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 our work. In this extended abstract, we focus on a small part of this aim considering only the defence principle and proposing explanations in order to explain why a subset of arguments defends all its elements. These explanations are visual, in the sense that they take the form of subgraphs of the initial argumentation framework. They form a class, whose properties are investigated.

KEYWORDS
XAI; Visual Explanations; Formal Abstract Argumentation

Extended Abstract

Visual Explanations for Defence in Abstract Argumentation

Sylvie Doutre
IRIT, Toulouse 1 University
Toulouse, France
sylvie.doutre@irit.fr

Théo Duchatelle
IRIT, Toulouse 3 University
Toulouse, France
theo.duchatelle@irit.fr

Marie-Christine Lagasquie-Schiex
IRIT, Toulouse 3 University
Toulouse, France
lagasq@irit.fr

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 our work. In this extended abstract, we focus on a small part of this aim considering only the defence principle and proposing explanations in order to explain why a subset of arguments defends all its elements. These explanations are visual, in the sense that they take the form of subgraphs of the initial argumentation framework. They form a class, whose properties are investigated.

Visual Explanations for Defence in Abstract Argumentation

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

Our aim is thus to build up on the approach of [2] and to go further by defining classes of explanations following a generic methodology. Due to space limitations, we only consider here a single principle: the defence (Def). Additional principles and semantics related complete results can be found in [8].

Given an argumentation framework $\mathcal{A} = (\mathcal{A}, \mathcal{R})$ and some set $S \subseteq \mathcal{A}$, the questions we will define answers for are:

$\text{Q}_{\text{Def}}$: Why does (not) $S$ respect the principle $\text{Def}$?

Let us recall that in [2], the answer for this question is the graph $G_{\text{Def}}(S)$ defined as: given $\mathcal{A} = (\mathcal{A}, \mathcal{R})$, $S \subseteq \mathcal{A}$,

$$G_{\text{Def}}(S) = (\mathcal{A}[S \cup R^{-1}(S)], E)$$

where $E = \{(a, b) \in R \mid (a \in R^{-1}(S) \text{ and } b \in S) \text{ or } (a \in S \text{ and } b \in R^{-1}(S))\}$.

An interpretation of this subgraph using a "checking procedure", denoted $C_{\text{Def}}(G)$, has also been proposed: given $\mathcal{A} = (\mathcal{A}, \mathcal{R})$, $S \subseteq \mathcal{A}$, let $G$ be a subgraph of $\mathcal{A}$,

$C_{\text{Def}}(G) = "\text{There are no source vertices in } R^{-1}(S) \text{ in } G."$

Hence, the subgraph $G_{\text{Def}}$ associated with the checking procedure $C_{\text{Def}}$ provides an explanation that answers the question $\text{Q}_{\text{Def}}$ if a

$[2]$ is one of the only approaches which has addressed this problem so far, and which has provided answers for some acceptability semantics of [10] in the form of relevant subgraphs, as in [16–18] and following the methodology of [5]. 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. Moreover, in [2], the semantics are based on a modular definition (see [9]), which allows the explanations to be decomposed considering independently each principle.

A limitation of [2] is however that, for each semantic principle, a single explanation subgraph is defined. It would be more realistic to consider classes (sets) of explanations. 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, and in [3] for a parametric computation of explanations.

The reader can refer to [16] for the basic notions on Abstract Argumentation and to [11] for some information about the Verification Problem.

set $S$ respects the principle $Def$, then $G_{Def}$ verifies $C_{Def}$, otherwise it does not.

The definition of a “class” of explanations in place of a “single” one not only allows one to recover the explanations described in [2] but also results in the possibility of producing several explanations for the same question. Thus, it takes into account the different points of view that may emerge and focus on different aspects.

In the case of the defence principle, to decide whether a set $S$ of arguments defends all its elements, one must know whether or not this set defeats all its attackers. Thus, we firstly require an explanation to contain only arguments of $S$ and their attackers, and secondly to contain only attacks from $S$ to these attackers and vice versa. To make sure that 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 whereas 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 1. Let $\mathcal{A} = (A, R)$ and $S \subseteq A$. Consider $X = \{(a, b) \in R \mid b \in R^{-1}(S), a \in S\}$ 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 $a \in S$

The following results issued from [2] can be extended to all the subgraphs captured by our class of explanations.

