in $A \setminus S$ is attacked by S

Some semantics originally defined in [13] can be characterized as follows:

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

Adm	(*admissible*)	iff it satisfies each principle of the set {*CF, Def*}
Co	(*complete*)	iff it satisfies each principle of the set {*CF, Def, Reins*}
Sta	(*stable*)	iff it satisfies each principle of the set {*CF, CA*}

2.2 Graph Theory

This section recalls some graph-theoretic notions. These concern particular subgraphs and nodes, as well as the successor and predecessor functions.

Def. 4 (Subgraph, Induced subgraph, Partial subgraph)**.** Let $G = (V,E)$ and $G' = (V',E')$ be two graphs.

- G' is a *subgraph* of G iff $V' \subseteq V$ and $E' \subseteq E$.
- 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 *partial subgraph*[2] *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 (see Fig. 2). In a partial 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 (see Fig. 3). Note that the computation of an induced or a partial subgraph is obviously polynomial in the size of the original graph.

Induced and partial subgraphs are examples of ways to compute a graph from another single graph. We continue with a particular kind of graphs, bipartite graphs.

[1] Note that any unattacked argument of $\mathscr{A}$ belongs to $F_{\mathscr{A}}(S)$ whatever is S.

[2] The name *spanning subgraph* is also used in the literature.

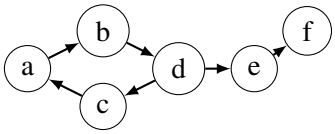

Figure 2: Induced subgraph of Fig. 1 by $\{a,b,c,d,e,f\}$

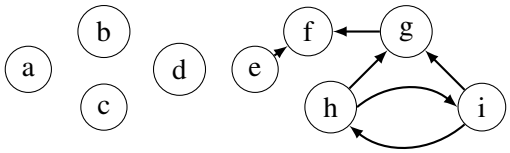

Figure 3: Partial subgraph of Fig. 1 by $\{(g,f),(e,f),(h,g),(i,g),(h,i),(i,h)\}$

Def. 5 (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).

Bipartite graphs are those 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.

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

Def. 6 (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, we recall some notions on vertices having a particular status in a graph.

Def. 7 (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.

Thus, *sources* (resp. *sinks*) are vertices that may only be origins (resp. endpoints) of arcs. *Isolated* vertices are those that are connected to no other vertices.

3 Explanations for Principles of Semantics

Given $\mathscr{A}=(A,R)$ an AF, $S\subseteq A$ a set of arguments, and $\sigma\in\{CF,Def,Reins,CA,Adm,Co,Sta\}$ an acceptability semantics or a principle used in such a semantics, the question we aim at answering is the following one:

Q_σ: "Why is S acceptable under σ in $\mathscr{A}$?"

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

Given the definitions presented in Sec. 2, an explanation of why S is acceptable under σ should show how S satisfies each of the principles which compose σ.[4]

This section will define explanations to the satisfaction by S of each of the four atomic acceptability principles, that is, we will answer Q_σ for $\sigma \in \{CF, Def, Reins, CA\}$. These explanations will take the form of subgraphs relative to S, and properties of these subgraphs which will depend on whether the principle is satisfied or not by S will be characterised. Properties of these explanations will also be investigated. Sec. 4 will use these explanations for the atomic acceptability principles, to explain the satisfaction by S of acceptability semantics which combine them, answering then Q_σ for $\sigma \in \{Adm, Co, Sta\}$.

Notice that Q_σ is the positive part of XVer_σ introduced in Sec. 1. This focus aims at enhancing the readability of the results which will follow. Whether the question is positive or negative, the explanation subgraphs will be the same, but their properties and interpretation will differ depending on whether S is acceptable or not.

On the form, in the explanation subgraphs, the nodes the question is about will be shaded (in blue). Depending on the principle to be explained, the nodes or arcs which may cause the principle not to be satisfied will appear with a bold line (in red).

3.1 Characterisations

***Conflict-freeness principle** (CF)* Fig. 4 illustrates why $\{d, f, h\}$ constitutes a conflict-free set in the AF of Fig. 1: if we focus only on those arguments, we see that they are linked with no arc. On the contrary, one can observe on Fig. 5 that $\{a, d, e\}$ is not conflict-free as the focus on those arguments highlights the presence of a conflict.

Figure 4: Why is $\{d, f, h\}$ conflict-free?

Figure 5: Why is $\{a, d, e\}$ not conflict-free?

Thus, an explanation for the property of conflict-freeness should show the conflicts inside a set, with the absence of said conflicts considered as a testimony for the property to hold. So the explanation for a given set is only the subgraph induced by this set:

Def. 8 (Explanation for conflict-freeness)**.** Let $\mathscr{A} = (A, R)$ and $S \subseteq A$. The subgraph $G_{CF}(S) = \mathscr{A}[S]_V$ is an answer to Q_{CF}.

The next theorem shows that this explanation enjoys the property of containing conflicts iff the queried set is not conflict-free.

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

In order to provide some properties about the minimality / maximality of our explanation, we need to define what constitutes a minimal or maximal explanation for conflict-freeness. The notion of minimal or maximal explanation revolves around what reason suffices to decide whether a set is conflict-free or not. So a minimal explanation is an explanation containing one such reason, while a maximal explanation contains them all.

[4] In order to explain why S is *not* acceptable under σ, at least one of the principles of σ which is not satisfied by S has to be shown.

Def. 9. Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $G_{CF}(S) = (A',R')$ is minimal iff $|R'| \leq 1$. It is maximal iff $\{(a,b) \in R \mid a,b \in S\} \subseteq R'$.

Sec. 3.2 will present results regarding the minimality, maximality, existence and unicity of our explanations.

***Defence principle** (Def)* In an AF, defence (by a set) is captured by the notion of acceptability (w.r.t. that set) which states that all attackers of an argument must in turn be attacked by an argument of the defending set. Fig. 6 illustrates why $\{b,c\}$ defends all its arguments in the AF of Fig. 1: all arguments that attack b or c, namely a and d, are also attacked by either b or c. On the other hand, Fig. 7 illustrates why $\{b,e\}$ does not defend all its arguments: a that attacks b is immediately spotted as receiving no attack from either b or e.

