Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning
KR Main Track

Axiomatic Characterisations of Argumentation Semantics

Leila Amgoud
CNRS – IRIT, France
leila.amgoud@irit.fr

Abstract

The evaluation of argument strength lies at the core of any argumentation system. Numerous semantics have been proposed for this purpose, along with a variety of principles (or axioms) that such semantics are expected to satisfy. Most existing semantics in the literature have been analyzed and compared in light of these principles. While this body of work marks a significant step toward establishing the *theoretical foundations* of argumentation semantics, it remains incomplete. In particular, characterizations of entire classes of semantics that uniquely satisfy specific subsets of axioms are still lacking, leaving open questions on the kind of semantics that can still be defined and their added values.

This paper addresses this gap by establishing *representation theorems* that explicitly relate subsets of principles to corresponding classes of semantics. These semantics are defined through two mathematical functions: an impact function and an aggregation operator, each satisfying specific structural properties. We demonstrate how these principles offer a uniform and concise explanatory framework for the identified semantics. Finally, we show that classical extension-based semantics do not belong to these classes.

1 Introduction

Argumentation is a reasoning paradigm that justifies claims through supporting arguments. Over the past few decades, it has become a key approach in artificial intelligence, enabling solutions to tasks such as non-monotonic reasoning (Dung 1995), decision making under uncertainty (Amgoud and Prade 2009), and negotiation (Dimopoulos, Mailly, and Moraitis 2019) (see (Baroni et al. 2018; Simari et al. 2021) for further applications).

An argumentation-based system is typically defined as a set of arguments, each assigned an initial weight and potentially involved in attacks against, or being attacked by, other arguments. A central question in such systems is how to evaluate the *strength* of an argument—that is, *the degree of support it provides to its associated claim*. Accurately evaluating argument strength is crucial, as it directly impacts the system's outcomes. For instance, a reasoning engine might infer the conclusions that are supported by the strongest arguments. As a result, a substantial body of research has been devoted to formalizing evaluation methods, commonly referred to as *semantics*. Two main families of semantics can be distinguished: *extension-based* semantics, initiated in (Dung 1995), and *gradual* semantics, introduced in (Cayrol and Lagasquie 2005). Instances of the former can be found in (Dung 1995; Baroni, Giacomin, and Guida 2005; Caminada 2006; Dung, Mancarella, and Toni 2007), while examples of the latter are defined in (Leite and Martins 2011; da Costa Pereira, Tettamanzi, and Villata 2011; Rago et al. 2016; Amgoud et al. 2017; Potyka 2019; Amgoud and Doder 2019; Libman, Oren, and Yun 2024).

The large number of proposed semantics has prompted the need for a deeper understanding of their foundational assumptions and for principled methods to compare them. This has led to the emergence of an *axiomatic approach*, initiated in (Amgoud and Ben-Naim 2013; Amgoud and Ben-Naim 2016), in which semantics are modeled as abstract functions expected to satisfy a set of axioms (also referred to as principles). Some of these axioms capture desirable properties that any reasonable semantics should uphold, while others formalize strategic choices that may vary depending on the nature of the arguments. For example, considering the number of attackers may be essential for analogical arguments, but not for deductive ones. Most existing semantics have been evaluated against these axioms and additional ones proposed in (Bonzon et al. 2016; Amgoud et al. 2017; Baroni, Rago, and Toni 2018).

Although the axiomatic approach is a powerful tool for establishing the theoretical foundations of semantics, it remains relatively underexplored in the literature. In particular, *representation theorems*—one-to-one correspondences between specific subsets of axioms and the classes of semantics that satisfy them—are still lacking. Likewise, impossibility results, which demonstrate that no semantics can satisfy certain combinations of axioms, are also lacking. Such results would offer a more comprehensive understanding of the landscape of definable semantics.

This paper complements the existing literature by providing formal results that establish the missing connections, and thus bridges the existing gap. We consider an axiomatic framework grounded on the set of principles proposed in (Amgoud et al. 2017) and a new principle, called *Locality*. We characterize the classes of semantics that satisfy subsets of the principles and analyse their explainability. More precisely the contributions of the paper are fivefold:

- We consider an existing set of principles, introduce a new principle, analyze their interrelationships, and demon-

strate the existence of at least one semantics that satisfies all of them.

- We provide representation theorems that connect subsets of principles to the corresponding classes of semantics that satisfy them. In particular, we characterize the class of semantics that satisfy the *Equivalence* principle, defined in (Amgoud and Ben-Naim 2016) and which states that an argument's strength should depend only on its initial weight and the strengths of its direct attackers. We show that any semantics satisfying this principle must be defined by an *evaluation method*—a pair of mathematical functions: an influence function and an aggregation operator, each meeting specific structural conditions.
- We additionally identify and characterize subclasses of the aforementioned semantics that adhere to further principles, and identify conditions under which certain principles are satisfied.
- We provide an impossibility result showing that the classical extension semantics from (Dung 1995) cannot be represented by evaluation methods. More precisely, no semantics based on an evaluation method can produce the *acceptability statuses* of arguments provided by those extension semantics.
- We propose a novel approach for explaining semantics of the identified classes. Unlike existing explanation models that are semantics-dependent, our approach is based on principles satisfied by semantics, making it more general and effective for explaining the logic behind semantics.

The paper is structured as follows: Section 2 introduces the formal setting, Section 3 recalls existing principles, defines a new one and studies their properties, Section 4 presents the characterizations, Section 5 provides additional properties, Section 6 studies the case of extension semantics, Section 7 addresses the explainability issue, and Section 8 is devoted to concluding remarks and perspectives.

2 Background

In the paper, we consider *weighted argumentation graphs* (wAGs) whose nodes are *arguments*, each of which has an *initial weight*, which may represent different notions (eg., votes given by users or certainty degrees), and edges represent *attacks* (or conflicts) between pairs of arguments. We denote by $\texttt{args}$ the set of all possible arguments.

Definition 1 (wAG)**.** *A* weighted argumentation graph *is a tuple* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle$*, where* $\mathcal{A} \subseteq \texttt{args}$ *is non-empty and finite,* $\mathbf{w} : \mathcal{A} \to [0,1]$*,* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$*. Let* $\texttt{wAG}$ *be the set of all wAGs that can be built from* $\texttt{args}$*.*

Notations: For $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \texttt{wAG}$ and $a \in \mathcal{A}$, $\texttt{att}_{\mathbf{G}}(a) = \{b \in \mathcal{A} \mid (b,a) \in \mathcal{R}\}$, i.e., the *direct attackers* of a in $\mathbf{G}$. For $X \subseteq \mathcal{A}$, we denote by $\mathbf{G}_{\downarrow X}$ the wAG that is reduced to arguments of X, i.e., $\mathbf{G}_{\downarrow X} = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle$ such that $\mathcal{A}' = X$, $\mathcal{R}' = \mathcal{R} \cap (X \times X)$ and $\forall a \in X$, $\mathbf{w}'(a) = \mathbf{w}(a)$. Let now $\mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle \in \texttt{wAG}$ be such that $\mathcal{A} \cap \mathcal{A}' = \emptyset$. We define $\mathbf{G} \oplus \mathbf{G}' = \langle \mathcal{A} \cup \mathcal{A}', \mathbf{w}'', \mathcal{R} \cup \mathcal{R}' \rangle \in \texttt{wAG}$ such that $\forall a \in \mathcal{A}$ (resp. $a \in \mathcal{A}'$), $\mathbf{w}''(a) = \mathbf{w}(a)$ (resp. $\mathbf{w}''(a) = \mathbf{w}'(a)$).

