Classes of Explanations for the Verification Problem in Abstract Argumentation

Sylvie Doutre[1] **Théo Duchatelle**[2] **Marie-Christine Lagasquie-Schiex**[2]

[1] IRIT, Université Toulouse Capitole, France
[2] IRIT, Université Paul Sabatier, France
Sylvie.Doutre@irit.fr
Theo.Duchatelle@irit.fr
Marie-Christine.Lagasquie@irit.fr

Résumé

Le problème de vérification en argumentation abstraite consiste à déterminer si un ensemble est acceptable sous une sémantique donnée dans un graphe d'argumentation donné. Cet article s'attache à expliquer la réponse retournée. Des explications visuelles en termes de sous-graphes du cadre d'argumentation initial sont définies. Ces explications sont regroupées en classes, ce qui permet de sélectionner l'explication qui convient le mieux dans un contexte donné parmi l'ensemble des possibilités offertes. Des résultats montrent comment utiliser les aspects visuels de ces explications pour soutenir l'acceptabilité d'un ensemble d'arguments sous une sémantique. Les aspects computationnels d'explications spécifiques sont également étudiés.

Abstract

The Verification Problem in abstract argumentation consists in checking whether a set is acceptable under a given semantics in a given argumentation graph. Explaining why the answer is so is the challenge tackled by this paper. Visual explanations in the form of subgraphs of the initial argumentation framework are defined. These explanations are grouped into classes, allowing one to select the explanation that suits them best among the several offered possibilities. Results are provided on how to use the visual aspects of these explanations to support the acceptability of a set of arguments under a semantics. Computational aspects of specific explanations are also investigated.

1 Introduction

Abstract Argumentation is increasingly studied as a formal tool to provide explanations in the context of eXplainable Artificial Intelligence (XAI). The term argumentative XAI has emerged, with a number of application domains, ranging from machine learning, to decision, medicine or security (see [19] for an overview). [7] presents the current approaches of argumentative XAI and their open challenges, and underlines that explanations for the argumentative process itself are necessary too.

The basic argumentation process relies on an abstract structure which takes the form of a directed graph, whose nodes are arguments and edges represent attacks between arguments [10]. Characterising the acceptability of arguments can take the form of extension-based semantics : they define sets (extensions) of arguments which are collectively acceptable according to the semantics. The main questions which have been addressed so far in this context concern the global acceptability status of an argument or of a set of arguments, that is, why, under a given semantics, they belong to at least one extension (credulous acceptance) or to every extension (skeptical acceptance). The most common explanation approach consists in identifying set(s) of arguments which act as explanation(s), as in [12, 4, 5, 18, 13, 1]. However, 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. Moreover, this set approach does not highlight the attacks which are involved in the explanations.

Another question regarding the argumentation process concerns the *Verification Problem* Ver, defined as follows : given an Argumentation Framework $\mathcal{A}$, a set of arguments S and an extension-based semantics σ, "Is S an extension under σ in $\mathcal{A}$?". The answer to this problem is "yes" or "no". In order to explain why the answer is so, the *eXplanation Verification Problem* XVer can be defined using the question Q_σ : "Why is S (not) an extension under σ in

$\mathcal{A}$?".

[2] is one of the only approaches which has addressed this problem and which has provided answers for some acceptability semantics of [10] in the form of relevant subgraphs, as in [17, 15, 16]. Such a visual approach is particularly of interest for human agents, graphs having been shown to be helpful for humans to comply with argumentation reasoning principles [20]. This graph-based approach not only highlights arguments, but also attacks. In [2], properties that these answers satisfy have been established, depending on whether the answer to the corresponding verification problem is "yes" or "no". This methodology follows the line of [6] in that an explanation for a set S satisfying a semantic σ is a (set of) subgraph(s) G of $\mathcal{A}$ such that G satisfies a given graph property C. Another interesting point in [2] is that the considered semantics are based on a modular definition, which allows the explanations to be decomposed.

A limitation of [2] is however that, for each semantic principle, a *single* explanation subgraph is defined. It could be more realistic to consider classes (sets) of explanations. Indeed such classes would be particularly meaningful and useful when several agents, human or artificial, are involved around the explanation of a same problem, in that they offer a variety of answers, which all follow a same schema, but which may differ on their exact content. Any agent can choose or can be presented an explanation that suits them best, and any agent can understand an explanation given by another agent, different from theirs. Classes of explanations adapt to a wide set of agents.

As in [2], the approach that will be presented in this paper goes further, by considering the possibility that the answer to the Verification Problem is not known before an explanation be asked and given. In this case, the explanation graph and its interpretation offer at a same time the answer to the problem and a justification to this answer.

Only few related works can be found concerning this notion of classes of explanation. Such classes have already been proposed in [1] for the problem of credulous acceptance of an argument, where the authors consider explanation schemes made of several elements, one of them being fixed, the other ones varying from one explanation to another. Another related work is [4] in which the authors define a parametric computation of explanations. As such, it is more the computation processes that are grouped in classes, rather than the explanations (i.e. results of the processes) themselves.

Thus, our aim in the current paper is to define classes of explanations following a generic methodology, applied to classical semantics (conflict-free, admissible, stable, complete), by building up on the approach of [2]. Additional properties (emptyness, uniqueness, maximality, minimality, computation) of explanations on these new classes will be defined and investigated.

Sec. 2 recalls background notions relative to abstract argumentation, graph theory, and presents the explanation approach defined in [2]. Classes of explanations are defined in Sec. 3, Sec. 4 studies their properties; Sec. 5 shows how to compute their maximal and minimal explanations and illustrates the whole approach on an example. Sec. 6 concludes and presents some future works. Proofs of all the results can be found in [8]

2 Background notions

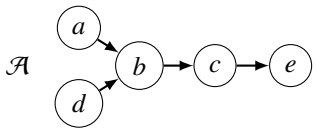
2.1 Argumentation and Graph Theory

We begin by recalling basic notions on Abstract Argumentation.