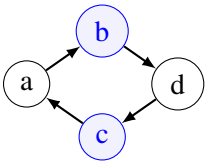

Figure 6: Why $\{b,c\}$ defends all its arguments

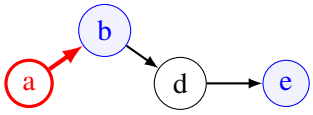

Figure 7: Why $\{b,e\}$ does not defend all its arguments

Thus, an explanation for defence should have the property of focusing on a set and its attackers, and on the attacks between these two groups of arguments. The fact that all the attackers receive an attack back is considered a testimony for the defence of the queried set. The next definition formalises this notion of explanation for defence.

Def. 10 (Explanation for defence)**.** Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. The subgraph $G_{Def}(S)$ is an answer to Q_{Def}:

$$G_{Def}(S) = (\mathscr{A}[S \cup R^{-1}(S)]_V)[\{(a,b) \in R \mid a \in R^{-1}(S) \text{ and } b \in S, \text{or } a \in S \text{ and } b \in R^{-1}(S)\}]_E$$

As such, the explanation for defence on a set S is simply the subgraph $G_{Def}(S)$ induced by S and its attackers, and in which we only keep the attacks between S and its attackers. In the following, we prove results showing that our notion of explanation for defence respects the properties we highlighted in the previous examples.[5]

Prop. 1. *Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. If S is conflict-free, $G_{Def}(S)$ is a bipartite graph and S can always be one of its parts.*

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

Here the notion of minimal or maximal explanation is defined as follows. Since a set is defended provided that all its attackers are attacked back, a minimal explanation is such that all attackers are the endpoint of at most one arc. A maximal explanation contains all possible arcs the set can use to defend itself.

Def. 11. Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $G_{Def}(S) = (A',R')$ is minimal iff for all $x \in R^{-1}(S)$, $|R'^{-}(x)| \leq 1$. It is maximal iff $\{(a,b) \in R \mid a \in S, b \in R^{-1}(S)\} \subseteq R'$.

[5]The results require the queried set to be conflict-free: this is not a strong limitation, defence being checked in combination with conflict-freeness in many semantics.

Reinstatement principle *(Reins)* This principle can be viewed as two sub-principles: $Reins_1$ meaning that all the unattacked arguments must be in the set and $Reins_2$ meaning that the set contains all the attacked arguments it defends. In the AF of Fig. 1, an empty subgraph can illustrate $Reins_1$ for $S = \{h\}$ or $S = \{b,c\}$ (since there is no unattacked argument in this AF, whereas Fig. 8 and 9 illustrate $Reins_2$ on the same two sets.

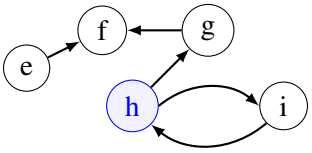

Figure 8: Why $\{h\}$ contains all the arguments it defends: h defends itself against i and although it defends f against g, h is ineffective against e. So, f is not defended by h and thus not part of the set

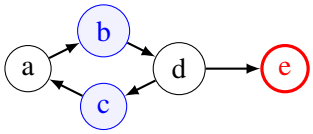

Figure 9: Why $\{b,c\}$ does not contain all the arguments it defends. Indeed, b defends e against d, its only attacker, but e is not part of the set

Through these examples we see that two subgraphs are necessary: the first one with the unattacked arguments of the graph; the second one with all the arguments that are in a range of 2 from the set following the relation. In the latter case, these arguments must (resp. must not) be in the set if they are (resp. are not) defended. [6] Thus the definition for an explanation of reinstatement:

Def. 12 (Explanation for reinstatement)**.** Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. The set of subgraphs $\{G_{Reins1}(S), G_{Reins2}(S)\}$ is an answer to Q_{Reins}, with $G_{Reins1}(S) = \mathscr{A}[\{a \in A | R^-(a) = \varnothing\}]_V$ and $G_{Reins2}(S) = (\mathscr{A}[S \cup R^2(S) \cup R^{-1}(R^2(S))]_V)$ $[\{(a,b) \in R \mid a \in R^{-1}(R^2(S)) \text{ and } b \in R^2(S), \text{ or } a \in S \text{ and } b \in R^{-1}(R^2(S))\}]_E$.

$G_{Reins1}(S)$ is induced by the unattacked arguments; $G_{Reins2}(S)$ is induced by S, the arguments defended by S and the attackers of those arguments, and in which the only attacks are from the attackers to the arguments defended by S and from S to the attackers.

In the following, we prove results showing that our notion of explanation for reinstatement respects the properties highlighted in the previous examples.

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

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

From Theo. 3 and 4 follows the next Corollary, which shows an equivalence result:

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

For $Reins_1$, a minimal G_{Reins1} must contain at most the unattacked arguments of S plus one other argument (not in S); for the maximality it must contain all the unattacked arguments. For $Reins_2$, the concept of

[6] If these two graphs were merged, it would not be possible to differentiate unattacked arguments in the merger of G_{Reins1} and G_{Reins2} which should be in S from those which should not be; hence the two graphs.

minimality/maximality of explanations for reinstatement relies on the same concept as for explanations for defence. So, we require that in G_{Reins2}, each attacker of $R^2(S)$ is the endpoint of at most one arc. For the maximality, we require that it includes all such arcs.

Def. 13. Let $\mathscr{A} = (A,R)$ and $S \subseteq A$.
$G_{Reins1}(S) = (A',R')$ is minimal iff $|A' \setminus S| \leq 1$. It is maximal iff $\{x | R^-(x) = \varnothing\} \subseteq A'$.
$G_{Reins2}(S) = (A',R')$ is minimal iff for all $x \in R^{-1}(R^2(S))$, $|R'^-(x)| \leq 1$. It is maximal iff $\{(a,b) \in R \mid a \in S, b \in R^{-1}(R^2(S))\} \subseteq R'$.

Complement Attack (CA) Fig. 10 illustrates why $\{a,d,f,h\}$ attacks its complement in the AF of Fig. 1: every argument that is not in $\{a,d,f,h\}$ receives an attack from either a, d, f or h. On the contrary, Fig. 11 illustrates why $\{b,c,h\}$ does not attack its complement: neither e or f that are not in $\{b,c,h\}$ are attacked by b, c or h.

