Visual Explanations for Defence in Abstract Argumentation

Extended Abstract

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.

KEYWORDS

XAI; Visual Explanations; Formal Abstract Argumentation

ACM Reference Format:
Sylvie Doutre, Théo Duchatelle, and Marie-Christine Lagasquie-Schiex. 2023. Visual Explanations for Defence in Abstract Argumentation: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

Abstract Argumentation is increasingly studied as a formal tool to provide explanations of decisions made using an Artificial Intelligence system in the context of eXplainable Artificial Intelligence (XAI). Several multi-agent extensions of Abstract Argumentation exist (see the survey by [21]). Another recent survey by [6] indicates that Argumentation can be used to generate explanations in various domains, notably in multi-agent systems as in [13], and that explanations for the argumentative process itself are also necessary.

The basic argumentation process relies on an abstract structure $\mathcal{A} = (A, R)$ which takes the form of a directed graph, whose nodes are arguments (the set A) and edges represent attacks between arguments (the binary relation R) [10]. In this context, an argument is acceptable if it belongs to an extension (set of arguments respecting some principles, e.g. conflict-freeness, defence). Several questions can be addressed with their corresponding explanations (see for instance [1, 3, 4, 12, 14, 19]). In this paper, we only consider a specific one regarding the argumentation process, the *eXplanation Verification Problem*, defined using the question Q_σ: let $\mathcal{A} = (A, R)$ and $S \subseteq A$, "Why is S (not) an extension under σ in $\mathcal{A}$?".[1]

[1]The reader can refer to [10] for the basic notions on Abstract Argumentation and to [11] for some information about the Verification Problem.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[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.[2] 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.

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 and related complete results can be found in [8].

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

Q_{Def}: `Why does (not) S respect the principle` Def?[3]

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

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

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

$C_{Def}(G)$ = "There are no source vertices in $R^{-1}(S)$ in G".[4]

Hence, the subgraph G_{Def} associated with the checking procedure C_{Def} provides an explanation that answers the question Q_{Def}: if a

[2]The reader is considered familiar with some basic notions of Graph Theory. Particularly the different kinds of subgraphs: Let G be a graph, a subgraph G' of G is a graph included in G. In an induced subgraph, $G' = G[S]_V$ 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, $G' = G[S]_E$ of G by a set of *edges* S, all the vertices of G are present but some edges of G can be missing.
[3]Recall that a set $S \subseteq A$ defends all its elements iff $\forall a \in S, \forall b \in A$ with $(b, a) \in R$ then $\exists c \in S$ st $(c, b) \in R$.
[4]That means that any attacker of an element of S must be attacked in G.

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 it 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 $X = \{(b, a) \in R \mid b \in R^{-1}(S), a \in S\}$ 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 $a \in S$

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:[5][6]

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 M is the set of all minimal explanations that answer Q_{Def}, then, $(A', R') = \bigcup_{G \in M} G$.*
- *$G_{Def}(S)$ 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.

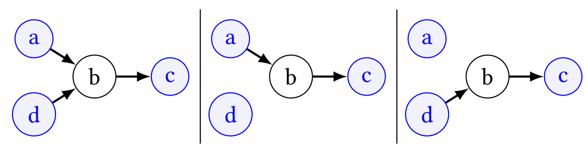
[5] We use here the classical notion of minimality and maximality: a minimal (resp. maximal) explanation is such that none of its strict subgraphs (supergraphs) is also an explanation. These notions have been introduced in several papers (see for instance [14]) and a discussion about this point could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

[6] Some of these results extend similar results given in [2], confirming that our approach generalises [2].

$\mathbf{Alg}_{Def}$ Computation of a minimal answer to Q_{Def}

Require: $\mathcal{A} = (A, R), S \subseteq A$
$(A', R') \leftarrow G_{Def}(S)$
for $y \in R^{-1}(S) \setminus S$ **do**
  while $|R'^{-1}(y)| > 1$ **do**
    $x \leftarrow choose(R'^{-1}(y))$
    $R' \leftarrow R' \setminus \{(x, y)\}$
  end while
end for
return (A', R')

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 $S = \{a, d, c\}$. 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:



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

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

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

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

[4] AnneMarie Borg and Floris Bex. 2021. Necessary and Sufficient Explanations for Argumentation-Based Conclusions. In *Proc. of ECSQARU (LNCS, Vol. 12897)*. Springer, Prague, Czech Republic, 45–58.

[5] Oana Cocarascu, Kristijonas Čyras, Antonio Rago, and Francesca Toni. 2018. Explaining with argumentation frameworks mined from data. In *Proc. of DEXAHAI*. Southampton, United Kingdom.

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

[7] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence IJCAI*, Zhi-Hua Zhou (Ed.). IJCAI Organization, Online Event / Montreal, Canada, 4392–4399. https://doi.org/10.24963/ijcai.2021/600

[8] Sylvie Doutre, Théo Duchatelle, and Marie-Christine Lagasquie-Schiex. 2022. *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.

[9] Sylvie Doutre and Jean-Guy Mailly. 2016. Quantifying the Difference Between Argumentation Semantics. In *Proc. of COMMA*, Vol. 287. IOS Press, Potsdam, Germany, 255–262.

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

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

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

[13] Yang Gao, Francesca Toni, Hao Wang, and Fanjiang Xu. 2016. Argumentation-Based Multi-Agent Decision Making with Privacy Preserved. In *Proc. of AAMAS*. ACM, Singapore, 1153–1161.

[14] Beishui Liao and Leendert van der Torre. 2020. Explanation Semantics for Abstract Argumentation. In *Proc. of COMMA*, Vol. 326. IOS Press, Perugia, Italy, 271–282.

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

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

[17] Teeradaj Racharak and Satoshi Tojo. 2021. On Explanation of Propositional Logic-based Argumentation System. In *Proc. of ICAART*, Vol. 2. SCITEPRESS, Online Streaming, 323–332.

[18] Zeynep Gozen Saribatur, Johannes Peter Wallner, and Stefan Woltran. 2020. Explaining Non-Acceptability in Abstract Argumentation. In *Proc. of ECAI*, Vol. 325. IOS Press, Santiago de Compostela, Spain, 881–888.

[19] Markus Ulbricht and Johannes Peter Wallner. 2021. Strong Explanations in Abstract Argumentation. In *Proc. of AAAI*. AAAI Press, Online event, 6496–6504.

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

[21] Liuwen Yu, Dongheng Chen, Lisha Qiao, Yiqi Shen, and Leendert van der Torre. 2021. A Principle-based Analysis of Abstract Agent Argumentation Semantics. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*. 629–639. https://doi.org/10.24963/kr.2021/60