Definition 1 ([10]) *A* Dung's argumentation framework *(AF) 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 nodes are the arguments and the edges correspond to the attack relation).

Example 1 *Let consider* $\mathcal{A} = (A = \{a, b, c, d, e\}, R = \{(a, b), (d, b), (b, c), (c, e)\})$. $\mathcal{A}$ *is depicted by the following figure :*



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 ([10]) *Let* $\mathcal{A} = (A, R)$. *An argument* $a \in A$ *is* acceptable *wrt* $S \subseteq A$ *iff for all* $b \in A$, *if* bRa *then* $\exists c \in S$ *st* cRb.

Definition 3 ([10]) *Given* $\mathcal{A} = (A, R)$, *a subset S of A is :*

- *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* stable extension *iff S is conflict-free and S attacks any* $a \in A \setminus S$.

Example 2 *Let consider again* $\mathcal{A}$ *given in Ex. 1. Here there is a unique complete and stable extension :* $\{a, d, c\}$ *whereas there are 6 admissible sets :* $\{\}$, $\{a\}$, $\{d\}$, $\{a, c\}$, $\{d, c\}$, $\{a, d, c\}$.

The Verification Problem for the four semantics given in Def. 3 can be solved in polynomial time, as indicated by [11].

Example 3 *Considering $\mathcal{A}$ given in Ex. 1, an instance of the Verification problem could be : "Is $\{\}$ a stable extension ?" ; in this case the answer will "no". Another instance would be : "Is $\{a, d, c\}$ a complete extension ?" ; in this case the answer will "yes".*

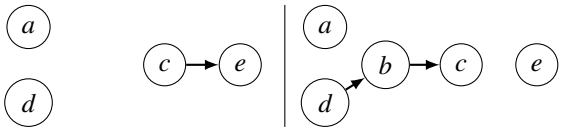
Since an AF can be represented using directed graphs, we also need to recall some basic notions of Graph Theory.

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]
- *G' is a* strict subgraph *of G iff it is a subgraph of G and either $V' \subset V$ or $E' \subset E$.* [2]
- *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.

Example 4 *Let consider $\mathcal{A}$ given in Ex. 1. An example of an induced (resp. spanning) subgraph of $\mathcal{A}$ is given in the left (resp. right) following figure :*



Induced and spanning subgraphs are examples of ways to compute a graph from another single graph. Another 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. The* union *of G_1 and G_2 is defined 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 :

1. G is then a *supergraph* of G'
2. G is then a *strict supergraph* of G'

Definition 6 (Bipartite Graph) *Let $G = (V, E)$ be a graph. G is* bipartite *(with* parts *T and U) iff there exist $T, U \subseteq V$ such that $T \cup U = V$ and $T \cap U = \emptyset$ (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).*

Some important functions can be defined over graphs.

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

Considering an argumentation framework, the successor (resp. predecessor) function represents the arguments that are attacked by (resp. are 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.

We then recall some notions on 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 $E^-(v) = \emptyset$ and it is said to be a* sink *iff $E^+(v) = \emptyset$. 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.

Example 5 *Let consider $\mathcal{A}$ given in Ex. 1. Argument a is a 3-step predecessor of e, whereas c is a predecessor of e (and obviously e is a 3-step successor of a, whereas e is a successor of c). Moreover, a and d are the sources of $\mathcal{A}$ and e is the sink of $\mathcal{A}$.*

2.2 Explanations in Argumentation

We recall the main definitions of what explanations are in [2] but only for those answering the questions about semantics results in abstract argumentation. These questions

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

are defined as follows : let σ represent a semantics among conflict-freeness, admissibility, completeness and stability, and given an argumentation framework $\mathcal{A} = (A, R)$ and some set $S \subseteq A$,

Q_σ : `Why is S (not) an extension under` σ `in` $\mathcal{A}$`?`

In order to answer these questions, and hence to provide explanations, [2] uses the decomposition of semantics into principles. The idea is to identify some properties that can be used to provide a modular characterization of semantics. We refer the reader to [9] for further details. Given a set S, the following principles are considered :

Conflict-freeness (CF) :	No internal conflicts in S
Defence (Def) :	$\forall x \in S$, x is acceptable wrt S
Reinstatement (Re) :	$\forall x$ acceptable wrt S, $x \in S$
Complement Attack (CA) :	S attacks all arguments not in S

Note that the reinstatement principle has been split into two sub-principles. Indeed, to decide whether a set S of arguments contains all the arguments acceptable wrt S, one must consider on the one hand the arguments that are unattacked and thus acceptable by lack of attackers (sub-principle denoted by Re_1), and on the other hand the arguments for which S defeats all the attackers (sub-principle denoted by Re_2).

The following has been proven in [9].

Proposition 1 *Let $\mathcal{A} = (A, R)$ and $S \subseteq A$. S is :*

Conflict-free	*iff S respects*	*{CF}*
Admissible	*iff S respects*	*{CF, Def}*
Complete	*iff S respects*	*{CF, Def, Re_1, Re_2}*
Stable	*iff S respects*	*{CF, CA}*

With this result, a straightforward answer arises for Q_σ : a set S is an extension under semantics σ because it respects all the principles listed for σ in Prop. 1. This moves the burden of explanation from semantics to principles. So, in order to answer Q_σ, we are going to answer intermediate questions on principles. Let $\pi \in \{CF, Def, Re_1, Re_2, CA\}$ represent a principle. Given an argumentation framework $\mathcal{A} = (A, R)$ and some set $S \subseteq A$, the questions we will define answers for are :

Q_π : `Why does (not) S respect principle` π`?`

[2] defines visual answers to these questions. These answers take the form of a graph. This allows for the answers to be drawn, as well as to study their visual (i.e. structural) properties. More precisely, as argumentation frameworks are graphs themselves, the answers given are subgraphs of an argumentation framework.