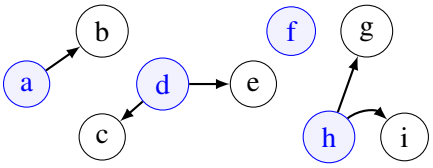

Figure 10: Why $\{a,d,f,h\}$ attacks its complement

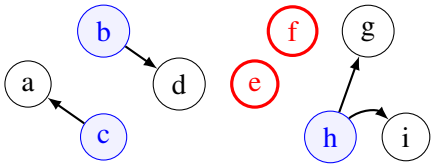

Figure 11: Why $\{b,c,h\}$ does not attack its complement

Hence, we see that an explanation for the attack of a set's complement should show the arguments attacked by that set among all the other arguments in the AF. If these two groups happen to coincide, then we can conclude that the set indeed attacks its complement. Thus the definition for an explanation of complement attack is simply the partial subgraph by the attacks from the set to arguments that are not in it:

Def. 14 (Explanation for complement attack)**.** Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. The subgraph $G_{CA}(S)$ is an answer to Q_{CA} with $G_{CA}(S) = \mathscr{A}[\{(a,b) \in R \mid a \in S \text{ and } b \notin S\}]_E$.

As such, the explanation for complement attack on a set is simply the partial subgraph by the attacks from the set to arguments that are not in it.

Next, we prove results which show that our notion of explanation respects the properties mentioned in the introductory examples.

Prop. 2. *Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $G_{CA}(S)$ is a bipartite graph, S can always be one of its parts and all vertices in S are sources in it.*[7]

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

Since a set S attacks its complement only when all other arguments are attacked by some argument of S, a minimal explanation is such that any argument not in S is the endpoint of at most one arc. A maximal explanation contains all possible arcs S can use to attack other arguments.

Def. 15. Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $G_{CA}(S) = (A',R')$ is minimal iff for all $x \notin S$, $|R'^-(x)| \leq 1$. It is maximal iff $\{(a,b) \in R \mid a \in S, b \notin S\} \subseteq R'$.

[7] 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.

Synthesis Table 1 sums up all the previous characterisations of our explanations with regard to the user's questions giving the answer we produce and their interpretation.[8]

σ	Question Q_σ: "Why is S acceptable under σ in $\mathscr{A}$?"	
	Answer	Interpretation of the Answer: $G_\sigma = (A', R')$
CF	G_{CF} (Def.8)	YES iff C_{CF} is satisfied by S (Th.1) with C_{CF}: "$R' = \varnothing$ in $G_{CF}(S)$"
Def	G_{Def} (Def.10)	when S is conflict-free, YES iff C_{Def} is satisfied by S (Th.2) with C_{Def}: "$\nexists b \in R^{-1}(S) \cap A'$ st $R'^{(}b) = \varnothing$ in $G_{Def}(S)$"
Reins	G_{Reins1} G_{Reins2} (Def.12)	YES if C_{Reins1} and C_{Reins2} are satisfied by S (Th.3) YES only if C_{Reins1} and C'_{Reins2} are satisfied by S (Th.4) when $R^2(S)$ is conflict-free, YES iff C_{Reins1} and C_{Reins2} are satisfied by S (Cor.1) with C_{Reins1}: "$A' \subseteq S$ in $G_{Reins1}(S)$" C_{Reins2}: "$\forall x \in R^2(S) \setminus S, \exists (b,x) \in R'$ st $R'^{-}(b) = \varnothing$ in $G_{Reins2}(S)$" C'_{Reins2}: "$\forall x \in R^2(S) \setminus S, \exists (b,x) \in R'$ st $R'^{-}(b) = \varnothing$ or $b \in R^2(S)$ in $G_{Reins2}(S)$"
CA	G_{CA} (Def.14)	YES iff C_{CA} is satisfied by S (Th.5) with C_{CA}: "$\forall x \in A' \setminus S, \exists (b,x)$ or $(x,b) \in R'$ in $G_{CA}(S)$"

Table 1: Synthesis of explanations for principles with their interpretation: "YES" means that no reason can be exhibited in our answer for proving that S is not acceptable under σ (and so our answer shows why S is acceptable under σ in $\mathscr{A}$)

To finish with, we would like to point out that, since all our explanations are defined using only induced subgraphs and partial subgraphs, they are efficiently computed.

3.2 Additional properties

Regarding uniqueness, our explanations for the five (sub-)principles described in the previous section (*CF*, *Def*, $Reins_1$, $Reins_2$, *CA*) are unique given a set of arguments:

Prop. 3. *Let $\mathscr{A} = (A,R)$, $S, S' \subseteq A$ and $\sigma \in \{CF, Def, Reins_1, Reins_2, CA\}$. If $S = S'$, then $G_\sigma(S) = G_\sigma(S')$.*

About the non-emptiness of explanations, the conditions are different depending on the concerned (sub-)principle: if for *CF*, *Def*, $Reins_2$ the explanations are empty only when the queried set is the empty set, it is when there exists no unattacked argument that the explanation for $Reins_1$ is empty and when the original graph is empty that the explanation for *CA* is empty.

Prop. 4. *Let $\mathscr{A} = (A,R)$, $S \subseteq A$. 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 $\mathscr{A} = (\varnothing, \varnothing)$.*

Regarding maximality, the explanations for the five (sub-)principles are maximal:

Prop. 5. *Let $\mathscr{A} = (A,R)$, $S \subseteq A$ and $\sigma \in \{CF, Def, Reins_1, Reins_2, CA\}$. $G_\sigma(S)$ is a maximal explanation for σ on S.*

Regarding minimality, a specific case for the explanation for the *CF* principle exists:

[8] For the negative version of these questions, *i.e.* "Why is S *not* acceptable under σ in $\mathscr{A}$?", the answer is obviously the same but with an opposite interpretation.

Prop. 6. *Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. If S is conflict-free, $G_{CF}(S)$ is a minimal explanation for conflict-freeness on S.*

For the four other (sub-)principles, our explanations may not be minimal and some simple counterexamples are enough for illustrating this point (see Fig. 12, 13 and 14).

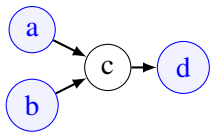

Figure 12: G_{Def} and G_{Reins2} not minimal: $\mathscr{A}$ and its G_{Def} and G_{Reins2} for $S = \{a,b,d\}$; either (a,c) or (b,c) is enough for a minimal explanation about the defence or the part 2 of reinstatement of S

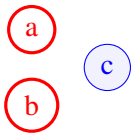

Figure 13: G_{Reins1} not minimal: $\mathscr{A}$ and its G_{Reins1} for $S = \{c\}$; either a or b is enough for a minimal explanation about the part 1 of reinstatement of S

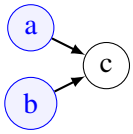

Figure 14: G_{CA} not minimal: $\mathscr{A}$ and its G_{CA} for $S = \{a,b\}$; either (a,c) or (b,c) is enough for a minimal explanation about S attacking its complement

4 Explanations for Semantics

Semantics have been defined as sets of principles they should all satisfy. Since we gave explanations concerning these principles (and sub-principles), explanations concerning semantics naturally arise as sets of explanations on (sub-)principles.