Theorem 1. Let $\mathcal{A} = (A, R), S \subseteq A$ be a conflict-free set of arguments and $(A', R')$ be an explanation to $Q_{Def}$. $S$ defends all its elements iff $C_{Def}(A', R')$ is satisfied by $S$. Moreover, if $S$ is conflict-free, $(A', R')$ is a bipartite graph and $S$ can always be one of its parts.

Some other interesting properties hold:\footnote{We use here the classical notion of minimality and maximality: a minimal (resp. maximal) explanation is such that none of its strict subgraphs (supergraph) is also an explanation. These notions have been introduced in several papers (see for instance [14]) and a discussion about this point could be an interesting future work.}

Theorem 2. Let $\mathcal{A} = (A, R), (A', R')$ be a subgraph of $\mathcal{A}$ and $S \subseteq A$.

- $(\varnothing, \varnothing)$ is an explanation that answers $Q_{Def}$ iff $S = \varnothing$.
- If $(\varnothing, \varnothing)$ is an explanation to $Q_{Def}$, then it is unique.
- If $(A', R')$ is a maximal explanation that answers $Q_{Def}$, then it is the unique maximal explanation that answers $Q_{Def}$.
- If $(A', R')$ is the maximal explanation that answers $Q_{Def}$ and $\mathcal{M}$ is the set of all minimal explanations that answer $Q_{Def}$, then $(A', R') = \bigcup_{G \in \mathcal{M}} G$.
- $G_{Def}(S)$ is the maximal explanation that answers $Q_{Def}$.

In order to compute the minimal explanations, we will start from the maximal explanation, and gradually remove elements until obtaining a minimal explanation. In the case of the defence, Algorithm Alg$_{Def}$ is sound and complete for the computation of minimal explanations.

\footnote{Some of these results extend similar results given in [2], confirming that our approach generalises [2].}

<table>
<thead>
<tr>
<th>Alg$<em>{Def}$ Computation of a minimal answer to $Q</em>{Def}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Require:</strong> $\mathcal{A} = (A, R), S \subseteq A$</td>
</tr>
<tr>
<td>1. $(A', R') \leftarrow G_{Def}(S)$</td>
</tr>
<tr>
<td>2. for $y \in R^{-1}(S) \setminus S$ do</td>
</tr>
<tr>
<td>3. while $</td>
</tr>
<tr>
<td>4. $x \leftarrow \text{choose}(R^{-1}(y))$</td>
</tr>
<tr>
<td>5. $R' \leftarrow R' \setminus {(x, y)}$</td>
</tr>
<tr>
<td>6. end while</td>
</tr>
<tr>
<td>7. end for</td>
</tr>
<tr>
<td>8. return $(A', R')$</td>
</tr>
</tbody>
</table>

As an illustration of the whole approach, consider that $\mathcal{A} = ((a, b, c, d, e), ((a, b), (d, b), (b, c), (c, e)))$ and $S = \{a, d, e\}$. There exist three explanations showing why $S$ defends all its elements, the first one corresponding to $G_{Def}(S)$ and the two others being minimal:

![Explanation Diagram](image)

Note that neither $e$ nor $(c, e)$ belong to an explanation for $S$. Moreover, applying $C_{Def}$ on each of these three explanations, we can see that each attacker of $S$ (here only $b$) is not a source vertex; so $S$ satisfies the defence principle.

Based on these results, the proposed approach is ready to be implemented. Like in any XAI approach (as underlined by [7]), this implementation should go along with an empirical assessment to decide to which extent these visual explanations actually are helpful for human agents. This is a first important future work, clearly related with the social process described in [15].

A second one is to take into account the notion of “realizability” or personalization/adaptability (see [15]) of an explanation: an agent 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 that an explanation that is personalized for the agent would be provided. In order to do so, a deeper investigation of the inner structure of our class of explanations, and more specifically of the links that we think it could have with lattices, may be of help.

Contrastive questions may also be addressed: generalising those proposed in [2] to classes of explanations, following the approach presented in the current paper, could be addressed. Extending XVer to additional semantics (preferred and grounded notably) may also be considered, and an attempt for producing a generic approach could be done.

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 the user already accepted).
REFERENCES