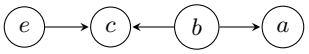
Property 1. *Let* $\{\mathbf{G}, \mathbf{G}'\} \subseteq \texttt{wAG}$ *be such that* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle$*,* $\mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle$ *and* $\mathcal{A} \cap \mathcal{A}' = \emptyset$*. For any* $a \in \mathcal{A}$ *(resp.* $a \in \mathcal{A}'$*),* $\texttt{att}_{\mathbf{G} \oplus \mathbf{G}'}(a) = \texttt{att}_{\mathbf{G}}(a)$ *(resp.* $\texttt{att}_{\mathbf{G} \oplus \mathbf{G}'}(a) = \texttt{att}_{\mathbf{G}'}(a)$*).*

We recall the notion of *path* in a graph and the useful idea of *attack structure* from (Amgoud, Doder, and Vesic 2022).

Definition 2 (Path - Attack Structure)**.** *Let* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \texttt{wAG}$ *and* $a \in \mathcal{A}$*.*

- *A* path *from* $b \in \mathcal{A}$ *to* a *is a finite non-empty sequence* $\langle x_1, \ldots, x_n \rangle$ *of arguments in* $\mathcal{A}$ *such that* $x_1 = b$*,* $x_n = a$*, and* $\forall 1 \leq i < n$*,* $(x_i, x_{i+1}) \in \mathcal{R}$*.*
- *The* attack structure *of* a *in* $\mathbf{G}$ *is* $\texttt{Str}_{\mathbf{G}}(a) = \{a\} \cup \{b \in \mathcal{A} \mid$ *there is a path from* b *to* a *in* $\mathbf{G}\}$*.*

Example 1. *The attack structures of* a *and* c *in the wAG depicted below are* $\{a, b\}$ *and* $\{b, c, e\}$*, respectively.*



Property 2. *Let* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \texttt{wAG}$ *and* $a \in \mathcal{A}$*. For any* $x \in \texttt{Str}_{\mathbf{G}}(a)$*,* $\texttt{att}_{\mathbf{G}}(x) \subseteq \texttt{Str}_{\mathbf{G}}(a)$*.*

The following definition recalls the notion of *isomorphism* between weighted graphs.

Definition 3 (Isomorphism)**.** *Let* $\mathbf{G}, \mathbf{G}' \in \texttt{wAG}$ *such that* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle$ *and* $\mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle$*. An* isomorphism *from* $\mathbf{G}$ *to* $\mathbf{G}'$ *is a bijective function* $h : \mathcal{A} \to \mathcal{A}'$ *such that:*

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

Let us now recall the notion of *preordered set*. It is a pair $(X, \succeq)$ consisting of a set X of objects and a preorder (*reflexive* and *transitive* binary relation) on X. For $x, y \in X$, $x \succeq y$ stands for "x is at least as strong as y". The symbol $\succ$ stands for the strict version of $\succeq$, that is, $x \succ y$ iff $x \succeq y$ and $y \not\succeq x$. The pair $(X, \succeq)$ is *totally preordered* iff $\forall x, y \in X$, $x \succeq y$ or $y \succeq x$ (all elements of X are comparable).

The next concept is of crucial importance for the purpose of this paper. It concerns the notion of *evaluation method*, which has been discussed in several works on gradual semantics (e.g., (Cayrol and Lagasquie 2005; Leite and Martins 2011; Egilmez, Martins, and Leite 2013; Amgoud and Doder 2019)). In the following, we consider its most basic form—namely, a pair of functions—without imposing any constraints on them.

Definition 4 (Evaluation Method)**.** *An* evaluation method *is a pair* $(\mathbf{f}, \mathbf{g})$ *such that:*

- $\mathbf{g} : \bigcup_{n=0}^{+\infty} [0,1]^n \to \mathbf{I}$ *where:*
  - $(\mathbf{I}, \succeq)$ *is a totally preordered set such that* $\texttt{e} \in \mathbf{I}$ *and* $\forall x \in \mathbf{I}$*,* $x \succ \texttt{e}$ *(i.e.,* $\texttt{e}$ *is the weakest element in* $\mathbf{I}$*).*
  - $\mathbf{g}() = \texttt{e}$
- $\mathbf{f} : [0,1] \times \mathbf{I} \to [0,1]$*.*

The codomain $\mathbf{I}$ of the function $\mathbf{g}$ is an ordered set that contains a **weakest element**. It may be the interval $[0, +\infty)$ (as in (Amgoud and Doder 2019)), the unit interval $[0,1]$, or any other numerical or quantitative scale.

$\mathbf{f}$	*f-Invariance*	$\forall x \in [0,1], \mathbf{f}(x, \mathtt{e}) = x$
	Robustness	$\forall x, y \in [0,1]$, if $x > 0$, then $\mathbf{f}(x,y) > 0$
	Increasing-F	$\forall x, x' \in [0,1], \forall y \in \mathbf{I}$, if $x > x'$, then $\mathbf{f}(x,y) > \mathbf{f}(x',y)$
	Decreasing-S	$\forall y \in [0,1], \forall x, x' \in \mathbf{I}$, if $x \succ x'$, then $\mathbf{f}(y,x) < \mathbf{f}(y,x')$
$\mathbf{g}$	*Identity*	$\forall x \in [0,1], \mathbf{g}(x) = x$
	Commutativity	If μ is a permutation on $\{1,\dots,n\}$, then $\mathbf{g}(x_1,\dots,x_n) = \mathbf{g}(x_{\mu(1)},\dots,x_{\mu(n)})$
	g-Invariance	$\mathbf{g}(x_1,\dots,x_n,0) = \mathbf{g}(x_1,\dots,x_n)$
	Monotonicity	If $y \geq z$, then $\mathbf{g}(x_1,\dots,x_n,y) \succeq \mathbf{g}(x_1,\dots,x_n,z)$
	Strict Monotonicity	If $y > z$, then $\mathbf{g}(x_1,\dots,x_n,y) \succ \mathbf{g}(x_1,\dots,x_n,z)$

Table 1: Properties of $\mathbf{f}$ and $\mathbf{g}$.

Table 1 lists properties that the two functions defining an evaluation method may satisfy. The f-invariance property states that the weakest element $\mathtt{e}$ of the set $\mathbf{I}$ serves as a neutral element for the function $\mathbf{f}$. The robustness property ensures that $\mathbf{f}$ returns a positive value whenever its first argument is positive. The remaining two properties specify that $\mathbf{f}$ is increasing in its first variable and decreasing in its second. The function $\mathbf{g}$ returns its input when given a singleton, is insensitive to the order of its arguments, and has 0 as its neutral element. The last two properties state that $\mathbf{g}$ is (strictly) increasing. We will show in the next sections that these properties are crucial for defining semantics that have a specific behaviour.

Property 3. *Let* $\mathbf{g} : \bigcup_{n=0}^{+\infty} [0,1]^n \to \mathbf{I}$.