Def. 16 (Explanation for semantics)**.** Let $\mathscr{A} = (A,R)$ and $S \subseteq A$.[9]

- $G_{Adm}(S)$ is an answer to Q_{Adm} with $G_{Adm}(S) = \{G_{CF}(S), G_{Def}(S)\}$
- $G_{Co}(S)$ is an answer to Q_{Co} with $G_{Co}(S) = \{G_{CF}(S), G_{Def}(S), G_{Reins1}(S), G_{Reins2}(S)\}$
- $G_{Sta}(S)$ is an answer to Q_{Sta} with $G_{Sta}(S) = \{G_{CF}(S), G_{CA}(S)\}$

Since explanations for semantics are composed of explanations for (sub-)principles, properties of the latter naturally extend to the former, provided that these properties are universally enjoyed by all explanations for (sub-)principles. This is for instance the case concerning properties of uniqueness and maximality.

Prop. 7. *Let $\mathscr{A} = (A,R)$, $S,S' \subseteq A$ and $\sigma \in \{Adm, Co, Sta\}$. If $S = S'$, then $G_\sigma(S) = G_\sigma(S')$. Moreover, $G_\sigma(S)$ is a maximal explanation for σ on S.*

In addition, all explanations for (sub-)principles are efficiently computed, thus it is also the case for explanations for semantics.

In a similar fashion, properties that are not universally enjoyed by all explanations for (sub-)principles are also not enjoyed by explanations for semantics. So, considering the minimality of explanations, Fig. 12, 13 and 14 can be used for showing that $G_{Adm}(S)$, $G_{Co}(S)$ and $G_{Sta}(S)$ are not minimal. To finish with, consider the emptyness property:

Prop. 8. *Let $\mathscr{A} = (A,R)$, $S \subseteq A$. $G_{Adm}(S) = (\varnothing,\varnothing)$ iff $S = \varnothing$. $G_{Co}(S) = (\varnothing,\varnothing)$ iff $S = \varnothing$ and $\{x \in A | R^-(x) = \varnothing\} = \varnothing$. $G_{Sta}(S) = (\varnothing,\varnothing)$ iff $\mathscr{A} = (\varnothing,\varnothing)$.*

[9] Using sets allows to not have a fixed order in the sequence of subgraphs.

5 Related Works

A similar approach to ours in providing explanation in Abstract Argumentation is [9]. The authors aim to explain the credulous non acceptance of some argument using strongly rejecting subframeworks, that is, induced subgraphs of the original AF s.t. neither this subgraph nor its supergraphs credulously accept the queried argument. As such, strongly rejecting subframeworks capture the core argumentative reasons for why an argument is not credulously accepted under a certain semantics. We focus on a different problematic, the status of a set of arguments as a (non-)extension. In addition, while strongly rejecting subframeworks are exclusively induced subgraphs, we use both induced and partial subgraphs as we consider attacks to be as important as arguments in an explanation. [10] also studies subgraphs as explanations for the credulous non acceptance of some argument for a given semantics using both induced and partial subgraphs. Nevertheless these subgraphs are not combined like we do in our work.

A specific kind of graph that is used in explaining argumentative results is *defence trees*: 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. While not being subgraphs technically speaking, one can easily retrieve the subgraph represented by a defence tree using the original AF. Some works, like [11], use defence trees as explanations for argumentative results. The authors of [11] 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 [2], use them to compute their notion of explanations. While not being subgraphs technically speaking, one can easily retrieve the subgraph represented by a defence tree using the original AF. Thus, a defence tree can be seen as a visual explanation (in the sense of a proof) of the (non-)acceptability of some argument w.r.t. a given set.

Some works explore the idea of explanations as elements to remove from an AF to modify a given result (e.g. an argument being credulously accepted or not). In [16], 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 to make the queried argument credulously accepted under admissibility in the resulting subgraph. This is also the method used in [17], which called such sets "*diagnosis*". [10] notes that, for explanations, diagnoses can be seen as the dual of induced and partial subgraphs. Indeed, each diagnosis infers an induced or partial subgraph, and conversely, each induced or partial subgraph is computed using (the complement of) a diagnosis.

Finally, the most widely used method to define explanations is as sets of arguments. Among this category, several works like [2, 3, 4, 15, 18, 5] are interested in explanations for the credulous/skeptical (non-)acceptance of some argument(s). In [2], the authors define an explanation semantics, called related admissibility, which provides all the reasons why an argument belongs to an admissible set. In [3, 4], the authors propose a basic framework to compute explanations as sets of arguments for the credulous/skeptical (non-)acceptance of an argument, and extend it in subsequent works [15, 18]. [5] proposes strong explanations for credulous acceptance of a set of arguments under a given semantics, that 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 contains the set to explain. Other works that use sets of arguments as explanations deal with the same question we focus on in this paper. In [6], the authors base their approach on the observation that each Strongly Connected Component (SCC) of an AF can be seen as making a choice for accepting conflict-free sets of arguments. From these choices results the rest of the accepted arguments, hence such choices can be seen as explanations for the rest of the arguments in a set. Using this idea and extending beyond choices of conflict-free sets in SCC, they explore explanation semantics and their properties. Similarly, in [7], the authors observe that complete and admissible semantics are computed firstly by computing 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.

6 Conclusion and Future Work

This paper provides explanations to why a set of arguments is or is not acceptable under a given semantics σ in Dung's framework (XVer_σ problem). Semantic *principles* and *semantics* which combine them are addressed. Graph properties satisfied by the *subgraphs* or sets of subgraphs defined as explanations for the problem, depending on the actual acceptability of the considered set, have been proven. Properties of the explanations themselves such as uniqueness, non-emptiness, maximality and minimality have been investigated. The *modularity* of the considered semantics make the explanations modular as well. The considered problem, the solution in term of induced and partial subgraphs, its properties and its modularity, as far as we know, make the proposed approach original.