Definition 9 ([2]) *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $\pi \in \{CF, Def, Re_1, Re_2, CA\}$. $G_\pi(S)$ is defined as :*

$$
\begin{aligned}
G_{CF}(S) &= \mathcal{A}[S]_V \\
G_{Def}(S) &= (\mathcal{A}[S \cup R^{-1}(S)]_V) \\
&\quad [\{(a,b) \in R \mid (a \in R^{-1}(S) \text{ and } b \in S) \\
&\qquad \text{or } (a \in S \text{ and } b \in R^{-1}(S))\}]_E \\
G_{Re_1}(S) &= \mathcal{A}[\{a \in A | R^-(a) = \emptyset\}]_V \\
G_{Re_2}(S) &= (\mathcal{A}[S \cup R^2(S) \cup R^{-1}(R^2(S))]_V) \\
&\quad [\{(a,b) \in R \mid (a \in R^{-1}(R^2(S)), b \in R^2(S)) \\
&\qquad \text{or } (a \in S, b \in R^{-1}(R^2(S)))\}]_E \\
G_{CA}(S) &= \mathcal{A}[\{(a,b) \in R \mid a \in S \text{ and } b \notin S\}]_E
\end{aligned}
$$

Moreover the interpretation of these subgraphs can be done using a "checking procedure" in order to explicitly identify if the given subset satisfies or not the concerned principle :

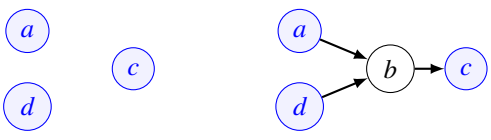
Definition 10 ([2]) *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $\pi \in \{CF, Def, Re_1, Re_2, CA\}$. Let G be a subgraph of $\mathcal{A}$. The checking procedure $C_\pi(G)$ is defined as :*

$$
\begin{aligned}
C_{CF}(G) &= \textit{no attacks in } G \\
C_{Def}(G) &= \textit{no source vertices in } R^{-1}(S) \textit{ in } G \\
C_{Re_1}(G) &= \textit{all vertices in } G \textit{ are in } S \\
C_{Re_2}(G) &= \textit{all vertices in } R^2(S) \setminus S \textit{ are endpoint of an arc whose origin is a source vertex in } G \\
C'_{Re_2}(G) &= \textit{all vertices in } R^2(S) \setminus S \textit{ are endpoint of an arc whose origin is a source vertex or is in } R^2(S), \textit{ in } G \\
C_{CA}(G) &= \textit{no isolated vertices in the complement part of } S \textit{ in } G
\end{aligned}
$$

For each principle π, [2] has proven that the subgraph G_π associated with the corresponding checking procedure C_π provides an explanation that answers question Q_π.[4] More precisely, if a set S respects a principle π, then G_π verifies C_π, otherwise it does not. When the principles are combined into a semantics σ, the answer to Q_σ is the corresponding set of subgraphs along with their corresponding checking procedures.

Example 6 *Let consider $\mathcal{A}$ given in Ex. 1 and $S = \{a, d, c\}$. The question we are interested in is : "Why is S an extension under admissibility in $\mathcal{A}$?". This question comes down to wondering : "Why S satisfies conflict-freeness CF and defense Def ?". So, an explanation of why S is admissible is a set which contains the explanation for CF and the explanation for Def.*

The $G_{CF}(S)$ and $G_{Def}(S)$ explanations are given in the following figure :



4. This result is slightly more complex in the case of reinstatement. See [2] and Sec. 3.3.

There is no attack in G_{CF}, hence C_{CF} is satisfied. And so we can conclude that S is conflict-free.

Concerning G_{Def}, note that neither e nor (c, e) belong to this explanation since they have no impact on the defence of S. Then applying C_{Def} on G_{Def}, we can see that each attacker of S (here only b) is not a source vertex; so S also satisfies the defence principle.

This allows this form of explanation to be used for two purposes as indicated in the introduction : when the answer to the corresponding verification problem is known, that is, when we know that a set is (resp. is not) acceptable under a given semantics or principle, G_π on which C_π is (resp. is not) verified, offers a visual explanation of the situation, answering XVer. When the answer to the verification problem is not known, G_π and the verification of whether C_π holds or not offers at the same time an answer to Ver and an explanation of this answer.

3 Classes of explanations

In this paper, we are interested in refining the notion of explanation proposed in [2] and recalled in Sec. 2.2. Indeed, considering Ex. 6 leads to the following remark : for explaining the respect of the defence principle it seems useless to consider the two defenders of c in G_{Def} (only one is enough for proving that c is defended). So, in order to propose a more flexible notion of explanation, another approach based on the notion of *classes of explanations* is presented in this section. Of course the definition of these classes allows to recover the explanations described in [2] but also it results in the *possibility of producing several explanations for the same question*.

Hence, for each principle π, we define our explanations so that they contain at least enough information to be able to decide whether or not S respects π. We then prove that our explanations can be used in conjunction with the checking procedures recalled in Def. 10.

