Interpretability of Gradual Semantics in Abstract Argumentation

Jérôme Delobelle, Serena Villata

Interpretability of Gradual Semantics in Abstract Argumentation

Jérôme Delobelle and Serena Villata

Université Côte d'Azur, Inria, CNRS, I3S, Sophia-Antipolis, France
jerome.delobelle@inria.fr villata@i3s.unice.fr

Abstract. Argumentation, in the field of Artificial Intelligence, is a formalism allowing to reason with contradictory information as well as to model an exchange of arguments between one or several agents. For this purpose, many semantics have been defined with, amongst them, gradual semantics aiming to assign an acceptability degree to each argument. Although the number of these semantics continues to increase, there is currently no method allowing to explain the results returned by these semantics. In this paper, we study the interpretability of these semantics by measuring, for each argument, the impact of the other arguments on its acceptability degree. We define a new property and show that the score of an argument returned by a gradual semantics which satisfies this property can also be computed by aggregating the impact of the other arguments on it. This result allows to provide, for each argument in an argumentation framework, a ranking between arguments from the most to the least impacting ones w.r.t. a given gradual semantics.

Keywords: Abstract Argumentation; Gradual Semantics; Interpretability

1 Introduction

The issue of interpreting the results obtained by Artificial Intelligence (AI) methods is receiving an increasing attention both in the AI community but also from a wider audience. In particular, the ability to interpret the rationale behind the results (e.g., classifications, decisions) returned by an artificial intelligent agent is of main importance to ensure the transparency of the interaction between the two entities in order to accomplish cooperative tasks. According to Miller [13], *interpretability* is the degree to which an observer can understand the cause(s) of a result. An algorithm, a program or a decision is said to be interpretable if it is possible to identify the elements or the features that have the greatest impact on (and thus lead to) the result. This term must not be confused with the term *explanation* which is the answer to a why-question or with the term *justification* which explains why a result is good, but does not necessarily aim to give an explanation of the process. Despite the numerous (formal and empirical) approaches [12, 11, 17, 9] to tackle the problem of interpretability of artificial intelligent systems, it is still an open research problem. As highlighted

by Mittelstadt et al. [14], artificial argumentation [3] may play an important role in addressing this open issue, thanks to its inner feature of combining decision making with the pro and con arguments leading to a certain decision.

In this paper, we aim to study, from a formal point of view, how to cast the notion of interpretability in abstract argumentation so that the reasons leading to the acceptability of one or a set of arguments in a framework may be explicitly assessed. More precisely, this research question breaks down into the following sub-questions: *(i)* how to formally define and characterise the notion of *impact* of an argument with respect to the acceptability of the other arguments in the framework? and *(ii)* how does this impact play a role in the interpretation process of the acceptability of arguments in the framework?

To answer these questions, we start from the family of graded semantics [6, 4], and we select two semantics which present different features so that we can show the generality of our approach to characterise the notion of impact. In particular, we select the h-categorizer semantics initially proposed by Besnard and Hunter [5] and the counting semantics from Pu et al. [16]. In both approaches, the acceptability of an argument, differently from standard Dung's semantics [10] where arguments are either (fully) *accepted* or *rejected*, is represented through an acceptability *degree* in the range $[0, 1]$. Roughly, we say that the impact of a certain argument (or a set of arguments) on the degree of acceptability of another argument can be measured by computing the difference between the current acceptability degree of the argument and its acceptability degree when the first argument is deleted. We study the formal properties of the notion of impact instantiated through these two graded semantics both for cyclic and acyclic abstract argumentation frameworks. Finally, we show that studying the impact of an argument on the other arguments allows us to answer to some main needs in terms of interpretability of argument-based decision maker's resolutions.

The remainder of the paper is as follows: in Section 2, we provide same basics about gradual semantics and more precisely, the h-categorizer [5] and the counting semantics [16], Section 3 discusses the notion of impact of an argument in an argumentation framework and its formal properties, Section 4 focuses on the balanced impact property, in Section 5 we highlight how the notion of impact and its properties play a role on the interpretability of abstract argumentation frameworks and the acceptability of the arguments. The discussion of the related literature and conclusions end the paper.

2 Preliminaries

An abstract argumentation framework (AF) is a set of abstract arguments connected by an attack relation.