These explanations should make the Verification Problem in Argumentation more intelligible to humans (be they ordinary users or specialists of argumentation who may have to use or develop argumentation solvers for instance). However, experiments with human users should be conducted, as in much of the XAI literature as [1] underlines, to check to which extend the proposed explanations actually make best sense.

Back to the problem, semantics which involve a *minimality* or a *maximality* principle (e.g. the grounded or the preferred semantics [13]) are more challenging to graphically explain. However, in the case where it is a set S output from a solver that a user may want to be explained, the modularity of our approach allows an explanation to be provided. Actually, when presented with a maximal (resp. minimal) set S satisfying a set P of principles, if a user wants S to be explained, it may be because, as advocated in [8], they think that S is not maximal (resp. minimal) satisfying P, hence that a superset (resp. subset) S' of S should be. The user may then ask why S' is not acceptable under the considered semantics. By the property of maximality (resp. minimality), the answer will show the principles of P the property applies to, which are not satisfied by S'.

Beyond maximality and minimality, the explanation problem should be extended to additional principles and semantics. It may even be extended to frameworks which enrich Dung's one. For an overview of such semantics and frameworks, see [19, 20].

Moreover, a *contrastive* variant of the problem may be considered: explaining why a set S is acceptable under a semantics σ and not under a semantics σ'. The modularity of the approach may here again be helpful: given σ and σ' defined as sets of principles, the explanation may consist in showing how S behaves on the principles on which σ and σ' differ. As an example, if σ is the admissible semantics and σ' the complete semantics, the explanation will show that the reinstatement principle is not satisfied by S.

Such extensions of the proposed approach, on principles, semantics, enriched frameworks and contrastive questions, are avenues for future works.

References

[1] Čyras K, Rago A, Albini E, Baroni P, Toni F. Argumentative XAI: A Survey. In: Proc. of IJCAI; 2021. p. 4392-9.

[2] Fan X, Toni F. On Computing Explanations in Argumentation. In: Proc. of AAAI; 2015. p. 1496-502.

[3] Borg A, Bex F. Necessary and Sufficient Explanations in Abstract Argumentation. Computing Research Repository (CoRR). 2020;abs/2011.02414.

[4] Borg A, Bex F. Necessary and Sufficient Explanations for Argumentation-Based Conclusions. In: Proc. of ECSQARU. vol. 12897 of LNCS. Springer; 2021. p. 45-58.

[5] Ulbricht M, Wallner JP. Strong Explanations in Abstract Argumentation. In: Proc. of AAAI; 2021. p. 6496-504.

[6] Liao B, van der Torre L. Explanation Semantics for Abstract Argumentation. In: Proc. of COMMA. vol. 326. IOS Press; 2020. p. 271-82.

[7] Baumann R, Ulbricht M. Choices and their Consequences - Explaining Acceptable Sets in Abstract Argumentation Frameworks. In: Proc. of KR; 2021. p. 110-9.

[8] Besnard P, Doutre S, Duchatelle T, Lagasquie-Schiex MC. Question-Based Explainability in Abstract Argumentation. IRIT, France; 2022. IRIT/RR–2022–01–FR.

[9] Saribatur ZG, Wallner JP, Woltran S. Explaining Non-Acceptability in Abstract Argumentation. In: Proc. of ECAI. vol. 325; 2020. p. 881-8.

[10] Niskanen A, Järvisalo M. Smallest Explanations and Diagnoses of Rejection in Abstract Argumentation. In: Proc. of KR; 2020. p. 667-71.

[11] Racharak T, Tojo S. On Explanation of Propositional Logic-based Argumentation System. In: Proc. of ICAART. vol. 2; 2021. p. 323-32.

[12] Vesic S, Yun B, Teovanovic P. Graphical Representation Enhances Human Compliance with Principles for Graded Argumentation Semantics. In: Proc. of AAMAS; 2022. .

[13] Dung PM. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence. 1995;77(2):321-57.

[14] Cocarascu O, Čyras K, Rago A, Toni F. Explaining with argumentation frameworks mined from data. In: Proc. of DEXAHAI; 2018. .

[15] Borg A, Bex F. A Basic Framework for Explanations in Argumentation. IEEE Intelligent Systems. 2021;36(2):25-35.

[16] Fan X, Toni F. On Explanations for Non-Acceptable Arguments. In: Proc. of TAFA. vol. 9524 of LNCS; 2015. p. 112-27.

[17] Ulbricht M, Baumann R. If Nothing Is Accepted - Repairing Argumentation Frameworks. JAIR. 2019;66:1099-145.

[18] Borg A, Bex F. Contrastive Explanations for Argumentation-Based Conclusions. Computing Research Repository (CoRR). 2021;abs/2107.03265.

[19] Baroni P, Gabbay D, Giacomin M, Van der Torre L. Handbook of formal argumentation. College Publications; 2018.

[20] Gabbay D, Giacomin M, Simari G. Handbook of formal argumentation - Volume 2. College Publications; 2021.

A Proofs

A.1 Proofs for the characterisation of principles

A.1.1 Conflict-freeness characterisation

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

Proof. (of Theo. 1) We prove both directions.

$\Rightarrow$: Suppose that S is conflict-free and that there is an attack (a,b) in $\mathscr{A}[S]_V$. By Definition 4, we have that $a,b \in S$ and that $(a,b) \in R$. This contradicts Definition 3 on conflict-freeness.

$\Leftarrow$: Suppose now that there are no attacks in $\mathscr{A}[S]_V$ and that S is not conflict-free. By Definition 3, there exists $a,b \in S$ such that $(a,b) \in R$. Thus, by Definition 4, (a,b) is in $\mathscr{A}[S]_V$. This contradicts the absence of attacks in $\mathscr{A}[S]_V$. □

A.1.2 Defence characterisation

Prop. 1 Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. If S is conflict-free, $G_{Def}(S)$ is a bipartite graph and S can always be one of its parts.

Proof. (of Prop. 1) Suppose S is conflict-free. Let $G_{Def}(S) = (A',R')$. By assumption, S is conflict-free, and thus $S \cap R^{-1}(S) = \varnothing$. Since by Definition 10 $A' = S \cup R^{-1}(S)$, S and $R^{-1}(S)$ then form a partition of A'. According to Definition 5, we must then show that for every $(a,b) \in R'$, $a \in S$ and $b \in R^{-1}(S)$ or $a \in R^{-1}(S)$ and $b \in S$. This is given by Definition 10. □

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

Proof. (of Theo. 2) We prove both directions.

