Measuring the Intensity of Attacks in Argumentation Graphs with Shapley Value

Leila Amgoud, Jonathan Ben-Naim, Srdjan Vesic

Measuring the Intensity of Attacks in Argumentation Graphs with Shapley Value

Leila Amgoud[1], **Jonathan Ben-Naim**[1], **Srdjan Vesic**[2]
[1] IRIT, CNRS - Université de Toulouse, France
[2] CRIL, CNRS - Université d'Artois, France
amgoud@irit.fr, bennaim@irit.fr, vesic@cril.fr

Abstract

In an argumentation setting, a semantics evaluates the overall acceptability of arguments. Consequently, it reveals the global *loss* incurred by each argument due to attacks. However, it does not say anything on the *contribution* of each attack to that loss. This paper introduces the novel concept of *contribution measure* for evaluating those contributions. It starts by defining a set of axioms that a reasonable measure would satisfy, then shows that the Shapley value is the unique measure that satisfies them. Finally, it investigates the properties of the latter under some existing semantics.

1 Introduction

An argumentation framework is a reasoning model based on the justification of claims by arguments. It is made of a *graph* and a *semantics*. The nodes of the graph are arguments, each of which is assigned a basic strength, and the arcs are attacks between pairs of arguments. The semantics is a function assigning to each argument of the graph a value representing its *overall strength* or *acceptability degree*. See [Simari and Rahwan, 2009] for an overview on argumentation in AI.

Recently, Amgoud et al. [2016; 2017] have argued that the acceptability degree of an argument should be equal to the basic strength of the argument, if the latter is not attacked. Otherwise, the argument is weakened by its attackers and thus looses weight, leading to an acceptability degree lower than the basic strength. Hence, from the outcome of a semantics, it is possible to compute the *global loss* undergone by each argument because of its attackers. It is the difference between the basic strength of the argument and its acceptability degree. However, it is not possible to say anything on the *contribution* of each attack to that loss. That contribution represents, in some sense, the *intensity* of the attack. The greater the contribution of an attack, the more harmful the attack.

Information on attacks' contributions is very useful since it allows a better understanding of the impact of each attack. Namely, it allows detecting *worthless* attacks (i.e., attacks that do not have any impact on the target), and *redundant* ones (i.e., attacks that lead to the same loss for their target).

Attacks' contributions allow also to rank order the attackers of each argument of a graph from the most to the least harmful ones. This ranking is very useful, especially in persuasion dialogues where agents have to choose the best counter-attack in order to win a dialogue. Assume a dialogue between two agents who exchange arguments in order to persuade each other. At each step of the dialogue, an agent presents a new argument attacking one of those uttered by the other party. For that purpose, the agent should choose i) which argument of the opponent to attack, and ii) with which argument. A reasonable strategy consists of targeting an argument that is very harmful for the agent's arguments.

This paper studies for the first time the question of measuring the contribution of each attack to the global loss of its target. It introduces the novel concept of *contribution measure*, which takes as input an argumentation framework, and returns as output a weight for each attack, representing the contribution of the attack. It considers a broad range of semantics including extension-based ones [Dung, 1995]. The paper starts by defining a set of axioms that a reasonable measure should satisfy. Then, it provides a characterization theorem, which states that Shapley value [Shapley, 1953] is the *unique* measure that satisfies the axioms. Finally, it investigates properties of that measure under extension semantics, and $\mathbf{h}$-categorizer semantics defined by Besnard and Hunter [2001].

The paper is structured as follows: Section 2 defines argumentation frameworks. Section 3 shows examples of semantics covered by the study. Section 4 introduces contribution measures as well as the set of axioms that they would satisfy. Section 5 provides our characterization result. Section 6 instantiates the Shapely measure with some existing semantics.

2 Argumentation Frameworks

An argumentation framework is made of an *argumentation graph* and an acceptability *semantics*. Throughout the paper, we focus on argumentation graphs whose nodes are arguments and arcs are attacks between arguments. An argument is an abstract entity whose internal structure is not specified, however, it has an initial value representing its *basic strength*. The latter may represent different issues, like certainty degree of argument's premises [Benferhat *et al.*, 1993], trustworthiness in argument's source [da Costa Pereira *et al.*, 2011], etc. Before defining formally argumentation frameworks, we start

by introducing the useful notion of *weighting*.

Definition 1 (Weighting) *A* weighting *on a set X is a function from X to $[0, 1]$.*

Let $\mathtt{Arg}$ be an infinite set of all possible arguments. An argumentation graph is defined as follows:

Definition 2 (Argumentation Graph) *An* argumentation graph *is an ordered tuple $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$, where $\mathcal{A}$ is a finite subset of $\mathtt{Arg}$, w is a weighting on $\mathcal{A}$, and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. Let $\mathtt{AG}$ be the universe of all argumentation graphs built on $\mathtt{Arg}$.*

For two arguments a, b of an argumentation graph $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$, $w(a)$ represents the *basic strength* of the argument a, and $(a, b) \in \mathcal{R}$ means that argument a *attacks argument* b.

Notations: Let $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$ be an argumentation graph. Elements of $\mathcal{R}$ will be denoted by $r_1, r_2, \ldots, r_n$. Note that $\mathcal{R}$ is finite since $\mathcal{A}$ is finite. $\mathtt{Sc}_{\mathbf{A}}(.)$ and $\mathtt{Tr}_{\mathbf{A}}(.)$ are two functions, which return respectively the *source* a and the *target* b of an attack $(a, b) \in \mathcal{R}$. $\mathtt{Att}_{\mathbf{A}}(.)$ is a function, which returns all the *attacks* on an argument (i.e. for $a \in \mathcal{A}$, $\mathtt{Att}_{\mathbf{A}}(a) = \{r \in \mathcal{R} \mid \mathtt{Tr}_{\mathbf{A}}(r) = a\}$). Let $X \subseteq \mathcal{R}$, $\mathbf{A} \ominus X = \langle \mathcal{A}, w, \mathcal{R} \setminus X \rangle$.