Definition 1 (AF). *An (abstract) argumentation framework (AF) is a tuple* $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ *where* $\mathcal{A}$ *is a finite and non-empty set of (abstract) arguments, and* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is a binary relation on* $\mathcal{A}$*, called the attack relation. For two arguments* $x, y \in \mathcal{A}$*, the notation* $(x, y) \in \mathcal{R}$ *(or* $x\mathcal{R}y$*) means that* x *attacks* y*.*

Definition 2 (Non-attacked set of arguments). *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF. The set of arguments $X \subseteq \mathcal{A}$ is non-attacked if $\forall x \in X, \nexists y \in \mathcal{A} \backslash X$ s.t. $(y, x) \in \mathcal{R}$.*

Notation 1 *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $x, y \in \mathcal{A}$. A **path** P from y to x, noted $P(y, x)$, is a sequence $\langle x_0, \ldots, x_n \rangle$ of arguments in $\mathcal{A}$ such that $x_0 = x$, $x_n = y$ and $\forall i < n, (x_{i+1}, x_i) \in \mathcal{R}$. The length of the path P is n (i.e., the number of attacks it is composed of) and is denoted by $l_P = n$. A **cycle** is a path from x to x and a **loop** is a cycle of length 1.*
*Let $\mathcal{R}_n^-(x) = \{y \mid \exists P(y, x)$ with $l_P = n\}$ be the multiset of arguments that are bound by a path of length n to the argument x. Thus, an argument $y \in \mathcal{R}_n^-(x)$ is a direct attacker (resp. defender) of x if $n = 1$ (resp. $n = 2$). More generally, y is an **attacker** (resp. **defender**) of x if n is odd (resp. even).*

A gradual semantics assigns to each argument in an argumentation framework a score, called *acceptability degree*, depending on different criteria. This degree must be selected among the interval $[0, 1]$.

Definition 3 (gradual semantics). *A gradual semantics is a function $\mathcal{S}$ which associates to any argumentation framework $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ a function $\mathtt{Deg}_F^{\mathcal{S}} : \mathcal{A} \to [0, 1]$. Thus, $\mathtt{Deg}_F^{\mathcal{S}}(x)$ represents the acceptability degree of $x \in \mathcal{A}$.*

h-categorizer semantics [5, 15] This gradual semantics uses a categorizer function to assign a value to each argument which captures the relative strength of an argument taking into account the strength of its attackers, which itself takes into account the strength of its attackers, and so on.

Definition 4. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework. The **categorizer function** $\mathtt{Deg}_F^{Cat} : \mathcal{A} \to]0, 1]$ is defined such that $\forall x \in \mathcal{A}$,*

$$\mathtt{Deg}_F^{Cat}(x) = \begin{cases} 1 & \text{if } \mathcal{R}_1^-(x) = \emptyset \\ \frac{1}{1 + \sum_{y \in \mathcal{R}_1^-(x)} \mathtt{Deg}_F^{Cat}(y)} & \text{otherwise} \end{cases}$$

Counting semantics [16] This gradual semantics allows to rank arguments by counting the number of their respective attackers and defenders. In order to assign a value to each argument, they consider an AF as a dialogue game between the proponents of a given argument x (i.e., the defenders of x) and the opponents of x (i.e., the attackers of x). The idea is that an argument is more acceptable if it has many arguments from proponents and few arguments from opponents. Formally, they first convert a given AF into a matrix $M_{n \times n}$ (where n is the number of arguments in AF) which corresponds to the adjacency matrix of AF (as an AF is a directed graph). The matrix product of k copies of M, denoted by M^k, represents, for all the arguments in AF, the number of defenders (if k is even) or attackers (if k is odd) situated at the beginning of a path of length k. Finally, a normalization factor N (e.g., the matrix infinite norm) is applied to M in order to guarantee the convergence, and a damping factor α is used to have a more refined treatment on different length of attacker and defenders (i.e., shorter attacker/defender lines are preferred).

Definition 5 (Counting model). *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework with $\mathcal{A} = \{x_1, \ldots, x_n\}$, $\alpha \in]0, 1[$ be a damping factor and $k \in \mathbb{N}$. The n-dimensional column vector v over $\mathcal{A}$ at step k is defined by,*

$$v_\alpha^k = \sum_{i=0}^{k} (-1)^i \alpha^i \tilde{M}^i \mathcal{I}$$

where $\tilde{M}$ is the normalized matrix such that $\tilde{M} = M/N$ with N as normalization factor and $\mathcal{I}$ the n-dimensional column vector containing only 1s.
The counting model of F is $v_\alpha = \lim_{k \to +\infty} v_\alpha^k$. The strength value of $x_i \in \mathcal{A}$ is the i^{th} component of v_α, denoted by $\mathtt{Deg}_F^{CS}(x_i)$.