$\Rightarrow$: Let $G_{Def}(S) = (A',R')$ and assume that $S \subseteq F_{\mathscr{A}}(S)$. Suppose now that there is a source vertex a in $R^{-1}(S)$ in $G_{Def}(S)$. By Definition 7, we have that $R'^{-}(a) = \varnothing$, which means there exists no $b \in A'$ such that $(b,a) \in R'$. However, by Definition 10 of $G_{Def}(S)$, we know that for all $(x,y) \in R'$, $x \in S$ and $y \in R^{-1}(S)$ or $x \in R^{-1}(S)$ and $y \in S$. This first means that, because S is conflict-free, $S \cap R^{-1}(S) = \varnothing$ and $a \in R^{-1}(S)$, it must be the case that $b \in S$. Hence, there exists no $b \in S$ such that $(b,a) \in R'$. In addition, since $R' \subseteq R$, there exists no $b \in S$ such that $(b,a) \in R$. As $a \in R^{-1}(S)$, there exists $c \in S$ such that $(a,c) \in R$. Hence, we know that there exists $a \in A$ with $(a,c) \in R$ for some $c \in S$ and such that there exists no $b \in S$ with $(b,a) \in R$. This contradicts the assumption that $S \subseteq F_{\mathscr{A}}(S)$.

$\Leftarrow$: Let $G_{Def}(S) = (A',R')$ and assume that there are no source vertices in $R^{-1}(S)$ in $G_{Def}(S)$. Suppose now that there is some $c \in S$ such that c is not acceptable w.r.t. S. By Definition 2, this means that there exists $a \in A$ such that $(a,c) \in R$ and there is no $b \in S$ with $(b,a) \in R$. First, notice that by Definition 6, $a \in R^{-1}(c)$ and so $a \in R^{-1}(S)$. Secondly, since $c \in S$, $a \in R^{-1}(S)$ and $(a,c) \in R$, by Definition 10, it holds that $c,a \in A'$ and $(a,c) \in R'$. Thus, by assumption, a is not a source vertex in $G_{Def}(S)$. Subsequently, there exists $b \in A'$ such that $(b,a) \in R'$. Moreover, by Definition 10, for all $(x,y) \in R'$, $x \in S$ and $y \in R^{-1}(S)$ or $x \in R^{-1}(S)$ and $y \in S$. Since $a \in R^{-1}(S)$ and S is conflict-free (i.e. $S \cap R^{-1}(S) = \varnothing$), it holds that $b \in S$ and $(b,a) \in R$. Thus, we have that $c \in S$ such that c is not acceptable w.r.t. S and for any $a \in A$ with $(a,c) \in R$, there is $b \in S$ with $(b,a) \in R$, a contradiction. □

A.1.3 Reinstatement characterisation

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

Proof. (of Theo. 3) Let $G_{Reins1}(S) = (A',R')$, $G_{Reins2}(S) = (A'',R'')$ and assume that $A' \subseteq S$ and that all vertices in $R^2(S) \setminus S$ are the endpoint of an arc whose origin is a source vertex in $G_{Reins2}(S)$. In other words, $A' \subseteq S$ and for every $x \in R^2(S) \setminus S$, there exists $y \in A'$ such that $(y,x) \in R'$ and $R'^{-}(y) = \varnothing$. Consider $a \in F_{\mathscr{A}}(S)$. This means that for every $b \in A$ such that $(b,a) \in R$, there exists $c \in S$ with $(c,b) \in R$. We must show that $a \in S$. Suppose first that a is not attacked in $\mathscr{A}$. That is to say, $R^-(a) = \varnothing$. By Definition 12, we have that $a \in A'$, and thus by assumption, $a \in S$. Suppose now that $R^-(a) \neq \varnothing$. By Definition 6, we have $a \in R^2(S)$ and for every $b \in A$ such that $(b,a) \in R$, $b \in R^{-1}(R^2(S))$. As such, by Definition 12, we have that $a,b,c \in A''$ and $(b,a),(c,b) \in R''$. Thus, for every $b \in A''$ such that $(b,a) \in R''$, $R''^{-}(b) \neq \varnothing$. Hence, all $b \in A''$ such that $(b,a) \in R''$ are not source vertices. Consequently, by assumption, $a \notin R^2(S) \setminus S$, but we know that $a \in R^2(S)$. It follows that $a \in R^2(S) \cap S$, and thus that $a \in S$. □

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

Proof. (of Theo. 4) Let $G_{Reins1}(S) = (A',R')$, $G_{Reins2}(S) = (A'',R'')$ and assume that $F_{\mathscr{A}}(S) \subseteq S$. Suppose now that either $A' \not\subseteq S$ or there is a vertex a in $R^2(S) \setminus S$ that is not the endpoint of an arc whose origin is a source vertex or in $R^2(S)$. In the first case, by Definition 12 we have that there exists $x \in A$ such that $R^-(x) = \varnothing$ and $x \notin S$. However, by Definition 2, this means that $x \in F_{\mathscr{A}}(S)$ and $x \notin S$, a contradiction. In the second case, we have $a \notin S$ and for every $b \in A''$ such that $(b,a) \in R''$, $R''^{-}(b) \neq \varnothing$ and $b \notin R^2(S)$. In other words, $b \notin R^2(S)$ and there exists $c \in A''$ with $(c,b) \in R''$. By Definition 12, $(x,y) \in R''$ if and only if $x \in R^{-1}(R^2(S))$ and $y \in R^2(S)$ or $x \in S$ and $y \in R^{-1}(R^2(S))$. Since $a \in R^2(S)$ and $b \notin R^2(S)$, we thus know that $b \in R^{-1}(R^2(S))$. In addition, also because $b \notin R^2(S)$, it must be the case that $c \in S$. So, for every $b \in A''$ such that $(b,a) \in R''$, there exists $c \in S$ with $(c,b) \in R''$. By Definition 12 again, we deduce that for every $b \in A$ such that $(b,a) \in R$, there exists $c \in S$ with $(c,b) \in R$. By Definition 2, this means that a is acceptable w.r.t. S and so that $a \in F_{\mathscr{A}}(S)$. Hence, by assumption, $a \in S$, a contradiction. □