A semantics is a function assigning to every argument in an argumentation graph an *acceptability degree*. The greater this degree, the more acceptable the argument. The degree is between 0 and the basic strength of the argument. The idea is: If an argument is not attacked, then it keeps its full basic strength, otherwise it may lose weight if its attackers are sufficiently strong. In [Amgoud *et al.*, 2017], we provided an axiomatic justification for this definition of semantics.

Definition 3 (Semantics) *A* semantics *is a function $\mathbf{S}$ transforming any argumentation graph $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$ into a weighting on $\mathcal{A}$ s.t. for any $a \in \mathcal{A}$, if $\mathtt{Att}_{\mathbf{A}}(a) = \emptyset$, then $\mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a) = w(a)$, else $\mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a) \in [0, w(a)]$. $\mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a)$ is the image of a by $\mathbf{S}(\mathbf{A})$, and is called* acceptability degree *of a.*

Throughout the paper, the semantics is left *unspecified*. However, without loss of generality, it satisfies the very basic *syntax-independence*, proposed in [Amgoud *et al.*, 2017], and *monotonicity* properties. The former ensures that the acceptability degree of an argument is independent of the argument's identity. Before defining formally the property, let us first recall the notion of isomorphism of graphs.

Definition 4 (Isomorphism) *Let $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle, \mathbf{A}' = \langle \mathcal{A}', w', \mathcal{R}' \rangle \in \mathtt{AG}$. An* isomorphism *from $\mathbf{A}$ to $\mathbf{A}'$ is a bijective function f from $\mathcal{A}$ to $\mathcal{A}'$ such that the following hold:*

- $\forall\, a \in \mathcal{A}$, $w(a) = w'(f(a))$,
- $\forall\, a, b \in \mathcal{A}$, $(a, b) \in \mathcal{R}$ *iff* $(f(a), f(b)) \in \mathcal{R}'$,

Definition 5 (Syntax-Indep.) *A semantics $\mathbf{S}$ is* syntax-independent *iff for all $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle, \mathbf{A}' = \langle \mathcal{A}', w', \mathcal{R}' \rangle \in \mathtt{AG}$, for any isomorphism f from $\mathbf{A}$ to $\mathbf{A}'$, the following holds: for any $a \in \mathcal{A}, \mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a) = \mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}'}(f(a))$.*

The monotonicity property ensures that attacks cannot be *beneficial* for arguments. It is worth pointing out that this property is different from the monotony axiom from [Amgoud and Ben-Naim, 2016; Amgoud *et al.*, 2017], which states the following: if the attackers of an argument a are also attackers of b, then a is at least as acceptable as b. This axiom assumes that a and b are in the *same graph*, thus the attackers of both arguments have fixed acceptability degrees. Our monotonicity axiom goes one step further by assuming that a and b are in *different graphs*. Hence, the acceptability degrees of their attackers may vary from one graph to the other. To sum up, it is possible for a semantics to satisfy one of the two forms of monotony and violates the other.

Definition 6 (Monotonicity) *A semantics $\mathbf{S}$ is* monotone *iff for any argumentation graph $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $\forall a \in \mathcal{A}$, $\forall X \subseteq \mathtt{Att}_{\mathbf{A}}(a)$, it holds that $\mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a) \leq \mathtt{Deg}_{\mathbf{S}}^{\mathbf{A} \ominus X}(a)$.*

Notation: Let $\mathtt{Sem}$ be the universe of all syntax-independent and monotone semantics defined on $\mathtt{AG}$.

When the acceptability degree of an argument is lower than its basic strength, the argument has lost strength due to its attackers. The total amount of that loss is defined as follows.

Definition 7 (Loss) *Let $\mathbf{S} \in \mathtt{Sem}$, $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, and $a \in \mathcal{A}$. The* loss *of a is $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a) = w(a) - \mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a)$.*

From the definitions, we get the following obvious results.

Property 1 *Let $\mathbf{S} \in \mathtt{Sem}$, $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $a \in \mathcal{A}$.*

- $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a) \in [0, 1]$.
- *If $\mathtt{Att}_{\mathbf{A}}(a) = \emptyset$, then $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a) = 0$.*

The following property follows also straightforwardly from the monotonicity of semantics of the set $\mathtt{Sem}$.

Property 2 *Let $\mathbf{S} \in \mathtt{Sem}$, $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, and $a \in \mathcal{A}$. For any $X \subseteq \mathtt{Att}_{\mathbf{A}}(a)$, $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a) \geq \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus X}(a)$.*

3 Examples of Covered Semantics

The previous section introduced the set $\mathtt{Sem}$ of semantics we consider in this paper, namely syntax-independent and monotone semantics. Before defining contribution measures that share the loss of an argument under such semantics among the argument's attacks, we need first to show that the set $\mathtt{Sem}$ is *not empty*. In other words, we should prove that there exist semantics that satisfy the two above properties. Hopefully, this is the case of at least Dung's extension semantics [1995] and Besnard and Hunter's **h**-Categorizer semantics [2001].

In his seminal paper, Dung assumed all arguments have the same basic strength. Thus, we consider argumentation graphs whose arguments have each the basic strength 1. An extension semantics starts by computing subsets of arguments (called *extensions*), which are *conflict free* (i.e., they do not contain two arguments that attack each others). Furthermore, they *defend* their elements (i.e., they attack any argument attacking one of their elements). Let $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$ be an argumentation graph such that for any $a \in \mathcal{A}$, $w(a) = 1$, and let $\mathcal{E} \subseteq \mathcal{A}$ be a conflict-free set.

- $\mathcal{E}$ is a *complete* extension iff it defends all its elements and contains any argument it defends.
- $\mathcal{E}$ is a *preferred* extension iff it is a maximal (w.r.t. set $\subseteq$) complete extension.
- $\mathcal{E}$ is a *stable* extension iff it attacks any $a \in \mathcal{A} \setminus \mathcal{E}$.

- $\mathcal{E}$ is a *grounded extension* iff it is a minimal (w.r.t. set $\subseteq$) complete extension.