3 Impact Measure

The impact of an argument on another argument can be measured by computing the difference when this argument exists and when it is deleted. To capture this notion of deletion, we need to define the complement operator which deletes a set of arguments from the initial argumentation framework w.r.t. a given argument (i.e., the targeted argument of the impact). These changes have also a direct impact on the set of attacks because the attacks directly related to the deleted arguments (attacking as well as attacked) are automatically deleted too.

Definition 6. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, $X \subseteq \mathcal{A}$ and $y \in \mathcal{A}$. The **complement operator** $\ominus$ is defined as $\mathsf{F} \ominus_y X = \langle \mathcal{A}', \mathcal{R}' \rangle$, where*

- $\mathcal{A}' = \mathcal{A} \backslash (X \backslash \{y\})$;
- $\mathcal{R}' = \{(x, z) \mid (x, z) \in \mathcal{R}$ *and* $x \in \mathcal{A} \backslash X, z \in \mathcal{A} \backslash X\}$.

Let us first formalise how to compute the impact of a non-attacked set of arguments on a given argument before generalising it for every set of arguments.

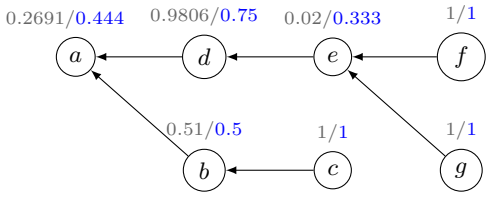
3.1 Impact of a non-attacked set of arguments

The impact of a non-attacked set of arguments X on the degree of acceptability of an argument y can be measured by computing the difference between the current acceptability degree of y and its acceptability degree when X is deleted.

Definition 7 (Impact of a non-attacked set of arguments). *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, $y \in \mathcal{A}$ and $X \subseteq \mathcal{A}$ be a non-attacked set of arguments. Let $\mathcal{S}$ be a gradual semantics. The impact of X on y is defined as follows:*

$$\mathtt{Imp}_F^{\mathcal{S}}(X, y) = \mathtt{Deg}_F^{\mathcal{S}}(y) - \mathtt{Deg}_{F \ominus_y X}^{\mathcal{S}}(y)$$

Globally, this definition is implicitly included in the formula of existing gradual semantics. A proof of this is that it is possible to compute the score of an argument by combining its basic score and the impact of each argument in the AF: $\mathtt{Deg}_{\mathrm{F}}^{\mathcal{S}}(y) = 1 + \mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\mathcal{A}, y)$. Figure 1 illustrates this idea where $\mathtt{Deg}_{\mathrm{F}}^{\mathrm{CS}}(a) = 1 + \mathtt{Imp}_{\mathrm{F}}^{\mathrm{CS}}(\mathcal{A}, a) = 1 + (\mathtt{Deg}_{\mathrm{F}}^{\mathcal{S}}(a) - \mathtt{Deg}_{\mathrm{F} \ominus_a \mathcal{A}}^{\mathcal{S}}(a)) = 1 - 0.7309 = 0.2691$.



$\mathtt{Imp}_{\mathsf{F}}^{\mathcal{S}}(X, a)$	CS	Cat
$X = \mathcal{A}$	-0.7309	-0.556
$X = \{e, f, g\}$	0.0138	0.044
$X = \{b, c\}$	-0.2547	-0.127
$X = \{f, g\}$	-0.2215	-0.056
$X = \{c\}$	0.2353	0.0808

Fig. 1. On the left hand side, an AF with, above each argument, its scores returned by the counting semantics (with $\alpha = 0.98$) and the h-categorizer semantics [CS/Cat]. On the right hand side, the table contains the impact of some non-attacked sets of arguments on the degree of acceptability of argument a.

Measuring the impact of these sets of arguments could be interesting for applications like the online debate platforms where people can argue on a given topic. A debate can be formalised with an AF which has, in many cases, a tree-shaped structure meaning that several sub-debates exist. For example, the arguments for/against the vegan diet can be divided into several categories like the environmental impact, health impact, psychological effects, etc. Checking the impact of these different categories (i.e., the sub-trees in the AF) on the topic implies to better know the influence of each part on the debate.