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

Proof. (of Cor. 1) From Theorem 3 we have that if all vertices in $G_{Reins1}(S)$ are in S and all vertices in $R^2(S) \setminus S$ are the endpoint of an arc whose origin is a source vertex in $G_{Reins2}(S)$, then $F_{\mathscr{A}}(S) \subseteq S$. Theorem 4 gives that if $F_{\mathscr{A}}(S) \subseteq S$ then all vertices in $G_{Reins1}(S)$ are in S and all vertices in $R^2(S) \setminus S$ are the endpoint of an arc whose origin is a source vertex, or is in $R^2(S)$, in $G_{Reins2}(S)$. However, as $R^2(S)$ is conflict-free, a vertex in $R^2(S) \setminus S$ cannot be the endpoint of an arc whose origin is in $R^2(S)$ in $\mathscr{A}$, and so in $G_{Reins2}(S)$. It follows that if $F_{\mathscr{A}}(S) \subseteq S$ then all vertices in $G_{Reins1}(S)$ are in S and all vertices in $R^2(S) \setminus S$ are the endpoint of an arc whose origin is a source vertex, in $G_{Reins2}(S)$. Hence we have the two directions of the equivalence. □

A.1.4 Complement attack characterisation

Prop. 2 Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $G_{CA}(S)$ is a bipartite graph, S can always be one of its parts and all vertices in S are sources in it.

Proof. (of Prop. 2) Let $G_{CA}(S) = (A',R')$. By Definition 4, we have that $R' \subseteq R$. An obvious partition of A based on S is of course S and $A \setminus S$. By Definition 14, we know that for every $(a,b) \in R'$, $a \in S$ and $b \in A \setminus S$. Subsequently, by Definition 5, $G_{CA}(S)$ is a bipartite graph. In addition, since there is no $(b,a) \in R'$ such that $b \in A \setminus S$ and $a \in S$, it holds that for every $a \in S$, $R'^{-}(a) = \varnothing$. Thus, by Definition 7, every vertex of S is a source vertex in $G_{CA}(S)$. □

Theo. 5 Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. $A \setminus S \subseteq R^{+}(S)$ iff C_{CA} is satisfied by S, with C_{CA}: "there are no isolated vertices in the complement part of S in $G_{CA}(S)$".

Proof. (of Theo. 5) We prove both directions.

$\Rightarrow$: Suppose that $A \setminus S \subseteq R^{+}(S)$. Let $G_{CA}(S) = (A',R')$. Suppose now that there is an isolated vertex a in $A' \setminus S$. By Definition 14, we know that $A' = A$. As such, there exists $a \in A \setminus S$ such that a is isolated in $G_{CA}(S)$. By Definition 7, this means in particular that $R'^{-}(a) = \varnothing$ and thus that there is no $b \in A'$ with $(b,a) \in R'$. Again, in particular, we have that there is no $b \in S$ with $(b,a) \in R'$. However, by Definition 14, we have that $(x,y) \in R'$ if and only if $(x,y) \in R$, $x \in S$ and $y \notin S$. Hence, we deduce that there is no $b \in S$ with $(b,a) \in R$. Since $a \in A \setminus S$, this contradicts the assumption that $A \setminus S \subseteq R^{+}(S)$.

$\Leftarrow$: Suppose now that there are no isolated vertices in $A' \setminus S$ in $G_{CA}(S)$ and that $A \setminus S \not\subseteq R^{+}(S)$. From the first assumption, by Definition 14, we have that there are no isolated vertices in $A \setminus S$ in $G_{CA}(S)$. In particular, by Definition 7, we know that there is no $a \in A \setminus S$ such that $R'^{-}(a) = \varnothing$, or equivalently, for every $a \in A \setminus S$, there exists $b \in A$ such that $(b,a) \in R'$. By Definition 14, we have that $(x,y) \in R'$ if and only if $(x,y) \in R$, $x \in S$ and $y \notin S$, thus we deduce that for every $a \in A \setminus S$, there exists $b \in S$ such that $(b,a) \in R'$. From the second assumption, we have that there exists some $c \in A \setminus S$ such that there is no $b \in S$ with $(b,c) \in R$. By Definition 14, and because $c \in A \setminus S$ and $b \in S$, we deduce that there exists some $c \in A \setminus S$ such that there is no $b \in S$ with $(b,c) \in R'$, a contradiction of the first assumption. □

A.2 Proofs for the additional properties about principles

Prop. 3 Let $\mathscr{A} = (A,R)$, $S,S' \subseteq A$ and $\sigma \in \{CF, Def, Reins_1, Reins_2, CA\}$. If $S = S'$, then $G_\sigma(S) = G_\sigma(S')$.

Proof. (of Prop. 3)

1. Suppose that $S = S'$ but that $G_{CF}(S) \neq G_{CF}(S')$. Let $G_{CF}(S) = (A_S,R_S)$ and $G_{CF}(S') = (A_{S'},R_{S'})$. Since $G_{CF}(S) \neq G_{CF}(S')$, we have either $A_S \neq A_{S'}$ or $A_S = A_{S'}$ and $R_S \neq R_{S'}$. In the first case, by Definitions 8 and 4 we have $A_S = S$ and $A_{S'} = S'$, thus $S \neq S'$ which contradicts our first hypothesis. In the second case, also by Definition 4, we deduce that either there exists $(a,b) \in R$ s.t. $a,b \in S$ and $(a,b) \notin R_{S'}$ or there exists $(x,y) \in R$ s.t. $x,y \in S'$ and $(x,y) \notin R_S$. Using $S = S'$ leads to a contradiction in both cases.

2. Suppose that $S = S'$ but that $G_{Def}(S) \neq G_{Def}(S')$. Let $G_{Def}(S) = (A_S,R_S)$ and $G_{Def}(S') = (A_{S'},R_{S'})$. Since $G_{Def}(S) \neq G_{Def}(S')$, we have either $A_S \neq A_{S'}$ or $A_S = A_{S'}$ and $R_S \neq R_{S'}$. In the first case, by Definitions 10 and 4 we have $A_S = S \cup R^{-1}(S)$ and $A_{S'} = S' \cup R^{-1}(S')$, thus $S \neq S'$ which contradicts our first hypothesis. In the second case, we deduce that either there exists $(a,b) \in R_S$ s.t. $(a,b) \notin R_{S'}$ or

there exists $(x,y) \in R_{S'}$ s.t. $(x,y) \notin R_S$. Thus, by Definition 10, either there exists $a,b \in S \cup R^{-1}(S)$ s.t. $a \in S$ and $b \in R^{-1}(S)$ or $a \in R^{-1}(S)$ and $b \in S$, and $(a,b) \notin R_{S'}$, or there exists $x,y \in S' \cup R^{-1}(S')$ s.t. $x \in S'$ and $y \in R^{-1}(S')$ or $x \in R^{-1}(S')$ and $y \in S'$, and $(x,y) \notin R_S$. Using $S = S'$ leads to a contradiction in both cases.