Let $\texttt{Ext}_x(\mathbf{A})$ denote the set of all extensions of $\mathbf{A}$ under semantics x, where $x \in \{\texttt{st}, \texttt{pr}, \texttt{gr}, \texttt{co}\}$ and st (resp. pr, gr, co) stands for stable (resp. preferred, grounded, complete) semantics. Once extensions are computed, acceptability degrees are assigned to arguments. In what follows, we slightly modify the definition given by Amgoud and Ben-Naim [2016], in particular the case where a graph has no stable extensions. Instead of assigning the degree 0.3 to each argument of the graph, we assume that arguments keep their basic strengths. The idea is that there is no reason for losing strength. Formally, for any $a \in \mathcal{A}$, if $\texttt{Ext}_x(\mathbf{A}) = \emptyset$, then $\texttt{Deg}_x^{\mathbf{A}}(a) = w(a) = 1$, otherwise:

- $\texttt{Deg}_x^{\mathbf{A}}(a) = 1$ iff $a \in \bigcap_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{A})} \mathcal{E}$.
- $\texttt{Deg}_x^{\mathbf{A}}(a) = 0.5$ iff $\exists \mathcal{E}, \mathcal{E}' \in \texttt{Ext}_x(\mathbf{A})$ s.t $a \in \mathcal{E}$, $a \notin \mathcal{E}'$.
- $\texttt{Deg}_x^{\mathbf{A}}(a) = 0.3$ iff $a \notin \bigcup_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{A})} \mathcal{E}$ and $\nexists \mathcal{E} \in \texttt{Ext}_x(\mathbf{A})$ s.t $\exists b \in \mathcal{E}$ and $(b, a) \in \mathcal{R}$.
- $\texttt{Deg}_x^{\mathbf{A}}(a) = 0$ iff $a \notin \bigcup_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{A})} \mathcal{E}$ and $\exists \mathcal{E} \in \texttt{Ext}_x(\mathbf{A})$ s.t $\exists b \in \mathcal{E}$ and $(b, a) \in \mathcal{R}$.

We show next that the four semantics are part of the set Sem. Indeed, they are in accordance with Definition 3 of semantics, and are all syntax-independent and monotone.

Proposition 1 *It holds that* $\{\texttt{st}, \texttt{gr}, \texttt{co}, \texttt{pr}\} \subseteq \texttt{Sem}$.

Besnard and Hunter [2001] proposed **h**-*categorizer* semantics for evaluating arguments in acyclic graphs. This semantics was extended by Pu et al. [2014] to any graph structure. It considers as input an argumentation graph whose arguments have all the same basic strength 1, and returns an acceptability degree in the interval $(0, 1]$ to each argument a as follows:

$$\texttt{Deg}_{\mathbf{h}}^{\mathbf{A}}(a) = \frac{1}{1 + \sum_{b:(b,a)\in\mathcal{R}} \texttt{Deg}_{\mathbf{h}}^{\mathbf{A}}(b)}$$

with $\sum_{b:b\mathcal{R}a} \texttt{Deg}_{\mathbf{h}}^{\mathbf{A}}(b) = 0$ if $\texttt{Att}_{\mathbf{A}}(a) = \emptyset$.

It was shown in [Amgoud and Ben-Naim, 2016] that **h**-*categorizer* is in accordance with Definition 3, and is syntax-independent. We implemented this semantics, and run several experiments, which all show that the semantics is monotone.

Conjecture 1 **h**-*categorizer semantics is monotone. It is thus a member of the set* Sem.

The two previous results show that the set of semantics investigated in the paper is not empty ($\texttt{Sem} \neq \emptyset$) and covers the main existing semantics.

4 Contribution Measures

A contribution measure takes as input an argumentation framework (an argumentation graph and a semantics, which evaluates the arguments of the graph), and assigns to each attack in the graph a value between 0 and 1. This value represents the *contribution* of the attack to the loss undergone by its target. In other words, for each attacked argument in the graph, the measure divides the total loss of the argument among the attacks received by the argument.

Definition 8 (Contribution Measure) *A* contribution measure *is a function* $\mathbf{C}$ *on* $\texttt{Sem} \times \texttt{AG}$ *such that,* $\forall \mathbf{S} \in \texttt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \texttt{AG}$, $\mathbf{C}(\mathbf{S}, \mathbf{A})$ *is a weighting* f *on* $\mathcal{R}$ *satisfying the following condition:* $\forall a \in \mathcal{A}$ *such that* $\texttt{Att}_{\mathbf{A}}(a) \neq \emptyset$,

$$\sum_{r \in \texttt{Att}_{\mathbf{A}}(a)} f(r) = \texttt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a). \quad (1)$$

$f(r)$ *is called the* contribution *of* r *to the loss of* $\texttt{Tr}(r)$. *Let* $\texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r)$ *denote* $f(r)$, *i.e. the image of* r *by* $\mathbf{C}(\mathbf{S}, \mathbf{A})$.

Equation 1 provides an *efficiency* condition, which ensures that the entire loss of an argument is divided among the argument's attacks. This leads to the following obvious property.

Property 3 *For any* $\mathbf{S} \in \texttt{Sem}$, *any* $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \texttt{AG}$, *any contribution measure* $\mathbf{C}$ *on* $(\mathbf{S}, \mathbf{A})$, *any* $a \in \mathcal{A}$ *such that* $\texttt{Att}_{\mathbf{A}}(a) \neq \emptyset$, *the following two properties hold:*

- *If* $\texttt{Att}_{\mathbf{A}}(a) = \{r\}$, *then* $\texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r) = \texttt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a)$.
- *If* $\texttt{Loss}_{\mathbf{S}}^{\mathbf{A}}(a) = 0$, *then* $\forall r \in \texttt{Att}_{\mathbf{A}}(a)$, $\texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r) = 0$.

In addition to the efficiency condition, the division of a loss should be both reasonable and fair. In what follows, we define properties (called axioms in the paper) that describe what a reasonable and fair measure is.

The first axiom guarantees syntax-independence. It states that the *identities* of arguments cannot change the outcome of a contribution measure.