3.1 Explanation about Conflict-freeness

To decide whether a set S of arguments is conflict-free, one must know whether or not there are attacks among its arguments. Thus, we firstly require our explanation to contain only arguments of S, and secondly to contain only attacks between these arguments. However, with only these two constraints, it may happen that no attacks are displayed on the explanation when there are some in the original framework, leading at best to an impossibility to decide or at worst, an incorrect decision. Hence, we add a third constraint, which is that if conflicts exist between arguments of S, then at least one must be present in the explanation.

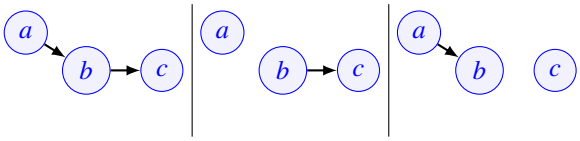
Definition 11 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $X = \{(a, b) \in R \mid a, b \in S\}$. The subgraph (A', R') of $\mathcal{A}$ is an* explanation to Q_{CF} *iff*
- $A' = S$
- $R' \subseteq X$
- *If $X \neq \emptyset$, then $R' \neq \emptyset$*

Note that the subgraph G_{CF} recalled in Def. 9 obviously belongs to the class of explanations for conflict-freeness. Moreover, in [2], a result concerning the structural property of explanations for conflict-freeness has been given : a set of arguments is conflict-free iff there is no attack in the subgraph corresponding to its explanation (checking procedure C_{CF} recalled in Def. 10). This result can be extended to all the subgraphs captured by our class of explanations.

Theorem 1 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and (A', R') be an explanation to Q_{CF}. S is conflict-free iff $C_{CF}(A', R')$ is satisfied by S.*

This provides a way of deciding whether a set is conflict-free based on an explanation to Q_{CF}. Note that this also provides a way of deciding whether a set is *not* conflict-free, hence the possibility of handling the negative version of Q_{CF}. The same goes for all the other equivalence results concerning the other principles.

Example 7 *Let consider $\mathcal{A}$ given in Ex. 1 and $S = \{a, b, c\}$. There are 3 explanations for Q_{CF}, each of them proving that S is not conflict-free :*



3.2 Explanation about Defence

To decide whether a set S of arguments contains only arguments that are acceptable wrt S, one must know whether or not this set defeats all its attackers. Thus, we firstly require our explanation to contain only arguments of S and its attackers, and secondly to contain only attacks from S to its attackers and vice versa. To make sure the attackers are spotted as such, we further require that all the attacks of the second type are contained in the explanation. However, with only these two constraints, it may happen that no attacks targeting a specific attacker are displayed on the explanation when there are some in the original framework. As we wish the explanation to show how S defends itself, this situation is certainly undesirable. Hence, we add a third constraint, which is that if an attacker is attacked by S, then at least one attack from S to this attacker must be present in the explanation.

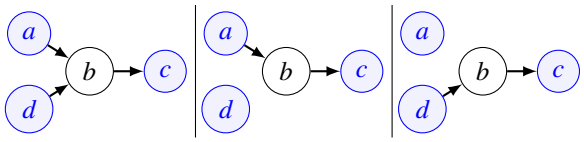
Definition 12 *Let $\mathcal{A} = (A, R)$ and $S \subseteq A$. Consider $X = \{(b, a) \in R \mid b \in R^{-1}(S), a \in S\}$ and $Y = \{(a, b) \in R \mid a \in S, b \in R^{-1}(S)\}$. The subgraph (A', R') of $\mathcal{A}$ is an* explanation to Q_{Def} *iff*

— $A' = S \cup R^{-1}(S)$
— $X \subseteq R' \subseteq X \cup Y$
— $\forall b \in R^{-1}(S)$, *if* $b \in R^{+1}(S)$, *then* $\exists (a, b) \in R'$ *with* $a \in S$

Note that the subgraph G_{Def} recalled in Def. 9 obviously belongs to the class of explanations for defence. Moreover it has been shown in [2] that a conflict-free set of arguments defends all its arguments iff there is no source vertex among its attackers in $G_{Def}(S)$ (checking procedure C_{Def} recalled in Def. 10). This result can be extended to all the subgraphs captured by our class of explanations.

Theorem 2 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ be a conflict-free set of arguments and (A', R') be an explanation to Q_{Def}. $S \subseteq F_{\mathcal{A}}(S)$ iff $C_{Def}(A', R')$ is satisfied by S.*

Example 8 *Let consider $\mathcal{A}$ given in Ex. 1 and $S = \{a, c, d\}$. There are 3 explanations for proving that S satisfies the defence principle :*



Additionally, the next result extends a similar result given in [2] providing more insight on the behavior of an explanation for defence : when computed using a conflict-free set, the explanation for defence takes the form of a bipartite graph.

Proposition 2 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and (A', R') be an explanation to Q_{Def}. If S is conflict-free, (A', R') is a bipartite graph and S can always be one of its parts.*

The two previous results can thus be used to decide whether a set of arguments effectively defends all its arguments or if it is not conflict-free.

3.3 Explanation about Reinstatement

The first part of the reinstatement principle concerns unattacked arguments. All these arguments are acceptable wrt S and should thus belong to S. Thus, we firstly require our explanation to contain only unattacked arguments, and secondly to contain no attacks (which results from the only arguments displayed being unattacked). However, with only these two constraints, it may happen that an unattacked argument not belonging to S is not displayed on the explanation. Hence, we add a third constraint, which is that if there exists unattacked arguments that are not in S, then at least one must be present in the explanation.