3.2 General impact

As it stands, the formula of the impact (Definition 7) cannot be used for an attacked set of arguments. Indeed, calculating the impact of $\{e\}$ on a in Fig. 1 reverts to compute the impact of $\{e, f, g\}$ on a because, by deleting e, the path from f and g (the direct attackers of e) to a are also removed implying to indirectly take into account the impact of f and g on a too.
In order to compute the impact of any set of arguments X on an argument y, we propose to consider the degree of acceptability of y when the arguments in X are the strongest (i.e., when their direct attackers are deleted). The fact that these arguments are attacked will be taken into account during the computation of the impact of these attackers on y.

Definition 8 (Impact). *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, $y \in \mathcal{A}$ and $X \subseteq \mathcal{A}$. Let $\mathcal{S}$ be a gradual semantics. The **impact** of X on y is:*

$$\mathtt{Imp}_{F}^{\mathcal{S}}(X, y) = \mathtt{Deg}^{\mathcal{S}}_{F \ominus_y (\bigcup_{x \in X} \mathcal{R}_1^-(x))}(y) - \mathtt{Deg}^{\mathcal{S}}_{F \ominus_y X}(y)$$

This definition generalises Definition 7 because if $\bigcup_{x \in X} \mathcal{R}_1^-(x) = \emptyset$ (meaning that X is non-attacked) then the two formulae are equivalent.
As the acceptability degree of an argument is between 0 and 1 (see Definition 3), the impact of a set of arguments on an argument is in the interval $[-1, 1]$.

Proposition 1. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, $y \in \mathcal{A}$ and $X \subseteq \mathcal{A}$. Let $\mathcal{S}$ be a gradual semantics. We have $\mathtt{Imp}_{F}^{\mathcal{S}}(X, y) \in [-1, 1]$.*

Three categories of impact can be defined, i.e., *positive*, *negative* and *neutral*.

Definition 9. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF, $y \in \mathcal{A}$ and $X \subseteq \mathcal{A}$. Let $\mathcal{S}$ be a gradual semantics. We say that X has a **positive impact** on y if $\mathtt{Imp}_F^{\mathcal{S}}(X, y) > 0$, X has a **negative impact** on y if $\mathtt{Imp}_F^{\mathcal{S}}(X, y) < 0$, X has a **neutral impact** on y if $\mathtt{Imp}_F^{\mathcal{S}}(X, y) = 0$.*

Note that the fact that a set of arguments has a specific impact (positive, negative or neutral) does not mean that all arguments belonging to this set also have this specific impact. For example, in Fig. 1, we can see that, when CS is used, the set $\{e, f, g\}$ has a positive impact whereas only e has a positive impact (f and g have a negative impact).
In order to be used for interpretability (Section 5), we define three notations to select the single arguments which have either a positive, negative or neutral impact on another argument.

Notation 2 *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $y \in \mathcal{A}$. Let $\mathcal{S}$ be a gradual semantics.*
$\mathtt{I}_{\mathcal{S}}^{+}(y) = \{x \in \mathcal{A} \mid \{x\}$ *has a positive impact on* $y\}$
$\mathtt{I}_{\mathcal{S}}^{-}(y) = \{x \in \mathcal{A} \mid \{x\}$ *has a negative impact on* $y\}$
$\mathtt{I}_{\mathcal{S}}^{=}(y) = \{x \in \mathcal{A} \mid \{x\}$ *has a neutral impact on* $y\}$.

Example 1 *Let us compute the impact of each single argument in the AF visualised in Fig. 1 on a when CS is used ($\alpha = 0.98$). Focusing on e, we have $\mathtt{Imp}_F^{CS}(\{e\}, a) = \mathtt{Deg}_{F \ominus_a \{f,g\}}^{CS}(a) - \mathtt{Deg}_{F \ominus_a \{e\}}^{CS}(a) = 0.4906 - 0.25530 = 0.2353$. For the other arguments, we have $\mathtt{Imp}_F^{CS}(\{a\}, a) = 0$, $\mathtt{Imp}_F^{CS}(\{b\}, a) = \mathtt{Imp}_F^{CS}(\{d\}, a) = -0.49$, $\mathtt{Imp}_F^{CS}(\{c\}, a) = 0.2353$ and $\mathtt{Imp}_F^{CS}(\{f\}, a) = \mathtt{Imp}_F^{CS}(\{g\}, a) = -0.1108$.
Thus, we have $\mathtt{I}_{CS}^{+}(a) = \{c, e\}$, $\mathtt{I}_{CS}^{-}(a) = \{b, d, f, g\}$ and $\mathtt{I}_{CS}^{=}(a) = \{a\}$.*

4 Balanced impact property

The definition of a new gradual semantics is often coupled with an axiomatic evaluation [1, 4]. Such axioms are mainly used to better understand the behaviour of gradual semantics in specific situations. The role and impact of an argument/attack are also discussed. Such axioms have the aim to answer questions like: Is an attack between two arguments killing (cf. Killing property [1]) or just weakening (cf. Weakening property [1]) the target of the attack? In addition, two semantics can both consider that an attack weakens its target (and then both satisfy the Weakening property) but with different levels of weakening. Unfortunately, this distinction cannot be captured with such axioms.