Axiom 1 (Anonymity) *A contribution measure* $\mathbf{C}$ *satisfies* anonymity *iff,* $\forall \mathbf{S} \in \texttt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \texttt{AG}$ *and* $\forall \mathbf{A}' = \langle \mathcal{A}', w', \mathcal{R}' \rangle \in \texttt{AG}$, *for any isomorphism* f *from* $\mathbf{A}$ *to* $\mathbf{A}'$, *it holds that* $\forall\, r \in \mathcal{R}$,

$$\texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r) = \texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}'}((f(\texttt{Sc}(r)), f(\texttt{Tr}(r)))).$$

A fair division of an argument's loss among the argument's attacks should take into account the effective impact of each attack. The second axiom concerns worthless attacks. It states that if an attack is not harmful to its target, then its contribution should be 0. An important question then is: what is a worthless, called here *dummy*, attack? A possible definition is: an attack whose source has an acceptability degree 0. Consider the following example.

Example 1 *Consider the semantics* $\mathbf{S}_1$ *defined as follows: for any argumentation graph* $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$, *for any* $a \in \mathcal{A}$,

$$\texttt{Deg}_{\mathbf{S}^1}^{\mathbf{A}}(a) = \begin{cases} w(a) & \text{if } \texttt{Att}_{\mathbf{A}}(a) = \emptyset \\ 0 & \text{else} \end{cases}.$$

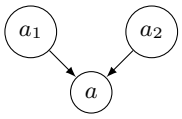
This semantics is clearly syntax-independent and monotone, thus it belongs to the set Sem. *Consider now the argumentation graph* $\mathbf{A}_1 = \langle \mathcal{A}_1, w_1, \mathcal{R}_1 \rangle$, *where* $\mathcal{A}_1 = \{a, b\}$, $\mathcal{R}_1 = \{(b, a)\}$, $w_1(a) = 1$ *and* $w_1(b) = 0$. *Clearly,* $\texttt{Deg}_{\mathbf{S}^1}^{\mathbf{A}_1}(a) = 0$. *Thus,* $\texttt{Loss}_{\mathbf{S}^1}^{\mathbf{A}_1}(a) = 1$ *meaning that* a *loses all its basic strength. This loss is due to the attack from* b *even if* $\texttt{Deg}_{\mathbf{S}^1}^{\mathbf{A}_1}(b) = 0$ *(since* $\texttt{Att}_{\mathbf{A}_1}(b) = \emptyset$ *and* $w_1(b) = 0$*). Hence,* (b, a) *is certainly not a dummy attack.*

It is worth noticing that a contribution measure should be rational whatever its input. Namely, it should be able to perform fair division whatever the semantics that is considered, even very basic ones like $\mathbf{S}_1$. From the example, it follows that the above definition of dummy attack is not suitable for semantics $\mathbf{S}_1$, and more generally for semantics that do not take into account the acceptability degrees of attackers. We will see in Section 6 that it is not suitable for some extension semantics as well. To sum up, a good definition of dummy attack should be independent of acceptability degrees of arguments. A natural candidate is a definition that is based on the *marginal contributions* of attacks. The marginal contribution of an attack is the difference between the loss undergone by the target and the loss of the target when the attack is removed from the argumentation graph. A dummy attack is an attack whose marginal contribution is 0. Let us illustrate this idea with an example.

Example 1 (Cont) Recall that $\texttt{Loss}_{\mathbf{S}_1}^{\mathbf{A}_1}(a) = 1$. Consider the argumentation graph $\mathbf{A}_1' = \mathbf{A}_1 \ominus \{(b,a)\}$. In $\mathbf{A}_1'$, a is not attacked, then $\texttt{Deg}_{\mathbf{S}^1}^{\mathbf{A}_1'}(a) = 1$ and $\texttt{Loss}_{\mathbf{S}^1}^{\mathbf{A}_1'}(a) = 0$. The attack (b,a) is not dummy since $\texttt{Loss}_{\mathbf{S}^1}^{\mathbf{A}_1}(a) - \texttt{Loss}_{\mathbf{S}^1}^{\mathbf{A}_1'}(a) \neq 0$.

Unfortunately, this new definition is still not fully satisfactory as shown in the following example.

Example 2 *Consider the argumentation graph $\mathbf{A}_2$ depicted below, where each argument has the basic strength 1.*



Let $r_1 = (a_1, a)$ and $r_2 = (a_2, a)$. Consider stable semantics (st) proposed by Dung [1995]. The graph $\mathbf{A}_2$ has one stable extension $\{a_1, a_2\}$. Thus, $\texttt{Deg}_{\texttt{st}}^{\mathbf{A}_2}(a) = 0$, and $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2}(a) = 1$. In order to check whether r_1 is a dummy attack of a, we consider the graph $\mathbf{A}_2' = \mathbf{A}_2 \ominus \{r_1\}$. Clearly, $\texttt{Deg}_{\texttt{st}}^{\mathbf{A}_2'}(a) = 0$ (since $\mathbf{A}_2'$ has one stable extension $\{a_1, a_2\}$) and $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2'}(a) = 1$. Note that $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2}(a) = \texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2'}(a)$. Thus, r_1 is a dummy attack, and for any contribution measure $\mathbf{C}$, $\texttt{Ctr}_{\mathbf{C}}^{\texttt{st},\mathbf{A}_2}(r_1) = 0$. However, it is easy to check that r_2 is also dummy, and its contribution is 0. This violates the efficiency condition of Equation 1. Indeed, $\texttt{Ctr}_{\mathbf{C}}^{\texttt{st},\mathbf{A}_2}(r_1) + \texttt{Ctr}_{\mathbf{C}}^{\texttt{st},\mathbf{A}_2}(r_2) = 0$ while $\texttt{Loss}_{\mathbf{A}_2}^{\texttt{st}}(a) = 1$. Furthermore, this is not intuitive since the argument a has lost its entire basic strength because of its two attackers.

In order to avoid the previous problems, we propose to check the marginal contribution of an attack in the initial graph as well as in all sub-graphs of the initial graph where some target's attacks are removed.