Definition 13 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $X = \{a \in A \mid R^{-1}(a) = \emptyset\}$. The subgraph (A', R') of $\mathcal{A}$ is an* explanation to Q_{Re_1} *iff*

— $S \cap X \subseteq A' \subseteq X$
— $R' = \emptyset$
— *If* $(A \setminus S) \cap X \neq \emptyset$, *then* $\exists a \in (A \setminus S) \cap X$ *with* $a \in A'$

The second part concerns arguments for which S defeats the attackers. These arguments must belong to S if S defeats all of their attackers. Thus, we firstly require our explanation to contain the arguments of S, the arguments that S defends (two steps of the attack relation from S), and the attackers of these arguments. Secondly, we require it contains only the attacks from S to the attackers and from the attackers to the arguments S defends. In addition, we require that all the attacks of the second type are displayed on the explanation, so that none is missed. However, with only these two constraints, it may happen that no attacks targeting a specific attacker are displayed on the explanation when there are some in the original framework. Hence, we add a third constraint, which is that if an attacker is attacked by S, then at least one attack from S to this attacker must be present in the explanation.

Definition 14 *Let $\mathcal{A} = (A, R)$ and $S \subseteq A$. Consider $X = \{(b, c) \in R \mid b \in R^{-1}(R^{+2}(S)), c \in R^{+2}(S)\}$ and $Y = \{(a, b) \in R \mid a \in S, b \in R^{-1}(R^{+2}(S))\}$. The subgraph (A', R') of $\mathcal{A}$ is an* explanation to Q_{Re_2} *iff*

— $A' = S \cup R^{+2}(S) \cup R^{-1}(R^{+2}(S))$
— $X \subseteq R' \subseteq X \cup Y$
— *For every* $b \in R^{-1}(R^{+2}(S))$, *if* $b \in R^{+1}(S)$, *then* $\exists (a, b) \in R'$ *with* $a \in S$

Note that the subgraph G_{Re_1} (resp. G_{Re_2}) recalled in Def. 9 obviously belongs to the class of explanations for the first (resp. second) part of the principle of reinstatement. Moreover in the case of reinstatement, two results have been proven in [2] and can be extended to all the subgraphs captured by our class of explanations.

The first one shows how to conclude that a set contains all the arguments that it effectively defends from both parts of the explanation on reinstatement.

Theorem 3 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$, (A', R') be an explanation to Q_{Re_1} and (A'', R'') be an explanation to Q_{Re_2}. If $C_{Re_1}(A', R')$ and $C_{Re_2}(A'', R'')$ are satisfied by S then $F_{\mathcal{A}}(S) \subseteq S$.*

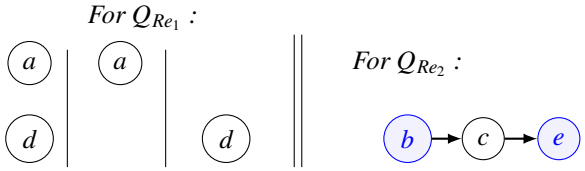
The second results shows the behavior of both parts of the explanation on reinstatement if computed on a set that contains all the arguments it effectively defends.

Theorem 4 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$, (A', R') be an explanation to Q_{Re_1} and (A'', R'') be an explanation to Q_{Re_2}. If $F_{\mathcal{A}}(S) \subseteq S$ then $C_{Re_1}(A', R')$ and $C'_{Re_2}(A'', R'')$ are satisfied by S.*

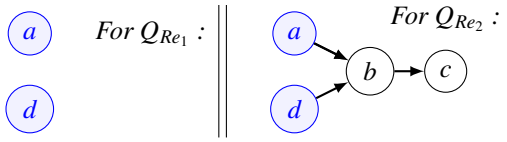
From Th. 3 and 4 follows the next corollary, which shows an equivalence result :

Corollary 1 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ such that $R^2(S)$ is conflict-free, (A', R') be an explanation to Q_{Re_1} and (A'', R'') be an explanation to Q_{Re_2}. $F_{\mathcal{A}}(S) \subseteq S$ iff $C_{Re_1}(A', R')$ and $C_{Re_2}(A'', R'')$ are satisfied by S.*

Example 9 *Let consider $\mathcal{A}$ given in Ex. 1 and $S = \{b, e\}$. There are 3 explanations for proving that S does not satisfy the first reinstatement principle (some unattacked arguments are not in S ; here it is the case for a and d) and one for proving that S satisfies the second reinstatement principle (the arguments defended by S are in S) :*



Let consider now $S = \{a, d\}$. There are one explanation for proving that S satisfies the first reinstatement principle (any unattacked argument is in S) and another one for proving that S does not satisfy the second reinstatement principle (some arguments defended by S are not in S ; here it is the case of c) :



3.4 Explanation about Complement Attack

To decide whether a set S of arguments attacks its complement, one must know whether or not all the arguments not in S are attacked by S. Thus, we firstly require our explanation to contain all the arguments of the original framework (S and its complement), and secondly to contain only attacks from S to arguments not in S. However, with only these two constraints, it may happen that no attacks targeting a specific argument outside of S are displayed on the explanation when there are some in the original framework. Hence, we add a third constraint, which is that if an argument not in S is attacked by S, then at least one attack from S to this argument must be present in the explanation.

Definition 15 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $X = \{(a, b) \in R \mid a \in S, b \notin S\}$. The subgraph (A', R') of $\mathcal{A}$ is an* explanation to Q_{CA} *iff*

- $A' = A$
- $R' \subseteq X$
- $\forall b \in A \setminus S$, *if* $b \in R^{+1}(S)$, *then* $\exists (a, b) \in R'$ *with* $a \in S$

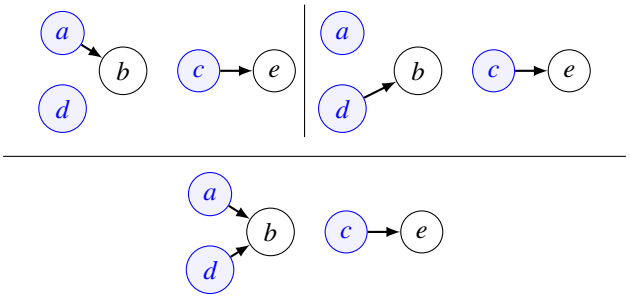
Note that the subgraph G_{CA} recalled in Def. 9 obviously belongs to the class of explanations for the principle of complement attack. Moreover concerning this principle, it was proven in [2] that a set of arguments attacks its complement iff there are no isolated vertices in $G_{CA}(S)$ and the explanation subgraph is always a bipartite graph with the arguments of S being the only possible origins for attacks. We extend these results to our class of explanations for complement attack.