- $\forall (x_1,\dots,x_n) \in [0,1]^n$, $\mathbf{g}(x_1,\dots,x_n) \succeq \mathtt{e}$.
- *If* $\mathbf{g}$ *satisfies g-Invariance, then the following hold:*
  - $\mathbf{g}(0) = \mathbf{g}(0,\dots,0)$.
  - *If* $\mathbf{g}$ *satisfies monotonicity, then:*

  a) $\mathbf{g}(x_1,\dots,x_n,y_1,\dots,y_k) \succeq \mathbf{g}(x_1,\dots,x_n)$.

  b) $\mathbf{g}(x_1,\dots,x_n) \succeq \mathbf{g}(0)$.

A fundamental concept in the field of argumentation is that of semantics—a formal method for evaluating the strength of arguments in weighted argumentation graphs. Semantics is typically defined as a function that assigns a value from an ordered scale to each argument, with higher values indicating stronger arguments. In the following discussion, we adopt the scale $[0,1]$, which is the most commonly used in the literature.

Definition 5 (Semantics)**.** *A* semantics *is a function* $\mathbf{S}$ *mapping any* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}$ *to* $\delta^{\mathbf{S}}_{\mathbf{G}} : \mathcal{A} \to [0,1]$. *For* $a \in \mathcal{A}$, $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ *is the* strength degree *of* a *in* $\mathbf{G}$ *under* $\mathbf{S}$.

Some semantics may be defined using specific evaluation methods. An example of such semantics is h-Categorizer from (Besnard and Hunter 2001; Pu, andY. Zhang, and Luo 2014). Other examples can be found in (Leite and Martins 2011; Amgoud and Doder 2019).

Definition 6. *Let* $\mathbf{S}$ *be a semantics and* $(\mathbf{f}, \mathbf{g})$ *an evaluation method.* $\mathbf{S}$ *is* based on $(\mathbf{f}, \mathbf{g})$ *iff for any* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}$*, for any* $a \in \mathcal{A}$*,*

$$\delta^{\mathbf{S}}_{\mathbf{G}}(a) = \mathbf{f}\Big(\mathbf{w}(a), \mathbf{g}\big(\delta^{\mathbf{S}}_{\mathbf{G}}(b_1),\dots,\delta^{\mathbf{S}}_{\mathbf{G}}(b_n)\big)\Big) \qquad (1)$$

where $\{b_1,\dots,b_n\} = \mathtt{att}_{\mathbf{G}}(a)$. *Let* $\mathtt{Eq}^{\mathbf{S}}_{\mathbf{G}}(a) = \{\delta^{\mathbf{S}}_{\mathbf{G}}(a), \delta^{\mathbf{S}}_{\mathbf{G}}(y_1),\dots,\delta^{\mathbf{S}}_{\mathbf{G}}(y_m)\}$ *denote the set of all equations invoked during the recursive computation of* $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$.

Here, $\mathbf{g}$ is an *aggregation* function that combines the strengths of all *direct attackers* of a into a single value, representing the overall strength of the attacking group. The function $\mathbf{f}$ is an *influence* function that determines how this aggregated value affects the initial weight of a. Note that the above recursive definition implies that, in order to compute the value of a, the semantics must solve a system of equations $\mathtt{Eq}^{\mathbf{S}}_{\mathbf{G}}(a) = \{\delta^{\mathbf{S}}_{\mathbf{G}}(a), \delta^{\mathbf{S}}_{\mathbf{G}}(y_1),\dots,\delta^{\mathbf{S}}_{\mathbf{G}}(y_m)\}$, thus it must also evaluate the arguments $y_1,\dots,y_m$, which include the direct $(b_1\dots,b_n)$ and indirect attackers of a.

We show that when a semantics is based on an evaluation method, the value assigned to an argument depends solely on a subgraph of the weighted argumentation graph (wAG), specifically the portion that captures the attack structure relevant to that argument. This result is significant, as it precisely identifies the set of arguments that influence—either directly or indirectly—the strength of a given argument.

Theorem 1. *Let* $\mathbf{S}$ *be a semantics based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*,* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}$*, and* $a \in \mathcal{A}$*. The following equivalence holds:*

$$\mathtt{Eq}^{\mathbf{S}}_{\mathbf{G}}(a) = \{\delta^{\mathbf{S}}_{\mathbf{G}}(a), \delta^{\mathbf{S}}_{\mathbf{G}}(y_1),\dots,\delta^{\mathbf{S}}_{\mathbf{G}}(y_m)\} \quad \textit{iff}$$

$$\mathtt{Str}_{\mathbf{G}}(a) = \{a, y_1,\dots,y_m\}.$$

Semantics based on evaluation methods solve a system of equations (of the form (1) in Definition 6), with one equation per argument. This system may, in general, admit multiple solutions, each representing a different semantics. Evaluation methods that guarantee a unique solution—and therefore define a unique semantics—are referred to as *rational*.

Definition 7 (Rationality)**.** *An evaluation method* $(\mathbf{f}, \mathbf{g})$ *is* rational *if there is a unique semantics that is based on* $(\mathbf{f}, \mathbf{g})$*, that is, if two semantics* $\mathbf{S}$ *and* $\mathbf{S}'$ *are both based on* $(\mathbf{f}, \mathbf{g})$*, then* $\mathbf{S} = \mathbf{S}'$.

The rationality of an evaluation method is of great importance, as it ensures that the semantics it defines is well-characterized and unambiguous. Obviously, if weighted graphs in $\mathtt{wAG}$ are acyclic (no cycles involved), then any evaluation method would be rational.

3 Principles

For a semantics to be reasonable, it should satisfy certain desirable properties, referred to as axioms in (Amgoud and Ben-Naim 2016) or *principles* in (Amgoud et al. 2017; Baroni, Rago, and Toni 2018; Baroni and Giacomin 2007). These properties capture high-level behavioral expectations that a semantics may be required to fulfill. In what follows, we consider the set of principles introduced in (Amgoud et al. 2017), to which we add a new one, called *Locality*. Let $\mathbf{S}$ be an arbitrary but fixed semantics. It satisfies a principle if it fulfills its conditions.

Anonymity: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle, \mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle \in \mathtt{wAG}$, for any isomorphism h from $\mathbf{G}$ to $\mathbf{G}'$, it holds that: $\forall\, a \in \mathcal{A}, \delta^{\mathbf{S}}_{\mathbf{G}}(a) = \delta^{\mathbf{S}}_{\mathbf{G}'}(h(a))$.

The Independence principle ensures that the value of an argument is invariant under changes to unrelated parts of the weighted graph.

Independence: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle, \mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle \in \mathtt{wAG}$ s.t $\mathcal{A} \cap \mathcal{A}' = \emptyset$, it holds: $\forall\, a \in \mathcal{A},\ \delta^{\mathbf{S}}_{\mathbf{G}}(a) = \delta^{\mathbf{S}}_{\mathbf{G} \oplus \mathbf{G}'}(a)$.

The principle of Directionality ensures that an argument can only be influenced by other arguments that are connected to it through a directed path.