Axiom 2 (Dummy) *A contribution measure satisfies* dummy *iff, $\forall \mathbf{S} \in \texttt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \texttt{AG}$, $\forall a \in \mathcal{A}$ s.t. $\texttt{Att}_{\mathbf{A}}(a) \neq \emptyset$, $\forall r \in \texttt{Att}_{\mathbf{A}}(a)$, if $\forall X \subseteq \texttt{Att}_{\mathbf{A}}(a) \setminus \{r\}$, $\texttt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus X}(a) = \texttt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r\})}(a)$, then $\texttt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r) = 0$. We say that r is a* dummy *attack.*

Example 2 (Cont) Even if one of the two attacks is sufficient to kill the argument a, none of them is dummy. Let us analyze r_1 (the same reasoning holds for r_2). We should check any $X \subseteq \texttt{Att}_{\mathbf{A}}(a) \setminus \{r_1\} = \{r_2\}$. There are two cases:

- $X = \emptyset$: $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2 \ominus X}(a) = 1$, $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2 \ominus (X \cup \{r_1\})}(a) = 1$.
- $X = \{r_2\}$: $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2 \ominus X}(a) = 1$, $\texttt{Loss}_{\texttt{st}}^{\mathbf{A}_2 \ominus (X \cup \{r_1\})}(a) = 0$.

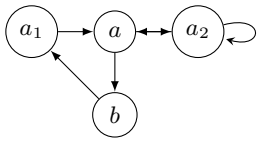
Note that when $X = \{r_2\}$, the argument a is not attacked in the graph $\mathbf{A}_2 \ominus (X \cup \{r_1\})$, and it is attacked only by a_1 in the graph $\mathbf{A}_2 \ominus X$. Thus, r_1 taken alone is harmful for a.

We show that a dummy attack cannot weaken its target even when it is the only attack received by its target.

Proposition 2 *Let $\mathbf{C}$ be a contribution measure, which satisfies Dummy. For any semantics $\mathbf{S} \in \texttt{Sem}$, any argumentation graph $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \texttt{AG}$, any $a \in \mathcal{A}$ s.t. $\texttt{Att}_{\mathbf{A}}(a) \neq \emptyset$, any $r \in \texttt{Att}_{\mathbf{A}}(a)$, if r is dummy, then $\texttt{Deg}_{\mathbf{S}}^{\mathbf{A}'}(a) = w(a)$ and $\texttt{Loss}_{\mathbf{S}}^{\mathbf{A}'}(a) = 0$, with $\mathbf{A}' = \mathbf{A} \ominus (\texttt{Att}_{\mathbf{A}}(a) \setminus \{r\})$.*

The next axiom defines when two attacks targeting the same argument should receive the *same contribution*. Such attacks are said to be *symmetric*. One might think that two attacks are symmetric if their sources have the same acceptability degree. However, the following example shows that this definition may lead to an unfair division of the loss.

Example 3 *Consider the argumentation graph $\mathbf{A}_3$ depicted below, where each argument has a basic strength equal to 1.*

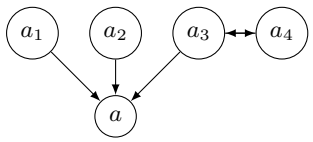


Under preferred semantics, pr, the empty set is the sole extension of $\mathbf{A}_3$. Thus, the four arguments get the same acceptability degree 0.3 ($\texttt{Deg}_{\texttt{pr}}^{\mathbf{A}_3}(.) = 0.3$). Consequently, $\texttt{Loss}_{\texttt{pr}}^{\mathbf{A}_3}(a) = 0.7$. Since the two attackers (a_1 and a_2) of a have the same degree, the previous definition declares their attacks as symmetric and assigns to them the same contribution (0.35 each). However, this division is unfair. Consider the two argumentation graphs $\mathbf{A}_3' = \mathbf{A}_3 \ominus \{(a_1, a)\}$ and $\mathbf{A}_3'' = \mathbf{A}_3 \ominus \{(a_2, a)\}$. It is easy to check that $\mathbf{A}_3'$ has one preferred extension, $\{a, a_1\}$. Thus, $\texttt{Deg}_{\texttt{pr}}^{\mathbf{A}_3'}(a) = 1$, $\texttt{Loss}_{\texttt{pr}}^{\mathbf{A}_3'}(a) = 0$, and the marginal contribution of the attack (a_1, a) is 0.7. However, the sole preferred extension of $\mathbf{A}_3''$ is the empty set. Hence, $\texttt{Deg}_{\texttt{pr}}^{\mathbf{A}_3''}(a) = 0.3$, $\texttt{Loss}_{\texttt{pr}}^{\mathbf{A}_3'}(a) = 0.7$, and the marginal contribution of the attack (a_2, a) is 0. This shows that the two attacks do not have the same impact on a.

The previous example suggests that the notion of symmetric attacks should not be defined on the basis of acceptability degrees, but rather be based on the comparison of the marginal contributions of the two attacks. However, like with dummy, the marginal contributions in the initial graph may lead to unfair divisions as shown by the following example.

Example 4 *Consider the argumentation graph $\mathbf{A}_4$ depicted below, where each argument has a basic strength equal to 1.*

Consider stable semantics st. *The graph* $\mathbf{A}_4$ *has two stable extensions* $\{a_1, a_2, a_3\}$ *and* $\{a_1, a_2, a_4\}$. *Thus,* $\mathtt{Deg}_{\mathtt{st}}^{\mathbf{A}_4}(a) = 0$ *and* $\mathtt{Loss}_{\mathtt{st}}^{\mathbf{A}_4}(a) = 1$.



Let $\mathbf{A}_{4i} = \mathbf{A}_4 \ominus \{(a_i, a)\}$, *with* $i \in \{1, 2, 3\}$. *For any* $i \neq j$, $\mathtt{Loss}_{\mathtt{st}}^{\mathbf{A}_{4i}}(a) = \mathtt{Loss}_{\mathtt{st}}^{\mathbf{A}_{4j}}(a)$. *Hence, the three attacks on* a *are symmetric, and get the same contribution (*$\frac{1}{3}$*). While* (a_1, a) *and* (a_2, a) *are clearly symmetric, this is not the case for* (a_1, a) *(respectively* (a_2, a)*) and* (a_3, a). *Indeed,* (a_1, a) *alone leads to a loss of 1 for* a, *while* (a_3, a) *alone (i.e., in graph* $\mathbf{A}_4 \ominus \{(a_1, a), (a_2, a)\}$*) leads only to a loss of 0.5.*

Symmetric attacks should then be defined by comparing the marginal contributions of those attacks in all the possible graphs where subsets of the target's attacks are removed.