For example, computing the impact of b and c on a in the three AFs visualised in Fig. 2 with the h-categorizer semantics shows that their impact on a is less important when they attack together ($\mathtt{Imp}_{\mathrm{F}_3}^{\mathrm{Cat}}(\{b, c\}, a) = -0.667$) than when they attack it separately ($\mathtt{Imp}_{\mathrm{F}_1}^{\mathrm{Cat}}(\{b\}, a) + \mathtt{Imp}_{\mathrm{F}_2}^{\mathrm{Cat}}(\{c\}, a) = -0.5 + -0.5 = -1$).
Conversely, for the counting semantics, both return the same result:
$\mathtt{Imp}_{\mathrm{F}_3}^{\mathrm{CS}}(\{b, c\}, a) = -0.98 = -0.49 + -0.49 = \mathtt{Imp}_{\mathrm{F}_1}^{\mathrm{CS}}(\{b\}, a) + \mathtt{Imp}_{\mathrm{F}_2}^{\mathrm{CS}}(\{c\}, a)$.

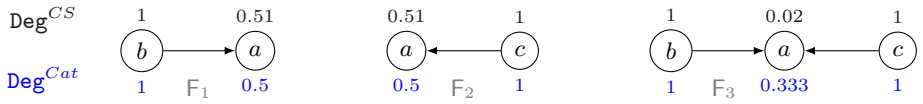


Fig. 2. Three argumentation frameworks F_1, F_2, F_3 showing the difference of impact among the counting semantics and the h-categorizer semantics.

To capture this idea, we define a new property, called Balanced Impact (BI), which states that the sum of the impact of two arguments alone on an argument y should be equal to the impact of these two arguments together on y.

Property 1 (Balanced Impact (BI)). A gradual semantics $\mathcal{S}$ satisfies Balanced Impact if and only if for any $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ and $x, y, z \in \mathcal{A}$,

$$\mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\{x\}, y) + \mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\{z\}, y) = \mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\{x, z\}, y)$$

Let us check which semantics (among CS and Cat) satisfies Balanced Impact.

Proposition 2. *The counting semantics satisfies Balanced Impact.*

Proposition 3. *The h-categorizer semantics does not satisfy Balanced Impact.*

Thus, this property allows to distinguish the semantics which distribute the impact of the arguments on another in a balanced way. Interestingly, this balance allows to go further because it is possible to compute the score of an argument w.r.t. a gradual semantics which satisfies BI from the impact of each single argument in the AF on this argument. Indeed, as explained in Section 3.1, the score of an argument y depends on the impact of all the arguments in the AF ($\mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\mathcal{A}, y)$), but thanks to the balanced impact property, we can split $\mathtt{Imp}_{\mathrm{F}}^{\mathcal{S}}(\mathcal{A}, y)$ into the impact of each individual argument in the AF. Let us first formally define it for the acyclic argumentation frameworks.

Definition 10. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an acyclic AF and $y \in \mathcal{A}$. Let $\mathcal{S}$ be a gradual semantics which satisfies BI. The score of y can be defined as follows:*

$$\mathtt{Deg}_{\mathsf{F}}^{\mathcal{S}}(y) = 1 + \sum_{x \in \mathcal{A}} \mathtt{Imp}_{\mathsf{F}}^{\mathcal{S}}(\{x\}, y)$$

Example 2 *Let us compute the score of a in the AF visualised in Fig. 1 using the impact of each single argument when CS is used.*

$$\begin{aligned}\mathtt{Deg}_{\mathsf{F}}^{CS}(a) &= 1 + (\mathtt{Imp}_{F}^{CS}(\{a\}, a) + \mathtt{Imp}_{F}^{CS}(\{b\}, a) + \mathtt{Imp}_{F}^{CS}(\{c\}, a) + \mathtt{Imp}_{F}^{CS}(\{d\}, a) \\ &\quad + \mathtt{Imp}_{F}^{CS}(\{e\}, a) + \mathtt{Imp}_{F}^{CS}(\{f\}, a) + \mathtt{Imp}_{F}^{CS}(\{g\}, a)) \\ &= 1 + (0 - 0.49 + 0.2353 - 0.49 + 0.2353 - 0.1108 - 0.1108) = 0.2691\end{aligned}$$