Directionality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}, \forall \mathbf{G}' = \langle \mathcal{A}', \mathbf{w}', \mathcal{R}' \rangle \in \mathtt{wAG}$ s.t. $\mathcal{A}' = \mathcal{A}$, $\mathbf{w}' = \mathbf{w}$, $\mathcal{R}' = \mathcal{R} \cup \{(a, b)\}$, it holds: $\forall x \in \mathcal{A}$, if there is no path from b to x, then $\delta^{\mathbf{S}}_{\mathbf{G}}(x) = \delta^{\mathbf{S}}_{\mathbf{G}'}(x)$.

The next principle states that the strength degree of an argument depends only on its initial weight and the strength degrees of its direct attackers.

Equivalence: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathbf{w}(a) = \mathbf{w}(b)$,
- there exists a bijective function h from $\mathtt{att}_{\mathbf{G}}(a)$ to $\mathtt{att}_{\mathbf{G}}(b)$ s.t. $\forall x \in \mathtt{att}_{\mathbf{G}}(a), \delta^{\mathbf{S}}_{\mathbf{G}}(x) = \delta^{\mathbf{S}}_{\mathbf{G}}(h(x))$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) = \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

Maximality states that the strength of a non-attacked argument is equal to its initial weight.

Maximality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}$, $\forall a \in \mathcal{A}$, if $\mathtt{att}_{\mathbf{G}}(a) = \emptyset$, then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) = \mathbf{w}(a)$.

Neutrality ensures that worthless attackers have no impact on their targets.

Neutrality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathbf{w}(a) = \mathbf{w}(b)$,
- $\mathtt{att}_{\mathbf{G}}(b) = \mathtt{att}_{\mathbf{G}}(a) \cup \{x\}$ such that $x \in \mathcal{A} \setminus \mathtt{att}_{\mathbf{G}}(a)$ and $\delta^{\mathbf{S}}_{\mathbf{G}}(x) = 0$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) = \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

The next principle highlights the negative role of attacks. It states that an argument loses strength if it is attacked by at least one serious attack.

Weakening: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a \in \mathcal{A}$, if

- $\mathbf{w}(a) > 0$,
- $\exists b \in \mathtt{att}_{\mathbf{G}}(a)$ s.t. $\delta^{\mathbf{S}}_{\mathbf{G}}(b) > 0$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) < \mathbf{w}(a)$.

(Strong) Proportionality states that argument strength is sensitive to the initial weight.

Proportionality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathtt{att}_{\mathbf{G}}(a) = \mathtt{att}_{\mathbf{G}}(b)$,
- $\mathbf{w}(a) \geq \mathbf{w}(b)$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) \geq \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

Strong Proportionality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathtt{att}_{\mathbf{G}}(a) = \mathtt{att}_{\mathbf{G}}(b)$,
- $\mathbf{w}(a) > \mathbf{w}(b)$,
- $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > 0$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

The Resilience principle ensures that an argument does not lose all of its strength unless its initial weight is zero.

Resilience: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a \in \mathcal{A}$, if $\mathbf{w}(a) > 0$, then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > 0$.

The Reinforcement principle ensures that the strength of an argument is sensitive to the strength of its attackers: the stronger the attacker, the greater its negative impact.

Reinforcement: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathbf{w}(a) = \mathbf{w}(b)$,
- $\mathtt{att}_{\mathbf{G}}(a) \setminus \mathtt{att}_{\mathbf{G}}(b) = \{x\}$,
- $\mathtt{att}_{\mathbf{G}}(b) \setminus \mathtt{att}_{\mathbf{G}}(a) = \{y\}$,
- $\delta^{\mathbf{S}}_{\mathbf{G}}(y) > \delta^{\mathbf{S}}_{\mathbf{G}}(x)$,
- $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > 0$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

The Counting principle states that the strength of an argument decreases as the number of serious attackers increases.

Counting: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall a, b \in \mathcal{A}$, if

- $\mathbf{w}(a) = \mathbf{w}(b)$,
- $\mathtt{att}_{\mathbf{G}}(b) = \mathtt{att}_{\mathbf{G}}(a) \cup \{x\}$ such that $x \in \mathcal{A} \setminus \mathtt{att}(a)$ and $\delta^{\mathbf{S}}_{\mathbf{G}}(x) > 0$,
- $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > 0$,

then $\delta^{\mathbf{S}}_{\mathbf{G}}(a) > \delta^{\mathbf{S}}_{\mathbf{G}}(b)$.

We now introduce a new principle, called *Locality*, which is a refined version of the classical Independence principle. It states that the value assigned to an argument depends only on its reachable subgraph. This property is particularly powerful, as it enables local reasoning and modular computation. Furthermore, it is useful for characterizing semantics.

Locality: $\forall \mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \mathtt{wAG}, \forall\, a \in \mathcal{A},\ \delta^{\mathbf{S}}_{\mathbf{G}}(a) = \delta^{\mathbf{S}}_{\mathbf{G}\downarrow \mathtt{Str}_{\mathbf{G}}(a)}(a)$.

Next, we show that independence follows from locality, and that locality follows from a combination of independence and directionality. These results apply to any semantics, whether or not they are based on evaluation methods.

Theorem 2. *Let* $\mathbf{S}$ *be a semantics.*

- *If* $\mathbf{S}$ *satisfies locality, then it satisfies independence.*

- *If* $\mathbf{S}$ *satisfies independence and directionality, then it satisfies locality.*

It is worth mentioning that directionality does not generally follow from locality. However, we show that, for semantics grounded in rational evaluation methods, satisfying locality entails satisfying directionality.

Theorem 3. *Let* $\mathbf{S}$ *be a semantics based on a rational evaluation method. If* $\mathbf{S}$ *satisfies locality, then it also satisfies directionality.*

The two previous results suggest an equivalence between the principle of locality and the combination of independence and directionality—assuming semantics based on rational evaluation methods.

Corollary 1. *Let* $\mathbf{S}$ *be a semantics based on a rational evaluation method.* $\mathbf{S}$ *satisfies locality* ***iff*** $\mathbf{S}$ *satisfies independence and directionality.*

It was shown in (Amgoud et al. 2017) that the first twelve principles are compatible, meaning they can all be satisfied simultaneously by a semantics. We show below that all thirteen principles are likewise compatible.

Theorem 4. *There exists at least one semantics that satisfies the thirteen principles.*

Finally, we show that the maximum possible strength of an argument is its initial weight, provided the semantics satisfies certain mandatory axioms: independence, directionality, maximality, neutrality, and weakening. This result aligns with the claims made by Pollock in (Pollock 2001), where he argues that the strength of an argument corresponds to its initial weight—computed in a specific way—which diminishes as the argument becomes the target of serious attacks.

Theorem 5. *Let* $\mathbf{S}$ *be a semantics that satisfies independence, directionality, maximality, neutrality and weakening. For any* $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \texttt{wAG}$*, for any* $a \in \mathcal{A}$*,* $\delta^{\mathbf{S}}_{\mathbf{G}}(a) \in [0, \mathbf{w}(a)]$.

4 Characterization Results

In this section, we characterize the classes of semantics that satisfy subsets of the presented principles. Such results are of great importance as they clarify the space of possible semantics, provide formal specifications, that is, precise descriptions of semantics's structure, behavior, and properties.