Axiom 3 (Symmetry) *A contribution measure* $\mathbf{C}$ *satisfies* symmetry *iff,* $\forall \mathbf{S} \in \mathtt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $\forall a \in \mathcal{A}$ *with* $|\mathtt{Att}_{\mathbf{A}}(a)| \geq 2$, $\forall r_i, r_j \in \mathtt{Att}_{\mathbf{A}}(a)$ *with* $r_i \neq r_j$, *if* $\forall X \subseteq \mathtt{Att}_{\mathbf{A}}(a) \setminus \{r_i, r_j\}$, $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r_i\})}(a) = \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r_j\})}(a)$, *then* $\mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r_i) = \mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r_j)$. *We say that the two attacks* r_i *and* r_j *are* symmetric.

Example 4 (Cont) Under stable semantics, a contribution measure satisfying symmetry declares the two attacks (a_1, a), (a_2, a) symmetric, and both are not symmetric with (a_3, a).

A fair division of an argument's loss among attacks should take into account the power of each attack. In $\mathbf{A}_4$, the attack (a_3, a) is less harmful than both (a_1, a) and (a_2, a). Thus, the former should receive a lower part than the latter. The next axiom captures this idea. It states that the greater the marginal contribution of an attack, the greater its contribution.

Axiom 4 (Dominance) *A contribution measure* $\mathbf{C}$ *satisfies* dominance *iff,* $\forall \mathbf{S} \in \mathtt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $\forall a \in \mathcal{A}$ *with* $|\mathtt{Att}_{\mathbf{A}}(a)| \geq 2$, $\forall r_i, r_j \in \mathtt{Att}_{\mathbf{A}}(a)$ *with* $r_i \neq r_j$, *if*

- $\exists X \subseteq \mathtt{Att}_{\mathbf{A}}(a) \setminus \{r_i, r_j\}$ *such that* $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r_i\})}(a) > \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r_j\})}(a)$, *and*
- $\forall X' \subseteq \mathtt{Att}_{\mathbf{A}}(a) \setminus \{r_i, r_j\}$ *such that* $X' \neq X$, $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X' \cup \{r_i\})}(a) \geq \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X' \cup \{r_j\})}(a)$

then $\mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r_j) > \mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r_i)$.

All the previous axioms assume a fixed semantics, and discuss how weights are assigned to attacks under that semantics. The last axiom shows how those weights may vary from one semantics to another. It states that if the marginal contribution of an attack is the same under two different semantics, then it will receive the same contribution in both cases. This axiom ensures that the contribution of an attack depends solely on its marginal contribution to the loss of its target. This prevents measures from taking into account features of semantics like its name, whether it is extension-based, whether it is binary (i.e., it allows two degrees), etc.

Axiom 5 (Coherence) *A contribution measure* $\mathbf{C}$ *satisfies* coherence *iff,* $\forall \mathbf{S}, \mathbf{S}' \in \mathtt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $\forall a \in \mathcal{A}$ *with* $\mathtt{Att}_{\mathbf{A}}(a) \neq \emptyset$, $\forall r \in \mathtt{Att}_{\mathbf{A}}(a)$, *if* $\forall X \subseteq \mathtt{Att}_{\mathbf{A}}(a) \setminus \{r\}$, $\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus X}(a) - \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A} \ominus (X \cup \{r\})}(a) = \mathtt{Loss}_{\mathbf{S}'}^{\mathbf{A} \ominus X}(a) - \mathtt{Loss}_{\mathbf{S}'}^{\mathbf{A} \ominus (X \cup \{r\})}(a)$, *then* $\mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S},\mathbf{A}}(r) = \mathtt{Ctr}_{\mathbf{C}}^{\mathbf{S}',\mathbf{A}}(r)$.

Example 2 (Cont) Consider the argumentation graph $\mathbf{A}_2$. Any contribution measure satisfying the axiom of Coherence assigns the same value to r_1 (respectively r_2) under stable, grounded, complete, and preferred semantics [Dung, 1995].

The five axioms are not fully *independent*. Indeed, Anonymity, Dummy and Dominance follow from the two other axioms (i.e., from Coherence and Symmetry).

Proposition 3 *Let* $\mathbf{C}$ *be a contribution measure.*

- *If* $\mathbf{C}$ *satisfies Coherence, then* $\mathbf{C}$ *satisfies Dummy.*
- *If* $\mathbf{C}$ *satisfies Coherence and Symmetry, then* $\mathbf{C}$ *satisfies Anonymity and Dominance.*

Coherence and Symmetry are however independent, i.e., none of them is implied by the other.

Proposition 4 *Symmetry and Coherence are independent.*

Hopefully, the five axioms are *compatible*, i.e., they can all be satisfied by a given contribution measure.

Proposition 5 *Symmetry and Coherence are compatible.*

From propositions 3 and 5, it follows that the five axioms are compatible.

5 Shapley Contribution Measure

The previous section introduced the notion of contribution measure, and five axioms that fair and reasonable measures should satisfy. The following questions raise then naturally:

- *Existence*: is there a measure which satisfies the axioms?
- *Uniqueness*: if yes, is it unique?

Hopefully, the answers to both questions are positive. Before presenting the formal results, let us start by introducing the key function Sh, called *Shapley measure* in the paper.

Definition 9 (Sh**)** Sh *is the function on* $\mathtt{Sem} \times \mathtt{AG}$ *such that* $\forall \mathbf{S} \in \mathtt{Sem}$, $\forall \mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, $\mathtt{Sh}(\mathbf{S}, \mathbf{A})$ *is the function* $\mathbf{s}$ *from* $\mathcal{R}$ *to* $\mathbb{R}$ *defined as follows:* $\forall r = (b, a) \in \mathcal{R}$,

$$\mathbf{s}(r) = \sum_{X \subseteq Y} \frac{|X|!(n - |X| - 1)!}{n!} (\mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}_1}(a) - \mathtt{Loss}_{\mathbf{S}}^{\mathbf{A}_2}(a))$$

where $Y = \mathtt{Att}_{\mathbf{A}}(a) \setminus \{r\}$, $n = |\mathtt{Att}_{\mathbf{A}}(a)|$, $\mathbf{A}_1 = \mathbf{A} \ominus X$, *and* $\mathbf{A}_2 = \mathbf{A} \ominus (X \cup \{r\})$.