Algorithm 1: Transformation function ACY

Data: $\mathsf{F} = \langle \mathcal{A} = \{x_1, \ldots, x_n\}, \mathcal{R} \rangle$ and $x_1 \in \mathcal{A}$ the targeted argument.
Result: $\mathsf{F}' = \langle \mathcal{A}', \mathcal{R}' \rangle$ the infinite acyclic AF of F
$C = \{x_1\}$; $\mathcal{A}' = \{x_1^0\}$; $\mathcal{R}' = \emptyset$ // *x_1^0 is called the universal sink vertex of* F'
for *every argument* x_i *in* C **do**
  $C = C \backslash \{x_i\}$
  $m_1 \leftarrow$ maximum value of m among $x_i^m \in \mathcal{A}'$
  for *every argument* x_j *in* $\mathcal{R}_1^-(x_i)$ **do**
    $C = C \cup \{x_j\}$
    if $x_j^0 \notin \mathcal{A}'$ **then**
      $\mathcal{A}' = \mathcal{A}' \cup x_j^0$; $\mathcal{R}' = \mathcal{R}' \cup (x_j^0, x_i^{m_1})$
    else
      $m_2 \leftarrow$ (maximum value of m among $x_j^m \in \mathcal{A}'$) + 1
      $\mathcal{A}' = \mathcal{A}' \cup x_j^{m_2}$; $\mathcal{R}' = \mathcal{R}' \cup (x_j^{m_2}, x_i^{m_1})$

In order to generalise this definition for any AF, a preprocessing step is required. Indeed, deleting an argument in a cycle removes as well its impact as the ones of other arguments in the cycle. As the method works for acyclic AFs, we propose to transform a cyclic AF into an infinite acyclic AF[1] focused on a given argument a. Thus, as visualised in Fig. 3, we obtain a tree-shaped AF where the root node is a itself, its parent nodes are its direct attackers, the parent nodes of its parent nodes are its direct defenders, and so on. Algorithm 1 details the transformation mechanism called ACY.

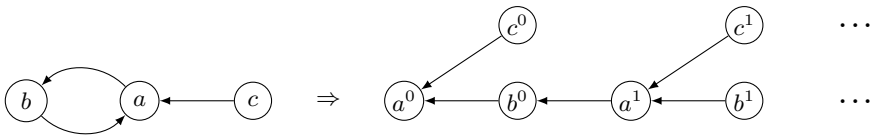


Fig. 3. Cyclic AF transformed into its infinite acyclic AF

We can now use the transformation of an AF, denoted by F, to define the impact of any argument x on a given argument y as the sum of the impact of all the sub-arguments of x $(x^0, x^1, \ldots)$ on y^0 (the universal sink vertex) in $\mathsf{ACY}_y(\mathsf{F})$.

Definition 11. *Let* $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ *be an AF with* $y \in \mathcal{A}$. *Let* $\mathsf{F}' = \mathsf{ACY}_y(\mathsf{F})$ *and* $\mathcal{X} = \{x^0, x^1, \ldots\}$ *be the sub-arguments of* $x \in \mathcal{A}$ *in* F'. *Let* $\mathcal{S}$ *be a gradual semantics which satisfies BI. The impact of* x *on* y *is 0 if* $\mathcal{X} = \emptyset$, *otherwise it is*

[1] From a computational point of view, the scores of each argument are computed using a fixed-point approach. If the function used in the gradual semantics converges, the number of iterations needed for convergence can also be used to define the maximal depth of the tree-shaped AF.

defined as follows:

$$\texttt{Imp}_{\mathsf{F}}^{\mathcal{S}}(\{x\}, y) = \sum_{x^i \in \mathcal{X}} \texttt{Imp}_{\mathsf{F}'}^{\mathcal{S}}(\{x^i\}, y^0)$$

This new definition of impact can then be used in Definition 10 to compute the score of a given argument.