The following representation theorem establishes a one-to-one correspondence between the *equivalence* principle and the class of semantics that satisfy it. In particular, it shows that every semantics of this class can be defined by an evaluation method with a commutative aggregation function — i.e. one that treats the set of attackers as unordered, so their order has no effect on the outcome. To put it differently, semantics that satisfy the equivalence principle are defined as described in equation 1 of Definition 6.

Theorem 6. *A semantics* $\mathbf{S}$ *satisfies equivalence* ***iff*** $\mathbf{S}$ *is based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*, where the aggregation function* $\mathbf{g}$ *is commutative.*

Unsurprisingly, existing semantics that satisfy the equivalence principle all rely on an evaluation method. This is particularly the case for the $h-$*Categorizer* semantics (Besnard and Hunter 2001; Pu, andY. Zhang, and Luo 2014), the *trust-based* semantics from (da Costa Pereira, Tettamanzi, and Villata 2011), the *social* semantics from (Leite and Martins 2011), and the three gradual semantics ($\texttt{Hbs}, \texttt{Mbs}, \texttt{Cbs}$) defined in (Amgoud et al. 2017).

The following results focus on those semantics grounded in evaluation methods—i.e., exactly the ones satisfying the equivalence principle (Theorem 6)—and then pinpoint the sub-classes of these semantics that obey other principles.

The first result characterizes the semantics—defined via an evaluation method $(\mathbf{f}, \mathbf{g})$—that satisfy Maximality. Specifically, it shows that such a semantics satisfies Maximality if and only if the function $\mathbf{f}$ treats the base value $\mathbf{g}()$ as a neutral element. Importantly, this result holds regardless of whether the evaluation method $(\mathbf{f}, \mathbf{g})$ is rational.

Theorem 7. *Let* $\mathbf{S}$ *be a semantics based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*.* $\mathbf{S}$ *satisfies maximality* ***iff*** $\mathbf{f}$ *satisfies* f*-invariance (i.e.,* $\forall x \in [0,1]$*,* $\mathbf{f}(x, \mathtt{e}) = x$*).*

We likewise characterize the subclass of semantics—based on evaluation methods $(\mathbf{f}, \mathbf{g})$—that satisfy the Resilience principle. Once again, the key condition depends on a specific property of the function $\mathbf{f}$, and the characterization holds even when $(\mathbf{f}, \mathbf{g})$ is not rational.

Theorem 8. *Let* $\mathbf{S}$ *be a semantics based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*.* $\mathbf{S}$ *satisfies resilience* ***iff*** $\mathbf{f}$ *satisfies robustness (i.e.,* $\forall x \in [0,1]$*,* $\forall y \in \mathbf{I}$*,* $\mathbf{f}(x, y) > 0$ *whenever* $x > 0$*).*

The next characterisation also concerns semantics based on evaluation methods $(\mathbf{f}, g)$-whether rational or not. We show that these semantics satisfy proportionality if and only if their functions $\mathbf{f}$ are non-decreasing in their first variable.

Theorem 9. *Let* $\mathbf{S}$ *be a semantics that is based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*.* $\mathbf{S}$ *satisfies proportionality* ***iff*** $\forall x, x' \in [0,1]$*,* $\forall y \in \mathbf{I}$*, if* $x \geq x'$*, then* $\mathbf{f}(x, y) \geq \mathbf{f}(x', y)$*.*

Strong proportionality is satisfied by a semantics if and only if the latter is based on an evaluation method $(\mathbf{f}, \mathbf{g})$ whose function $\mathbf{f}$ is strictly increasing in its first variable.

Theorem 10. *Let* $\mathbf{S}$ *be a semantics that is based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*.* $\mathbf{S}$ *satisfies strong proportionality* ***iff*** $\forall x, x' \in [0,1]$*,* $\forall y \in \mathbf{I}$*, if* $x > x'$*, then* $\mathbf{f}(x, y) > \mathbf{f}(x', y)$*.*

We now introduce a class of semantics grounded in evaluation methods that satisfy specific constraints.

Definition 8. *We define* $\mathbb{S}$ *as the set of all semantics* $\mathbf{S}$ *based on an evaluation method* $(\mathbf{f}, \mathbf{g})$ *that satisfies the following properties:*

- $\forall x \in [0,1]$*,* $\mathbf{f}(x, \mathtt{e}) = x$*.*
- $\forall x, x' \in [0,1]$*,* $\forall y \in \mathbf{I}$*, if* $x > x'$*, then* $\mathbf{f}(x, y) > \mathbf{f}(x', y)$*.*
- $\forall x \in [0,1]$*,* $\forall y \in \mathbf{I}$*,* $\mathbf{f}(x, y) > 0$ *whenever* $x > 0$*.*
- $\mathbf{g}$ *is commutative.*

Building on the previous results, we establish a representation theorem that connects the principles of equivalence, maximality, proportionality, strong proportionality, and resilience to the class $\mathbb{S}$. Specifically, the theorem demonstrates a one-to-one correspondence between these principles and the class of semantics that satisfy them.

Theorem 11. *A semantics* $\mathbf{S}$ *satisfies equivalence, maximality, proportionality, strong proportionality and resilience* ***iff*** $\mathbf{S} \in \mathbb{S}$.

5 Properties of Semantics in $\mathbb{S}$

In this section, we examine additional properties of semantics within the class $\mathbb{S}$. We show that any semantics in $\mathbb{S}$ based on a rational evaluation method satisfies locality, and therefore also satisfies independence and directionality. Recall that the rationality of a method is crucial for it to characterise its corresponding semantics.

Theorem 12. *If a semantics* $\mathbf{S} \in \mathbb{S}$ *is based on a rational evaluation method, then* $\mathbf{S}$ *satisfies locality, independence and directionality.*

The above result follows mainly from the rationality condition. It is worth mentioning that the satisfaction of the three principles (locality, independence, and directionality) by a semantics that is based on an evaluation method $(\mathbf{f}, \mathbf{g})$ does not necessarily imply that $(\mathbf{f}, \mathbf{g})$ is rational. Indeed, *trust-based semantics*, as defined in (da Costa Pereira, Tettamanzi, and Villata 2011), satisfies the three principles despite being based on an evaluation method that violates rationality. It can be shown that this semantics is based on $(\mathbf{f}, \mathbf{g})$, where $\mathbf{f}(x, y) = \mathtt{min}(x, 1 - y)$ and $\mathbf{g}(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$. Consider now a graph where $\mathcal{A} = \{a, b\}$, $\mathbf{w}(a) = \mathbf{w}(b) = 1$, and $\mathcal{R} = \{(a, b), (b, a)\}$. It can be checked that the system of the two equations $S(a) = \mathbf{f}(\mathbf{w}(a), \mathbf{g}(S(b)))$ and $S(b) = \mathbf{f}(\mathbf{w}(b), \mathbf{g}(S(a)))$ has an infinite number of solutions, showing that $(\mathbf{f}, \mathbf{g})$ is not rational.

The next result concerns the principle of *anonymity*, which is satisfied by *all* existing semantics in the literature. It shows that this principle is satisfied by any semantics based on a rational evaluation method.

Theorem 13. *If a semantics* $\mathbf{S} \in \mathbb{S}$ *is based on a rational evaluation method, then* $\mathbf{S}$ *satisfies anonymity.*

The two preceding results hold under the assumption of rational evaluation methods. In contrast, the following results do not require this assumption. The first shows that neutrality is satisfied by any semantics in the set $\mathbb{S}$ whose evaluation method uses an aggregation function $\mathbf{g}$ for which 0 is a neutral element, i.e., $\mathbf{g}$ satisfies g-invariance.