Theorem 5 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and (A', R') be an explanation to Q_{CA}.*

$A \setminus S \subseteq R^{+1}(S)$ iff $C_{CA}(A', R')$ is satisfied by S.

(A', R') is a bipartite graph, S can always be one of its parts and all vertices in S are sources in it.

Example 10 *Let consider $\mathcal{A}$ given in Ex. 1 and $S = \{a, b, c\}$. There are three explanations to Q_{CA} proving that S satisfies the principle of complement attack :*



4 Properties of Explanations

We now turn to the definition of explanation properties and to a formal study of our classes of explanations according to them. This will allow to highlight some particular kinds of explanations, as well as to better understand their behavior. The properties that we will consider are : minimality, maximality, emptyness and uniqueness.

4.1 Some specific explanations

In this section, we identify some specific properties that could be respected by our explanations.

Minimality, Maximality A minimal (resp. maximal) explanation is an explanation which contains the least (resp. all the) possible amount of information. In a sense, a minimal explanation only provides what is required to explain whereas a maximal explanation in fact provides everything that might be relevant to explain, even if it might be redundant.

Definition 16 *Let $\mathcal{A} = (A, R)$ and $S \subseteq A$. The subgraph (A', R') of $\mathcal{A}$ is a* minimal *(resp.* maximal*)* explanation *that answers Q_π iff there is no subgraph (A'', R'') of $\mathcal{A}$ which is also an explanation that answers Q_π such that (A'', R'')*

is a strict subgraph of (A', R') *(resp.* (A', R') *is a strict subgraph of* (A'', R'')*).*

Example 7 (cont'd) *In this example, the maximal explanation is the first one and the two other ones are minimal.*

Emptyness The notion of an empty explanation is one that should be avoided when providing explanations, in the sense that it somewhat represents the incapacity of the system to answer the question that has been asked.

Definition 17 *Let* $\mathcal{A} = (A, R)$ *and* $S \subseteq A$*. The subgraph* (A', R') *is an* empty explanation that answers Q_π *iff* $(A', R') = (\emptyset, \emptyset)$.

Uniqueness We consider an explanation to be unique when there is only one of its kind. Although we defined classes of explanations in an attempt to represent all the different points of view that could emerge as to how to answer a question, in some situations, there can only be one way to answer that question.

Definition 18 *Let* $\mathcal{A} = (A, R)$ *be a graph. The subgraph* (A', R') *is a* unique explanation that answers Q_π *iff there is no subgraph* (A'', R'') *with* $(A'', R'') \neq (A', R')$ *which is also an explanation that answers* Q_π.

Example 9 (cont'd) *In this example, the explanations for the second reinstatement principle are unique (for* $S = \{b, e\}$ *or* $S = \{a, d\}$*) whereas the explanation for the first reinstatement principle is unique for* $S = \{a, d\}$ *but not for* $S = \{b, e\}$.

Minimality and uniqueness are seen as explanation principles in [13]. However, these two notions are defined differently in [13], relatively to another concept of explanation based on sets of arguments, not on subgraphs, as we do.

4.2 Properties of specific explanations

Here, we provide the results of our formal study on our explanations using the aforementioned properties. We begin with empty explanations. The results show that, although empty explanations can occur, they only do so in very specific situations.

The following theorem establishes a characterisation of empty explanations, which generalises a similar result given in [2]. Moreover if this empty explanation occurs, it is the only possible one.

Theorem 6 *Let* $\mathcal{A} = (A, R)$ *and* $S \subseteq A$*.* $(\emptyset, \emptyset)$ *is an explanation that answers*

1. Q_π *with* $\pi \in \{CF, Def, Re_2\}$ *iff* $S = \emptyset$.
2. Q_{Re_1} *iff* $\{a \in A \mid R^{-1}(a) = \emptyset\} = \emptyset$.
3. Q_{CA} *iff* $\mathcal{A} = (\emptyset, \emptyset)$.

If $(\emptyset, \emptyset)$ *is an explanation to* Q_π *with* $\pi \in \{CF, Def, Re_1, Re_2, CA\}$*, then it is unique.*

Now, we turn to our study of maximal explanations. The next theorem states for each principle that there is only one possible maximal explanation.

Theorem 7 *Let* $\mathcal{A} = (A, R)$ *and* $S \subseteq A$*. If* (A', R') *is a maximal explanation that answers* Q_π *with* $\pi \in \{CF, Def, Re_1, Re_2, CA\}$*, then it is the unique maximal explanation that answers* Q_π.

In the worst case, the number of explanations can be exponential in the size of some specific sets of elements, depending on the type of explanation (for instance the set of the attacks between the arguments belonging to the extension S in the case of explanations for the conflict-free principle). Thus considering only minimal explanations is a first step towards a computationally efficient method.

Nevertheless, as it turns out, there can be multiple minimal explanations in general for each principle. The next theorem studies the relation between minimal and maximal explanations and shows that the maximal explanation is exactly the union of all the minimal explanations.