The function Sh assigns a real value to each attack, which is the weighted sum of its marginal contributions to the global loss of the targeted argument. It is easy to show that the previous definition of Sh is equivalent to the following one.

Theorem 1 *Let* $\mathbf{S} \in \mathtt{Sem}$, $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle \in \mathtt{AG}$, *and* $\mathbf{s} = \mathtt{Sh}(\mathbf{S}, \mathbf{A})$. *It holds that* $\forall (b, a) \in \mathcal{R}$,

$$\mathbf{s}((b, a)) = \sum_{X \subseteq Y} \frac{|X|!(n - |X| - 1)!}{n!} (\mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}_2}(a) - \mathtt{Deg}_{\mathbf{S}}^{\mathbf{A}_1}(a)),$$

where $Y = \mathtt{Att}_{\mathbf{A}}(a) \setminus \{(b, a)\}$, $n = |\mathtt{Att}_{\mathbf{A}}(a)|$, $\mathbf{A}_1 = \mathbf{A} \ominus X$, *and* $\mathbf{A}_2 = \mathbf{A} \ominus (X \cup \{(b, a)\})$.

It is worth noticing that $\texttt{Sh}(\mathbf{S}, \mathbf{A})$ is the well-known *Shapley value*, proposed as a solution for transferable utility games (TU games) by Shapley [1953]. A TU game is a set of *agents* and a real-valued *characteristic function* assigning a value to each subset of agents. Each subset is a coalition and its value represents how much the coalition can get for itself. The key problem is then, how the agents of the game divide the value of the grand coalition (the one made of all agents). Shapley value provides a unique division for each game.

In an argumentation context, each pair $(\mathbf{S}, \mathbf{A})$, with $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$, can be seen as a *set* of TU games (one per argument in $\mathcal{A}$). The agents of the game corresponding to an argument $a \in \mathcal{A}$ are the attacks on a, and the characteristic function is $\texttt{Loss}$. The latter evaluates the loss of a under semantics $\mathbf{S}$ when a is attacked by any subset of its attackers. Those subsets play the role of coalitions.

Shapley characterized his value with axioms, one of which is *additivity*. The latter shows how different games can be combined. Despite the relationship between a contribution measure and TU games, Additivity has no counter-part in the previous section since it does not make sense in the argumentation context. Shubik [1962] has shown, however, that symmetry follows from Shapley's axiom of anonymity. Note that the latter is based on permutations in the roles of agents, and thus is different from our Anonymity axiom.

We show next that $\texttt{Sh}$ is a contribution measure. Indeed, it assigns a value in the interval $[0, 1]$ to each attack. Furthermore, it satisfies the efficiency condition of Definition 8.

Theorem 2 $\texttt{Sh}$ *is a contribution measure.*

Let us illustrate this measure on the running examples.

Example 1 (Cont) In the graph $\mathbf{A}_1$, $\texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S}_1,\mathbf{A}_1}((b, a)) = 1$.

Example 2 (Cont) In the graph $\mathbf{A}_2$, $\texttt{Ctr}_{\texttt{Sh}}^{x,\mathbf{A}_2}(r_1) = \texttt{Ctr}_{\texttt{Sh}}^{x,\mathbf{A}_2}(r_2) = \frac{1}{2}$ for $x \in \{\texttt{gr}, \texttt{co}, \texttt{pr}, \texttt{co}\}$. Consider now $\mathbf{h}$-categorizer semantics. $\texttt{Deg}_{\mathbf{h}}^{\mathbf{A}_2}(a) = \frac{1}{3}$, and thus $\texttt{Loss}_{\mathbf{h}}^{\mathbf{A}_2}(a) = \frac{2}{3}$. So, $\texttt{Ctr}_{\texttt{Sh}}^{\mathbf{h},\mathbf{A}_2}(r_1) = \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{h},\mathbf{A}_2}(r_2) = \frac{1}{3}$.

Example 3 (Cont) In the graph $\mathbf{A}_3$, $\texttt{Ctr}_{\texttt{Sh}}^{\texttt{pr},\mathbf{A}_3}((a_1, a)) = 0.7$, $\texttt{Ctr}_{\texttt{Sh}}^{\texttt{pr},\mathbf{A}_3}((a_2, a)) = 0$. Note that (a_2, a) is a dummy attack under preferred semantics. However, as we will see in the next section, it is not dummy under $\mathbf{h}$-categorizer semantics.

Example 4 (Cont) In the graph $\mathbf{A}_4$, $\texttt{Ctr}_{\texttt{Sh}}^{x,\mathbf{A}_4}((a_1, a)) = \texttt{Ctr}_{\texttt{Sh}}^{x,\mathbf{A}_4}((a_2, a)) = \frac{5}{12}$ and $\texttt{Ctr}_{\texttt{Sh}}^{x,\mathbf{A}_4}((a_3, a)) = \frac{1}{6}$, where $x \in \{\texttt{pr}, \texttt{st}\}$. The two first attacks are more harmful than (a_3, a).

The following result is of great importance since it positively answers the two questions on the *existence* and *uniqueness* of a contribution measure that satisfies the axioms. Indeed, it provides a characterization, which states that not only $\texttt{Sh}$ satisfies the axioms (*existence*), but also it is the *unique* contribution measure that satisfies them.

Theorem 3 *A contribution measure* $\mathbf{C}$ *satisfies the two axioms of Symmetry and Coherence if and only if* $\mathbf{C} = \texttt{Sh}$.

From the previous result and the links between the five axioms, it follows that $\texttt{Sh}$ satisfies the five axioms.