Theorem 14. *For any semantics* $\mathbf{S} \in \mathbb{S}$ *based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*, if* $\mathbf{g}$ *satisfies g-invariance, then* $\mathbf{S}$ *satisfies neutrality.*

The next result concerns the principle of weakening, which captures the role of attacks. It is satisfied by any semantics in $\mathbb{S}$ whose evaluation method uses a function $\mathbf{f}$ that is strictly decreasing in its second variable and satisfies f-invariance, and a function $\mathbf{g}$ that is strictly increasing and satisfies g-invariance (see Table 1 for definitions).

Theorem 15. *Let* $\mathbf{S} \in \mathbb{S}$ *be based on an evaluation method* $(\mathbf{f}, \mathbf{g})$*. If* $\mathbf{f}$ *satisfies* f*-invariance and decreasing-S,* $\mathbf{g}$ *satisfies* g*-invariance and strict monotonicity, then* $\mathbf{S}$ *satisfies weakening.*

The principle of reinforcement is satisfied by any semantics in the set $\mathbb{S}$, provided its evaluation method uses a function $\mathbf{f}$ that is strictly decreasing in its second variable and a function $\mathbf{g}$ that is strictly increasing. Moreover, if $\mathbf{g}$ also satisfies g-invariance, then the semantics additionally satisfies the counting principle.

Theorem 16. *Let* $\mathbf{S} \in \mathbb{S}$ *be based on an evaluation method* $(\mathbf{f}, \mathbf{g})$ *such that:* $\mathbf{f}$ *satisfies decreasing-S and* $\mathbf{g}$ *satisfies strict monotonicity. The following properties hold:*

- $\mathbf{S}$ *satisfies reinforcement.*
- *If* $\mathbf{g}$ *satisfies* g*-invariance, then* $\mathbf{S}$ *satisfies counting.*

Summary: We have characterized a main class of semantics—those that satisfy the principle of equivalence. Within this class, we identified and characterized notable subclasses, each defined by additional principles such as maximality, resilience, and (strong) proportionality.

- $\mathbf{S}$ satisfies equivalence **iff** $\mathbf{S}$ is based on an evaluation method $(\mathbf{f}, \mathbf{g})$. (main class)
  - $\mathbf{S}$ satisfies maximality **iff** $\mathbf{f}$ satisfies f-invariance.
  - $\mathbf{S}$ satisfies resilience **iff** $\mathbf{f}$ satisfies robustness.
  - $\mathbf{S}$ satisfies proportionality **iff** $\mathbf{f}$ is non-decreasing.
  - $\mathbf{S}$ satisfies strong proportionality **iff** $\mathbf{f}$ satisfies increasing-F.

We have shown that semantics of the main class satisfy the remaining principles under **specific conditions**:

- Anonymity, Locality, Independence, Directionality **if** $(\mathbf{f}, \mathbf{g})$ is rational.
- Neutrality **if** $\mathbf{g}$ satisfies g-invariance.
- Weakening **if** $\mathbf{f}$ satisfies f-invariance and decreasing-S, $\mathbf{g}$ satisfies g-invariance and strict monotonicity.
- Reinforcement **if** $\mathbf{f}$ satisfies decreasing-S and $\mathbf{g}$ satisfies strict monotonicity.
- Counting **if** $\mathbf{f}$ satisfies decreasing-S, $\mathbf{g}$ satisfies strict monotonicity and $g-$invariance.

---

6 Consequences of the Results

Extension semantics, originally introduced in (Dung 1995), have been extended to deal with weighted argumentation graphs. The central idea is to define a new **defeat relation** that combines the initial attack relation with the arguments' initial weights. Classical (unweighted) semantics are then applied to the resulting flat graph. Two main approaches can be distinguished: the *contraction-based* approach and the *change-based* approach.

The contraction-based approach, studied in (Amgoud and Cayrol 2002; Bench-Capon 2003), removes attacks from an argument if it targets a stronger one. Formally, given a

weighted argumentation graph $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle \in \texttt{wAG}$, a new binary relation $\mathcal{R}'$ is defined such that for all $a, b \in \mathcal{A}$,

$$(a, b) \in \mathcal{R}' \text{ iff } (a, b) \in \mathcal{R} \text{ and } \mathbf{w}(a) \leq \mathbf{w}(b).$$

In contrast, the change-based approach, introduced in (Amgoud and Vesic 2011), reverses attacks from weaker arguments targeting stronger ones. Formally, for $a, b \in \mathcal{A}$,

$$(a, b) \in \mathcal{R}' \text{ iff } (b, a) \in \mathcal{R} \text{ and } \mathbf{w}(a) > \mathbf{w}(b).$$

Both approaches yield a new (flat) graph $\langle \mathcal{A}, \mathcal{R}' \rangle$, on which any classical extension-based semantics—including preferred, stable, grounded, and complete—can be applied. The latter are based on two key concepts: *conflict-freeness* and *defence*.

- *Conflict-freeness*: A set $\mathcal{E}$ of arguments is conflict-free iff $\nexists a, b \in \mathcal{E}$ such that $(a, b) \in \mathcal{R}'$.
- *Defence*: A set $\mathcal{E}$ of arguments defends an argument a iff for any argument b such that $(b, a) \in \mathcal{R}'$, $\exists c \in \mathcal{E}$ such that $(c, b) \in \mathcal{R}'$.

Semantics generate extensions, sets of arguments that satisfy specific conditions. Let $\mathbf{G} = \langle \mathcal{A}, \mathcal{R}' \rangle$ be the graph induced from $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle$ and let $\mathcal{E} \subseteq \mathcal{A}$.

- $\mathcal{E}$ is a *complete extension* iff it is a conflict-free set which defends all its elements and contains any argument it defends.
- $\mathcal{E}$ is a *grounded extension* iff it the subset-minimal complete extension.
- $\mathcal{E}$ is a *preferred extension* iff it is a subset-maximal complete extension.
- $\mathcal{E}$ is a *stable extension* iff $\mathcal{E}$ is conflict-free and $\forall a \in \mathcal{A} \backslash \mathcal{E}$, $\exists b \in \mathcal{E}$ such that $(b, a) \in \mathcal{R}'$.

Let $\texttt{co}, \texttt{gr}, \texttt{pr}, \texttt{st}$ denote complete, grounded, preferred, and stable respectively, and $\texttt{Ext}_x(\mathbf{G})$ the set of all extensions under semantics x, with $x \in \{\texttt{co}, \texttt{gr}, \texttt{pr}, \texttt{st}\}$.

Once extensions are computed, a strength degree is assigned to every argument by aggregating the different extensions. We recall below the assignment used in the literature (eg., (Amgoud and Ben-Naim 2016)). Let $0 < \alpha < \beta < 1$ and $x \in \{\texttt{co}, \texttt{gr}, \texttt{pr}, \texttt{st}\}$.