Theorem 8 *Let* $\mathcal{A} = (A, R)$ *and* $S \subseteq A$*. Consider* $\pi \in \{CF, Def, Re_1, Re_2, CA\}$ *and let* (A', R') *be the maximal explanation that answers* Q_π *and* M *be the set of all minimal explanations that answers* Q_π*. Then,* $(A', R') = \bigcup_{G \in M} G$.

This result opens the way to algorithmic solutions since, for a given principle, a maximal explanation covers all the possible explanations (the minimal ones but also all the intermediate explanations).

5 Computation of Explanations

This section investigates how to compute the maximal and minimal explanations of a class.

Maximal Explanations It turns out that maximal explanations exactly correspond to the explanations defined in [2] (recalled in Def. 9) :

Proposition 3 *Let* $\mathcal{A} = (A, R)$*,* $S \subseteq A$ *and* $\pi \in \{CF, Def, Re_1, Re_2, CA\}$*.* $G_\pi(S)$ *is the maximal explanation that answers* Q_π.

This result entails that maximal explanations can be computed using only the graph operators of induced and spanning subgraphs, thus ensuring an efficient computation.

Note that Prop. 3 aggregated with Th. 7 allows to recover a unicity result given in [2].

From Maximal to Minimal Explanations In order to compute the minimal explanations for each principle π, we start from the maximal explanation :

Given $\mathcal{A} = (A, R)$ and $S \subseteq A$, $(A', R') \leftarrow G_\pi(S)$

Then, we gradually remove elements until obtaining a minimal explanation. This leads to five algorithms Alg_π (one for each principle π) that are built following the same schema. They also use the same condition for stopping the removal : "it remains at most one element to remove". The only differences between these algorithms concern the "nature" of the removed elements : [5]

For *CF*, removal of attacks between elements of S :

While $|R'| > 1$
$(x, y) \leftarrow choose(R')$; $R' \leftarrow R' \setminus \{(x, y)\}$

For *Def*, for each attacker of S that is not in S, removal of attacks that target it :

For $y \in R^{-1}(S) \setminus S$
While $|R'^{-1}(y)| > 1$
$x \leftarrow choose(R'^{-1}(y))$; $R' \leftarrow R' \setminus \{(x, y)\}$

For Re_1, removal of unattacked arguments not in S :

While $|A' \setminus S| > 1$
$x \leftarrow choose(A' \setminus S)$; $A' \leftarrow A' \setminus \{x\}$

For Re_2, for each argument that is an attacker of the arguments S defends and that is not defended by S, removal of attacks that target it :

For $y \in R^{-1}(R^{+2}(S)) \setminus R^{+2}(S)$
While $|R'^{-1}(y)| > 1$
$x \leftarrow choose(R'^{-1}(y))$; $R' \leftarrow R' \setminus \{(x, y)\}$

For *CA*, for each argument that is not in S, removal of attacks that target it :

For $y \in A \setminus S$
While $|R'^{-1}(y)| > 1$
$x \leftarrow choose(R'^{-1}(y))$; $R' \leftarrow R' \setminus \{(x, y)\}$

Our algorithms are sound and complete for the computation of minimal explanations as shown by the following proposition.

Proposition 4 *Let $\mathcal{A} = (A, R)$, $S \subseteq A$ and $\pi \in \{CF, Def, Re_1, Re_2, CA\}$. Algorithm Alg_π using $\mathcal{A}$ and S as inputs is sound and complete for the computation of a minimal explanation that answers Q_π.*

The computation of minimal explanations thus relies on the computation of maximal explanations, and the removal of some arcs (or arguments) in them. The computation of maximal explanations is already known to be polynomial (see [2]). Moreover the complexity of the removal operation in the worst case is linear in the number of removed elements and this number is either quadratic in the number of vertices in the graph when these elements are attacks (so for any principle except the one for the first part of reinstatement), or linear in the number of vertices in the graph when these elements are vertices (for the first part of reinstatement). From these considerations, our algorithms can be considered as computationally efficient.

Note also that a slight adaptation of these algorithms could produce random intermediate explanations (so neither minimal, nor maximal). This could be done by randomly stopping the removal process after a parametric number of steps. It is also the way to create more specific explanations responding to certain constraints given by users (for instance, explanations containing only x elements of a given type among the $y \geq x$ existing ones).

6 Conclusion and Future Work

This paper has defined *classes of explanations* for principles and semantics for the explanation Verification Problem XVer in Abstract Argumentation. These classes of explanations have been studied according to general properties such as maximality, minimality, emptyness and uniqueness. They extend and generalize the single explanations of [2], allowing more flexibility in the choice of explanations that could be presented to potential users. Moreover we have established that the explanations of [2] correspond to the maximal explanations of the defined classes, thus providing a way to compute them using graph operators. A procedure to compute minimal explanations from the maximal ones has also been provided and proven sound and complete for each class of explanations.

These results make an implementation of the proposed approach ready to be done. From this implementation, like in any XAI approach, as underlined by [7], an empirical evaluation should be conducted to assess to which extent these visual explanations actually are helpful for human agents to understand the answer to the Verification Problem. This is a first important future work, clearly related with the explainability social process described in [14].

Moreover, this evaluation could also provide a first study about what is a "best explanation" and how to select it. It is therefore also related to a second important future work : how take into account the issue of the "realizability", or personalization of an explanation. Indeed, one may have in mind parts of an explanation (some arguments, some attacks), but not a correct and complete explanation ; determining whether there exists such an explanation, and providing it, would ensure a personalized answer. In order to do so, a deeper investigation of the inner structure of the classes of explanation, and more specifically of the links they could have with lattices, may be of help.

This contribution and its research avenues will be of help in any application which uses computational abstract argumentation [19, 7].