Example 3 *By focusing on the AF visualised in Fig. 3, the impact of b on a is $\texttt{Imp}_{\mathsf{F}}^{CS}(\{b\}, a) = \texttt{Imp}_{\mathsf{ACY}_a(\mathsf{F})}^{CS}(\{b^0\}, a^0) + \texttt{Imp}_{\mathsf{ACY}_a(\mathsf{F})}^{CS}(\{b^1\}, a^0) + \cdots \simeq -0.63$. We also have $\texttt{Imp}_{\mathsf{F}}^{CS}(\{c\}, a) \simeq -0.63$ and $\texttt{Imp}_{\mathsf{F}}^{CS}(\{a\}, a) \simeq 0.3$.*
We obtain $\texttt{Deg}_{\mathsf{F}}^{CS}(a) \simeq 0.04 = 1 + 0.3 - 0.63 - 0.63 = 1 + \sum_{x \in \{a,b,c\}} \texttt{Imp}_{\mathsf{F}}^{CS}(\{x\}, a)$.

5 Interpretability of gradual semantics

One of the goals of interpretability for gradual semantics is to identify the elements which have an impact on the score assigned by the selected gradual semantics on each argument. Definition 9 allows to assess whether an argument has a positive, negative or neutral impact on the acceptability degree of an argument. It allows to answer questions about the impact of certain arguments on the others, like in the following example about the AF (F) in Fig. 1:

Q: *Which arguments have a positive impact on a in F when CS is used?*
A: c and e have a positive impact on a. $\texttt{I}^{+}_{\text{CS}}(a) = \{c, e\}$

Through the impact values (see Definition 8), it is possible to provide, for each argument, a ranking between the arguments from the most positive to the most negative impacting ones w.r.t. a given gradual semantics.

Definition 12 (Impact ranking). *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $\mathcal{S}$ be a gradual semantics. The **impact ranking** $\succeq_y^{\mathcal{S}}$ on $\mathcal{A}$ with respect to $y \in \mathcal{A}$ is defined such that $\forall x, z \in \mathcal{A}$, $x \succeq_y^{\mathcal{S}} z$ iff $\texttt{Imp}_F^{\mathcal{S}}(\{x\}, y) \geq \texttt{Imp}_F^{\mathcal{S}}(\{z\}, y)$.*

This ranking allows us to select, for each argument, its most positive and negative impacting arguments, if they exist.

Definition 13. *Let $\mathsf{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an AF and $\mathcal{S}$ be a gradual semantics. The **most positive** (resp. **negative**) **impacting arguments** on the acceptability degree of $y \in \mathcal{A}$ are defined as follows:*

$$PI_F^{\mathcal{S}}(y) = argmax_{x \in \texttt{I}^{+}_{\mathcal{S}}(y)} |\{z \in \texttt{I}^{+}_{\mathcal{S}}(y) \mid x \succeq_y^{\mathcal{S}} z\}|$$
$$NI_F^{\mathcal{S}}(y) = argmax_{x \in \texttt{I}^{-}_{\mathcal{S}}(y)} |\{z \in \texttt{I}^{-}_{\mathcal{S}}(y) \mid z \succeq_y^{\mathcal{S}} x\}|$$

Example 4 *Let us consider the AF depicted in Fig. 1. The impact ranking of argument a, when CS is used, is $c \simeq_a^{CS} e \succ_a^{CS} a \succ_a^{CS} f \simeq_a^{CS} g \succ_a^{CS} b \simeq_a^{CS} d$. Consequently, we have $PI_F^{CS}(a) = \{c, e\}$ and $NI_F^{CS}(a) = \{b, d\}$.*

In addition to providing a better understanding of the scores assigned to each argument, this information can also be used to develop strategies during a debate. For example, if someone wants to defend a point of view (i.e., increase the degree of acceptability of an argument in a debate), she can identify the argument(s) with the most negative impact and therefore look for solutions to attack them by introducing some counter-arguments.

6 Related Work

Interpretability has already been studied in the context of extension-based semantics in formal argumentation. Fan and Toni [11] first studied how to give explanations for arguments that are acceptable w.r.t. the admissible semantics in terms of arguments defending them, before formalising explanations for arguments that are not acceptable w.r.t. the admissible semantics by using a dispute tree [12]. Although the extension-based semantics and the gradual semantics share the same goal (i.e., evaluating the arguments), the two approaches are different (see the discussion in [7] for more details). Consequently, the investigation of the notion of interpretability for these two families of semantics also differs.

Concerning the gradual semantics, Amgoud and al. [2] have introduced the concept of contribution measure for evaluating the intensity of each attack in an argumentation graph. The Shapley value is used as contribution measure. However, only a specific family of gradual semantics is considered (i.e., the ones which satisfy the syntax-independent and monotonicity properties like the h-categorizer semantics). Moreover, unlike our method which checks the impact of all arguments in the framework, their method only measures the contribution of direct attacks on an argument which is coherent for the family of semantics studied in this work, but it is not necessarily the case for all existing semantics.