- $\delta^x_{\mathbf{G}}(a) = 1$ iff $a \in \bigcap_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{G})} \mathcal{E}$.
- $\delta^x_{\mathbf{G}}(a) = \beta$ iff $\exists \mathcal{E}, \mathcal{E}' \in \texttt{Ext}_x(\mathbf{G})$ s.t $a \in \mathcal{E}$ and $a \notin \mathcal{E}'$.
- $\delta^x_{\mathbf{G}}(a) = \alpha$ iff $a \notin \bigcup_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{G})} \mathcal{E}$ and $\nexists \mathcal{E} \in \texttt{Ext}_x(\mathbf{G})$ s.t. some $b \in \mathcal{E}$, $(b, a) \in \mathcal{R}'$.
- $\delta^x_{\mathbf{G}}(a) = 0$ iff $a \notin \bigcup_{\mathcal{E} \in \texttt{Ext}_x(\mathbf{G})} \mathcal{E}$ and $\exists \mathcal{E} \in \texttt{Ext}_x(\mathbf{G})$ and $\exists b \in \mathcal{E}$ s.t. $(b, a) \in \mathcal{R}'$.

We next show that the four extension-based semantics recalled above are not grounded in any evaluation method. In other words, there exists no semantics based on an evaluation method that is equivalent to any of these four classical extension-based semantics.

Theorem 17. *There is no semantics* $\mathbf{S}$ *that is based on an evaluation method such that* $\mathbf{S} \equiv \mathbf{S}'$ *and* $\mathbf{S}' \in \{\texttt{st}, \texttt{pr}, \texttt{gr}, \texttt{co}\}$.

The above negative result stems from the fact that these semantics violate the Equivalence principle. While Equivalence requires that only the initial weights and the set of attackers be taken into account, these semantics also consider the *structure of the graph*.

Remark: It is worth mentioning that extensions themselves can be computed by an equational approach as shown in (Gabbay 2012). However, their aggregation to produce strength degrees of arguments (as required in the Definition 5 of a semantics) is impossible.

7 On Explainability of Semantics in $\mathbb{S}$

Explaining the outcomes of artificial intelligence models has attracted significant attention over the past decade (Miller 2019). In the field of argumentation, numerous works have focused on explaining the evaluations produced by various semantics. Most of these efforts concentrate on extension-based semantics, aiming to answer the question: "Why is an argument (not) accepted under a given semantics?" (e.g., (Doutre, Duchatelle, and Lagasquie-Schiex 2023; Fan and Toni 2015b; Fan and Toni 2015a; Zeng et al. 2019; Liao and van der Torre 2020; Saribatur, Wallner, and Woltran 2020; Kampik, Cyras, and Alarcón 2024; Borg and Bex 2024)). Generated explanations typically intend to describe the **reasoning process** of the semantics by identifying sets of arguments or subgraphs deemed responsible for the acceptance or rejection of a particular argument.

A smaller body of work has addressed the explainability of gradual semantics (e.g., (Amgoud, Ben-Naim, and Vesic 2017; Delobelle and Villata 2019)), where the proposed approach consists of assessing how attacks influence the strength degree of each argument.

Regardless of the family of semantics—whether extension-based or gradual—most existing explanations often fail to make the underlying logic of the semantics explicit. For example, identifying a subgraph that leads to the acceptance of an argument under the stable semantics, as defined in (Dung 1995), does not clarify the two core requirements: that stable extensions must be conflict-free and must attack all arguments not included in them. Moreover, explanations are generally semantics-dependent—they are developed separately for each semantics, and few efforts aim to provide a unifying framework applicable across multiple classes of semantics. One exception is (Amgoud 2024), which introduces a general approach to post-hoc explanations. However, that work does not capture the internal logic of semantics, but rather correlations between argumentation graphs and outcomes of semantics.

We argue that the most effective way to explain the logic underlying a semantics—or, more broadly, any AI model—is through the *principles* or axioms it satisfies. Principles represent the foundational assumptions upon which a semantics is built. As such, explanations based on principles are general and apply uniformly to all semantics within

the same class. In this paper, we focus on the class $\mathbb{S}$, identified through representation theorems in the previous sections. Our aim is to define explanations induced by the principles satisfied by its semantics. More specifically, we investigate possible answers to the following question:

[Q:] *Why argument a from a weighted graph $\mathbf{G} = \langle \mathcal{A}, \mathbf{w}, \mathcal{R} \rangle$ gets value $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ under semantics $\mathbf{S} \in \mathbb{S}$?*

Let us focus on **local explanations**, which concentrate on the immediate neighborhood of an argument—that is, the argument itself and its direct attackers. Such explanations are induced by the *equivalence* principle, which states that the strength of an argument depends solely on its initial weight and the strengths of its direct attackers. These elements thus form a **unique**, sufficient and self-contained explanation.

Definition 9. *A* complete explanation *of $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ is the pair* $\mathbf{E_c}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \langle \mathbf{w}(a), \mathtt{att}_{\mathbf{G}}(a) \rangle$.

Example 1 (Cont) Consider again the example.

- $\mathbf{E_c}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \langle \mathbf{w}(a), \{b\} \rangle$.
- $\mathbf{E_c}(\delta^{\mathbf{S}}_{\mathbf{G}}(c)) = \langle \mathbf{w}(c), \{b, e\} \rangle$.

It is worth noting that complete explanations are unique, which significantly reduces the number of possible explanations compared to traditional approaches. Furthermore, its size is bounded by the number of attackers, which is highly appreciated in the XAI literature. However, a complete explanation is not necessarily subset-minimal, since worthless attackers—those with no impact on their targets as indicated by the *neutrality* principle—can still be included. Therefore, a more concise explanation would consist only of the argument's initial weight and its serious attackers.

Definition 10. *A* relevant explanation *of $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ is the pair* $\mathbf{E_r}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \langle \mathbf{w}(a), \{b \in \mathtt{att}_{\mathbf{G}}(a) \mid \delta^{\mathbf{S}}_{\mathbf{G}}(b) \neq 0\} \rangle$.

Example 1 (Cont) Assume that $\mathbf{w}(b) = 0$ and $\mathbf{w}(e) > 0$. Hence, *maximality* would lead to $\delta^{\mathbf{S}}_{\mathbf{G}}(b) = 0$ and *resilience* implies $\delta^{\mathbf{S}}_{\mathbf{G}}(e) > 0$. Consequently, we get the following relevant explanations:

- $\mathbf{E_r}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \langle \mathbf{w}(a), \emptyset \rangle$.
- $\mathbf{E_r}(\delta^{\mathbf{S}}_{\mathbf{G}}(c)) = \langle \mathbf{w}(c), \{e\} \rangle$.

A relevant explanation is considered *optimal*—meaning it includes *all* and *only* the relevant information—when the semantics satisfies the *counting* principle. This principle ensures that every serious attack contributes to reducing the strength of its target. In contrast, semantics that violate this principle, such as `Mbs` from (Amgoud et al. 2017) and the Trust-based semantics from (da Costa Pereira, Tettamanzi, and Villata 2011), base their evaluations on only a subset of attackers, typically prioritizing the strongest ones. In such cases, those influential attackers must be explicitly identified as part of the explanation. In case of the two semantics, select the strongest attackers.