5. Note that these elements are generally attacks except in the case of the principle Re_1.

In addition, the approach may be extended in several directions :

— to some semantics that use additional principles like maximality/minimality for set inclusion, for instance, the preferred or grounded semantics ; in this case, some new visual criteria must be identified *in order to be able to explain why* a given set is or is not a preferred or a grounded extension ; note that the visualization difficulty is not related to the complexity of the underlying problem (since the Ver problem for the grounded semantics is a polynomial problem whereas it is an exponential one for the preferred semantics) ;

— to contrastive questions : single explanations to such questions have been proposed in [2] ; their generalisation to classes of explanations may be studied using the work presented here since, very often, a contrastive question can be viewed as the conjunction of some specific single questions.

Moreover, extending XVer to additional semantics and additional questions can be considered as an attempt to produce a generic approach for the computation of explanations, on the model of the approach of [3].

Finally, more notions of Graph Theory may be investigated in order to provide other kinds of visual explanations. In particular, the notion of graph isomorphism seems of great interest, especially to provide ways of reasoning by association (explaining a result via a structurally identical argumentation framework that one already accepted).

Références

[1] Baumann, Ringo et Markus Ulbricht: *Choices and their Consequences - Explaining Acceptable Sets in Abstract Argumentation Frameworks*. Dans *Proc. of KR*, pages 110–119, Online event, 2021. IJCAI Organization.

[2] Besnard, Philippe, Sylvie Doutre, Théo Duchatelle et Marie Christine Lagasquie-Schiex: *Explaining Semantics and Extension Membership in Abstract Argumentation*. Intelligent Systems with Applications, 16 :200118, 2022.

[3] Besnard, Philippe, Sylvie Doutre, Théo Duchatelle et Marie Christine Lagasquie-Schiex: *Generic logical encoding for argumentation*. Journal of Logic and Computation, 2022, ISSN 0955-792X. https://doi.org/10.1093/logcom/exac039.

[4] Borg, AnneMarie et Floris Bex: *A Basic Framework for Explanations in Argumentation*. IEEE Intelligent Systems, 36(2) :25–35, 2021.

[5] Borg, AnneMarie et Floris Bex: *Necessary and Sufficient Explanations for Argumentation-Based Conclusions*. Dans *Proc. of ECSQARU*, tome 12897 de *LNCS*, pages 45–58, Prague, Czech Republic, 2021. Springer.

[6] Cocarascu, Oana, Kristijonas Čyras, Antonio Rago et Francesca Toni: *Explaining with argumentation frameworks mined from data*. Dans *Proc. of DEXAHAI*, Southampton, United Kingdom, 2018.

[7] Čyras, Kristijonas, Antonio Rago, Emanuele Albini, Pietro Baroni et Francesca Toni: *Argumentative XAI : A Survey*. Dans *Proc. of IJCAI*, pages 4392–4399, Online Event / Montreal, Canada, 2021. IJCAI Organization.

[8] Doutre, Sylvie, Théo Duchatelle et Marie Christine Lagasquie-Schiex: *Classes of Explanations for the Verification Problem in Abstract Argumentation*. Research Report IRIT/RR–2022–09–FR, IRIT : Institut de Recherche en Informatique de Toulouse, France, 2022.

[9] Doutre, Sylvie et Jean Guy Mailly: *Quantifying the Difference Between Argumentation Semantics*. Dans *Proc. of COMMA*, tome 287, pages 255–262, Potsdam, Germany, 2016. IOS Press.

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

[11] Dvorák, Wolfgang et Paul E Dunne: *Computational problems in formal argumentation and their complexity*. Handbook of formal argumentation, 4 :631–688, 2018.

[12] Fan, Xiuyi et Francesca Toni: *On Computing Explanations in Argumentation*. Dans *Proc. of AAAI*, pages 1496–1502, Austin, Texas, USA, 2015. AAAI Press.

[13] Liao, Beishui et Leendert van der Torre: *Explanation Semantics for Abstract Argumentation*. Dans *Proc. of COMMA*, tome 326, pages 271–282, Perugia, Italy, 2020. IOS Press.

[14] Miller, Tim: *Explanation in artificial intelligence : Insights from the social sciences*. Artificial Intelligence, 267 :1–38, 2019.

[15] Niskanen, Andreas et Matti Järvisalo: *Smallest Explanations and Diagnoses of Rejection in Abstract Argumentation*. Dans *Proc. of KR*, pages 667–671, Rhodes, Greece, 2020. IJCAI Organization.

[16] Racharak, Teeradaj et Satoshi Tojo: *On Explanation of Propositional Logic-based Argumentation System*. Dans *Proc. of ICAART*, tome 2, pages 323–332, Online Streaming, 2021. SCITEPRESS.

[17] Saribatur, Zeynep Gozen, Johannes Peter Wallner et Stefan Woltran: *Explaining Non-Acceptability in Abstract Argumentation*. Dans *Proc. of ECAI*, tome 325, pages 881–888, Santiago de Compostela, Spain, 2020. IOS Press.

[18] Ulbricht, Markus et Johannes Peter Wallner: *Strong Explanations in Abstract Argumentation*. Dans *Proc. of AAAI*, pages 6496–6504, Online event, 2021. AAAI Press.

[19] Vassiliades, Alexandros, Nick Bassiliades et Theodore Patkos: *Argumentation and explainable artificial intelligence : a survey*. The Knowledge Engineering Review, 36, 2021.

[20] Vesic, Srdjan, Bruno Yun et Predrag Teovanovic: *Graphical Representation Enhances Human Compliance with Principles for Graded Argumentation Semantics*. Dans *Proc. of AAMAS*, pages 1319–1327, Auckland, New Zealand, 2022. IFAAMAS.