7 Conclusion

In this paper, we have presented a formal framework to interpret the results of gradual semantics in abstract argumentation. More precisely, we have considered the h-categorizer and the counting semantics, and we have formally studied the notion of impact of an argument with respect to the acceptability degree of another argument in the framework both for cyclic and acyclic frameworks. The impact of arguments on the acceptability degree of the other arguments is then employed to interpret the rationale behind the resulting ranking, and to provide a further understanding of the reasons why attacking one argument rather than another may be a strategically better choice.

Two main open issues will be considered as future work: first, in this paper we do not consider the *support* relation [8] between arguments but we aim to extend our formal framework to capture this relation too given its importance in many practical applications, and second, we plan to extend our analysis to the other gradual semantics proposed in the literature to provide a complete overview of the properties of the impact notion over such semantics.

8 Acknowledgements

This work benefited from the support of the project DGA RAPID CONFIRMA.

References

1. Amgoud, L., Ben-Naim, J.: Axiomatic foundations of acceptability semantics. In: Proc. of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR'16). pp. 2–11 (2016)
2. Amgoud, L., Ben-Naim, J., Vesic, S.: Measuring the intensity of attacks in argumentation graphs with shapley value. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17). pp. 63–69 (2017)
3. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G.R., Thimm, M., Villata, S.: Towards artificial argumentation. AI Magazine **38**(3), 25–36 (2017), https://www.aaai.org/ojs/index.php/aimagazine/article/view/2704
4. Baroni, P., Rago, A., Toni, F.: From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. Int. J. Approx. Reasoning **105**, 252–286 (2019). https://doi.org/10.1016/j.ijar.2018.11.019, https://doi.org/10.1016/j.ijar.2018.11.019
5. Besnard, P., Hunter, A.: A logic-based theory of deductive arguments. Artificial Intelligence **128**(1-2), 203–235 (2001)
6. Bonzon, E., Delobelle, J., Konieczny, S., Maudet, N.: A Comparative Study of Ranking-based Semantics for Abstract Argumentation. In: Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI'16). pp. 914–920 (2016)
7. Bonzon, E., Delobelle, J., Konieczny, S., Maudet, N.: Combining extension-based semantics and ranking-based semantics for abstract argumentation. In: Proc. of the 16th International Conference on Principles of Knowledge Representation and Reasoning (KR'18). pp. 118–127 (2018)
8. Cayrol, C., Lagasquie-Schiex, M.: Bipolarity in argumentation graphs: Towards a better understanding. Int. J. Approx. Reasoning **54**(7), 876–899 (2013). https://doi.org/10.1016/j.ijar.2013.03.001, https://doi.org/10.1016/j.ijar.2013.03.001
9. Cyras, K., Birch, D., Guo, Y., Toni, F., Dulay, R., Turvey, S., Greenberg, D., Hapuarachchi, T.: Explanations by arbitrated argumentative dispute. Expert Syst. Appl. **127**, 141–156 (2019). https://doi.org/10.1016/j.eswa.2019.03.012, https://doi.org/10.1016/j.eswa.2019.03.012
10. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–358 (1995)
11. Fan, X., Toni, F.: On computing explanations in argumentation. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 1496–1502 (2015)
12. Fan, X., Toni, F.: On explanations for non-acceptable arguments. In: Theory and Applications of Formal Argumentation - Third International Workshop, TAFA 2015, Buenos Aires, Argentina, July 25-26, 2015, Revised Selected Papers. pp. 112–127 (2015)
13. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
14. Mittelstadt, B.D., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019. pp. 279–288. ACM (2019). https://doi.org/10.1145/3287560.3287574, https://doi.org/10.1145/3287560.3287574

15. Pu, F., Luo, J., Zhang, Y., Luo, G.: Argument ranking with categoriser function. In: Proc. of the 7th International Conference on Knowledge Science, Engineering and Management, (KSEM'14). pp. 290–301 (2014)
16. Pu, F., Luo, J., Zhang, Y., Luo, G.: Attacker and defender counting approach for abstract argumentation. In: Proc. of the 37th Annual Meeting of the Cognitive Science Society, (CogSci'15) (2015)
17. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: Fantastic explanations and how to find them. In: Lang, J. (ed.) Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. pp. 1949–1955. ijcai.org (2018). https://doi.org/10.24963/ijcai.2018/269, https://doi.org/10.24963/ijcai.2018/269