Properties	**P1**	**P2**	**P3**	**P4**	**P5**
Stable	no	no	no	no	no
Preferred	no	no	no	no	no
Complete	no	yes	no	no	no
Grounded	no	yes	no	no	no
$\mathbf{h}$-Categorizer	yes	yes	yes	no	no

Table 1: Properties

6 Properties Related to Semantics

Shapley measure $\texttt{Sh}$ can be applied to any semantics in $\texttt{Sem}$. We show that the contributions it assigns to attacks respect however, properties of the underlying semantics. For that purpose, we consider the semantics defined in Section 3, a given argumentation graph $\mathbf{A} = \langle \mathcal{A}, w, \mathcal{R} \rangle$, and $a \in \mathcal{A}$ such that $\texttt{Att}_{\mathbf{A}}(a) \neq \emptyset$. Let $\texttt{Att}_{\mathbf{A}}(a) = \{r_1, \ldots, r_n\}$ and $\texttt{Sc}(r_i) = a_i$. Hence, the attackers of a are $\{a_1, \ldots, a_n\}$. We focus on the following properties: For any $i, j \in \{1, \ldots, n\}$,

(P1) $\texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_i) > 0$.

(P2) $\texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_i) > 0 \Rightarrow \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_i) > 0$.

(P3) $\texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_i) = 0 \Rightarrow \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_i) = 0$.

(P4) $\texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_i) > \texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_j) \Rightarrow \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_i) > \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_j)$.

(P5) $\texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_i) = \texttt{Deg}_{\mathbf{S}}^{\mathbf{A}}(a_j) \Rightarrow \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_i) = \texttt{Ctr}_{\texttt{Sh}}^{\mathbf{S},\mathbf{A}}(r_j)$.

(P1) states that there is no dummy attack. **(P2)** follows from **(P1)** and ensures that each serious attacker contributes to the loss of a. **(P3)** ensures that killed attackers do not contribute to the loss of their target. The two last properties state that the order between acceptability degrees is preserved.

Proposition 6 *Table 1 resumes the properties that are satisfied/violated by the semantics introduced in Section 3.*

The results show that under $\mathbf{h}$-Categorizer semantics, there is no dummy attack. They also confirm the necessity of defining contribution measures on the basis of marginal contributions of attackers rather than on their acceptability degrees.

7 Conclusion

The paper introduced a novel concept, contribution measure, which evaluates the intensity of each attack in an argumentation graph. It followed an axiomatic approach. Indeed, it defined axioms that characterize reasonable measures. Then, it showed that there is a unique measure which satisfies them, and it is the well-known Shapley value.

Future work consists of investigating the properties of Shapley measure under semantics like the ones proposed in [Dung *et al.*, 2007; Grossi and Modgil, 2015; Gabbay and Rodrigues, 2015; da Costa Pereira *et al.*, 2011], those proposed in probabilistic argumentation settings [Hunter, 2013; Li *et al.*, 2011], or in weighted argumentation graphs, i.e., graphs where attacks are assigned weights (see [Cayrol *et al.*, 2010; Dunne *et al.*, 2011; 2010]).

Acknowledgments

This work was supported by ANR-13-BS02-0004 and ANR-11-LABX-0040-CIMI.

References

[Amgoud and Ben-Naim, 2016] Leila Amgoud and Jonathan Ben-Naim. Axiomatic foundations of acceptability semantics. In *Proceedings of the 15th International Conference Principles of Knowledge Representation and Reasoning, KR'2016*, pages 2–11, 2016.

[Amgoud *et al.*, 2017] Leila Amgoud, Jonathan Ben-Naim, Dragan Doder, and Srdjan Vesic. Acceptability semantics for weighted argumentation frameworks. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'2017*, 2017.

[Benferhat *et al.*, 1993] Salem Benferhat, Didier Dubois, and Henri Prade. Argumentative inference in uncertain and inconsistent knowledge bases. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence, UAI'93*, pages 411–419, 1993.

[Besnard and Hunter, 2001] Philippe Besnard and Anthony Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1-2):203–235, 2001.

[Cayrol *et al.*, 2010] Claudette Cayrol, Caroline Devred, and Marie-Christine Lagasquie-Schiex. Acceptability semantics accounting for strength of attacks in argumentation. In *European Conference on Artificial Intelligence, ECAI'2010*, pages 995–996, 2010.

[da Costa Pereira *et al.*, 2011] Celia da Costa Pereira, Andrea Tettamanzi, and Serena Villata. Changing one's mind: Erase or rewind? In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI'11*, pages 164–171, 2011.

[Dung *et al.*, 2007] P.M. Dung, P. Mancarella, and F. Toni. Computing ideal skeptical argumentation. *Artificial Intelligence*, 171:642–674, 2007.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77:321–357, 1995.

[Dunne *et al.*, 2010] Paul Dunne, Diego Martinez, Alejandro Garcia, and Guillermo Simari. Computation with varied-strength attacks in abstract argumentation frameworks. In *Computational Models of Argument, COMMA'2010*, pages 207–218, 2010.

[Dunne *et al.*, 2011] Paul Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.

[Gabbay and Rodrigues, 2015] Dov Gabbay and Odinaldo Rodrigues. Equilibrium states in numerical argumentation networks. *Logica Universalis*, 9(4):411–473, 2015.

[Grossi and Modgil, 2015] Davide Grossi and Sanjay Modgil. On the graded acceptability of arguments. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence, IJCAI'15*, pages 868–874, 2015.

[Hunter, 2013] Anthony Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.

[Li *et al.*, 2011] Hengfei Li, Nir Oren, and Timothy Norman. Probabilistic argumentation frameworks. In *First International Workshop on Theorie and Applications of Formal Argumentation, TAFA*, pages 1–16, 2011.

[Pu *et al.*, 2014] F. Pu, J. Luo, Y. Zhang, and G. Luo. Argument ranking with categoriser function. In *Proceedings of the 7th International Knowledge Science, Engineering and Management Conference, KSEM'14*, pages 290–301, 2014.

[Shapley, 1953] Lloyd Shapley. A values for n-person games. H.W. Kuhn and A.W. Tucker, eds., Contributions to the theory of games, vol. II. *Annal of Mathematics Studies*, (28), 1953.

[Shubik, 1962] Martin Shubik. Incentives, decentralized control, the assignment of joint costs, and internal pricing. *Management Science*, (8(3)):325–343, 1962.

[Simari and Rahwan, 2009] Guillermo Ricardo Simari and Iyad Rahwan, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.