3. Suppose that $S = S'$ but that $G_{Reins1}(S) \neq G_{Reins1}(S')$. Let $G_{Reins1}(S) = (A_S, R_S)$ and $G_{Reins1}(S') = (A_{S'}, R_{S'})$. Since $G_{Reins1}(S) \neq G_{Reins1}(S')$, we have either $A_S \neq A_{S'}$ or $A_S = A_{S'}$ and $R_S \neq R_{S'}$. In the first case, by Definitions 12 and 4 we have $A_S = \{x | R^-(x) = \varnothing\}$ and $A_{S'} = \{x | R^-(x) = \varnothing\}$, hence a contradiction arises. In the second case, since by definitions 12 we have $A_S = A_{S'} = \{x | R^-(x) = \varnothing\}$, we know that $R_S = R_{S'} = \varnothing$, hence a contradiction also arises.

4. Suppose that $S = S'$ but that $G_{Reins2}(S) \neq G_{Reins2}(S')$. Let $G_{Reins2}(S) = (A_S, R_S)$ and $G_{Reins2}(S') = (A_{S'}, R_{S'})$. Since $G_{Reins2}(S) \neq G_{Reins2}(S')$, we have either $A_S \neq A_{S'}$ or $A_S = A_{S'}$ and $R_S \neq R_{S'}$. In the first case, by Definitions 12 and 4 we have $A_S = S \cup R^2(S) \cup R^{-1}(R^2(S))$ and $A_{S'} = S' \cup R^2(S') \cup R^{-1}(R^2(S'))$, thus $S \neq S'$ which contradicts our first hypothesis. In the second case, we deduce that either there exists $(a,b) \in R_S$ s.t. $(a,b) \notin R_{S'}$ or there exists $(x,y) \in R_{S'}$ s.t. $(x,y) \notin R_S$. Thus, by Definition 12, either there exists $a,b \in S \cup R^2(S) \cup R^{-1}(R^2(S))$ s.t. $a \in R^{-1}(R^2(S))$ and $b \in R^2(S)$ or $a \in S$ and $b \in R^{-1}(R^2(S))$, and $(a,b) \notin R_{S'}$, or there exists $x,y \in S' \cup R^2(S') \cup R^{-1}(R^2(S'))$ s.t. $x \in R^{-1}(R^2(S'))$ and $y \in R^2(S')$ or $x \in S'$ and $y \in R^{-1}(R^2(S'))$, and $(x,y) \notin R_S$. Using $S = S'$ leads to a contradiction in both cases.

5. Suppose that $S = S'$ but that $G_{CA}(S) \neq G_{CA}(S')$. Let $G_{CA}(S) = (A_S, R_S)$ and $G_{CA}(S') = (A_{S'}, R_{S'})$. Since $G_{CA}(S) \neq G_{CA}(S')$, we have either $A_S \neq A_{S'}$ or $A_S = A_{S'}$ and $R_S \neq R_{S'}$. By Definition 10 we have $A_S = A$ and $A_{S'} = A$, hence it must be the case that $R_S \neq R_{S'}$. We deduce that either there exists $(a,b) \in R_S$ s.t. $(a,b) \notin R_{S'}$ or there exists $(x,y) \in R_{S'}$ s.t. $(x,y) \notin R_S$. Thus, by Definition 10, either there exists $a,b \in A$ s.t. $a \in S$ and $b \notin S$ and $(a,b) \notin R_{S'}$, or there exists $x,y \in A$ s.t. $x \in S'$ and $y \notin S'$ and $(x,y) \notin R_S$. Using $S = S'$ leads to a contradiction in both cases.

□

Prop. 4 Let $\mathscr{A} = (A,R)$, $S \subseteq A$. Let $\sigma \in \{CF, Def, Reins_2\}$: $G_\sigma(S) = (\varnothing, \varnothing)$ iff $S = \varnothing$. $G_{Reins1}(S) = (\varnothing, \varnothing)$ iff $\nexists a \in A$ st $R^-(a) = \varnothing$. $G_{CA}(S) = (\varnothing, \varnothing)$ iff $\mathscr{A} = (\varnothing, \varnothing)$.

Proof. (of Prop. 4) Immediate using Definition 4 and the definitions corresponding to each principle (Def. 8, Def. 10, Def. 12 and Def. 14). □

Prop. 5 Let $\mathscr{A} = (A,R)$, $S \subseteq A$ and $\sigma \in \{CF, Def, Reins_1, Reins_2, CA\}$. $G_\sigma(S)$ is a maximal explanation for σ on S.

Proof. (of Prop. 5) Immediate using Definition 4 and the definitions corresponding to each principle (Def. 8, Def. 10, Def. 12, Def. 14). □

Prop. 6 Let $\mathscr{A} = (A,R)$ and $S \subseteq A$. If S is conflict-free, $G_{CF}(S)$ is a minimal explanation for conflict-freeness on S.

Proof. (of Prop. 6) As S is conflict-free, Theorem 1 tells us that $R' = \varnothing$. Hence, $|R'| = 0$ and so $G_{CF}(S)$ is a minimal explanation for conflict-freeness on S. □

A.3 Proofs for the properties about semantics

Prop. 7 Let $\mathscr{A} = (A,R)$, $S,S' \subseteq A$ and $\sigma \in \{Adm, Co, Sta\}$. If $S = S'$, then $G_\sigma(S) = G_\sigma(S')$. Moreover, $G_\sigma(S)$ is a maximal explanation for σ on S.

Proof. (of Prop. 7) Immediate using Definitions 3, 16 and Propositions 3 and 5. □

Prop. 8 Let $\mathscr{A} = (A,R)$, $S \subseteq A$. $G_{Adm}(S) = (\varnothing,\varnothing)$ iff $S = \varnothing$. $G_{Co}(S) = (\varnothing,\varnothing)$ iff $S = \varnothing$ and $\{x \in A | R^-(x) = \varnothing\} = \varnothing$. $G_{Sta}(S) = (\varnothing,\varnothing)$ iff $\mathscr{A} = (\varnothing,\varnothing)$.

Proof. (of Prop. 8) Immediate using Definitions 3, 16 and Proposition 4. □