Complete and relevant explanations focus only on direct attackers. However, the latter may themselves be influenced by other arguments. Hence, one may want a **global explanation** that highlights all elements that influenced the strength $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ of an argument a. Such explanation is **unique** and provided by the *locality* principle by identifying the exact portion of the graph required to evaluate the argument—namely, its *attack structure*. Any element outside this set is for sure irrelevant to $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$. Formally, it contains all arguments in the attack structure and their initial weights.

Definition 11. *A* global explanation *of $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ is the set*

$$\mathbf{E_g}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \{\langle \mathbf{w}(x), x \rangle \mid x \in \mathtt{Str}_{\mathbf{G}}(a)\}.$$

This explanation highlights *all* the arguments that impacted the strength of a.

Example 1 (Cont) Assume a semantics $\mathbf{S} \in \mathbb{S}$ that evaluates the arguments of the weighted graph $\mathbf{G}$.

- $\mathbf{E_g}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \{\langle \mathbf{w}(a), a \rangle, \langle \mathbf{w}(b), b \rangle\}$.
- $\mathbf{E_g}(\delta^{\mathbf{S}}_{\mathbf{G}}(c)) = \{\langle \mathbf{w}(b), b \rangle, \langle \mathbf{w}(c), c \rangle, \langle \mathbf{w}(e), e \rangle\}$.

We can also define its refined version by considering only serious attackers.

Definition 12. *A* relevant global explanation *of $\delta^{\mathbf{S}}_{\mathbf{G}}(a)$ is:*

$$\mathbf{E_{gr}}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) = \{\langle \mathbf{w}(x), x \rangle \mid x \in \mathtt{Str}_{\mathbf{G}}(a) \text{ and } \delta^{\mathbf{S}}_{\mathbf{G}}(x) \neq 0\}.$$

The following property shows the links between the four explanations.

Property 4. *The following inclusions hold:*

- $\mathbf{E_r}((\delta^{\mathbf{S}}_{\mathbf{G}}(a))) \subseteq \mathbf{E_c}(\delta^{\mathbf{S}}_{\mathbf{G}}(a))$,
- $\mathbf{E_{gr}}(\delta^{\mathbf{S}}_{\mathbf{G}}(a)) \subseteq \mathbf{E_g}(\delta^{\mathbf{S}}_{\mathbf{G}}(a))$.

It is worth noting that the computational complexity of generating these explanations is linear, as any algorithm only needs to examine the strengths of the argument's direct attackers. Note also that the provided explanations contain arguments, however these can be replaced by their strengths.

8 Conclusion

The paper tackles a fundamental research challenge and offers semantic characterizations through representation theorems. This contribution advances the theoretical foundations of a field that has, until now, received limited attention in the literature. Indeed, several semantics have been proposed in the literature, and a large number of principles have also been defined in order to compare formally pairs of semantics. However, there is no attempt at fully characterizing the whole classes of semantics that satisfy subsets of principles. This paper provides the first contribution in that direction. It provided some representation theorems that establish one-to-one correspondences between subsets of principles and classes of semantics. The main result shows a bijection between the principle of equivalence and the large class of semantics that are defined using evaluation methods. This result allows to clearly distinguish extension semantics from gradual ones. Indeed, all existing gradual semantics satisfy equivalence, thus argument strength depends only on the argument's initial weight and the strengths of its direct attackers. Extension semantics violate the principle because they also take into account the topology of the graph.

The paper provided also some representation theorems delimiting the classes of evaluation methods that lead to the satisfaction of various principles, including maximality, resilience and (strong) proportionality.

This work lends itself to a number of developments, the most pressing of which is the characterization of rational evaluation methods. Another important direction is the development of representation theorems that fully characterize the class of extension-based semantics. We also aim to further investigate the principle-based approach to explainability, and to compare it with the two prevailing alternatives in the literature: the model-based and post-hoc approaches.

Acknowledgments

This work was supported by the AI Interdisciplinary Institute ANITI, funded by the French program "AI-Cluster".

References

Amgoud, L., and Ben-Naim, J. 2013. Ranking-based semantics for argumentation frameworks. In *SUM'13*, 134–147.

Amgoud, L., and Ben-Naim, J. 2016. Axiomatic foundations of acceptability semantics. In *KR-16*, 2–11.

Amgoud, L., and Cayrol, C. 2002. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence* 34(1-3):197–215.

Amgoud, L., and Doder, D. 2019. Gradual semantics accounting for varied-strength attacks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, 1270–1278. International Foundation for Autonomous Agents and Multiagent Systems.

Amgoud, L., and Prade, H. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence* 173:413–436.

Amgoud, L., and Vesic, S. 2011. A new approach for preference-based argumentation frameworks. *Annals of Mathematics and Artificial Intelligence* 63(2):149–183.

Amgoud, L.; Ben-Naim, J.; and Vesic, S. 2017. Measuring the intensity of attacks in argumentation graphs with shapley value. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 63–69.

Amgoud, L.; Ben-Naim, J.; Doder, D.; and Vesic, S. 2017. Acceptability semantics for weighted argumentation frameworks. In *IJCAI-17*, 56–62.

Amgoud, L.; Doder, D.; and Vesic, S. 2022. Evaluation of argument strength in attack graphs: Foundations and semantics. *Artificial Intelligence* 302:103607.

Amgoud, L. 2024. Post-hoc explanation of extension semantics. In *27th European Conference on Artificial Intelligence, ECAI*, volume 392, 3276–3283.

Baroni, P., and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence* 171(10-15):675–700.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation, Vol. 1*. College Publications.

Baroni, P.; Giacomin, M.; and Guida, G. 2005. Scc-recursiveness: a general schema for argumentation semantics. *Artif. Intell.* 168(1-2):162–210.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In *Proceedings of the Thirty-Second Conference on Artificial Intelligence, AAAI*.

Bench-Capon, T. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13(3):429–448.

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

Bonzon, E.; Delobelle, J.; Konieczny, S.; and Maudet, N. 2016. A comparative study of ranking-based semantics for abstract argumentation. In *AAAI-16*.

Borg, A., and Bex, F. 2024. Minimality, necessity and sufficiency for argumentation and explanation. *International Journal of Approximate Reasoning* 168:109143.

Caminada, M. 2006. Semi-stable semantics. In *Proceedings of the 1st International Conference on Computational Models of Argument, COMMA'06*, 121–130.

Cayrol, C., and Lagasquie, M. 2005. Graduality in argumentation. *J. of Artificial Intelligence Research* 23:245–297.

da Costa Pereira, C.; Tettamanzi, A.; and Villata, S. 2011. Changing one's mind: Erase or rewind? In *IJCAI'11*, 164–171.

Delobelle, J., and Villata, S. 2019. Interpretability of gradual semantics in abstract argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 15th European Conference, ECSQARU*, volume 11726 of *Lecture Notes in Computer Science*, 27–38. Springer.

Dimopoulos, Y.; Mailly, J.; and Moraitis, P. 2019. Argumentation-based negotiation with incomplete opponent profiles. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, 1252–1260.

Doutre, S.; Duchatelle, T.; and Lagasquie-Schiex, M. 2023. Classes of explanations for the verification problem in abstract argumentation. In *17èmes Journées d'Intelligence Artificielle Fondamentale, JIAF 2023*, 124–134.

Dung, P.; Mancarella, P.; and Toni, F. 2007. Computing ideal skeptical argumentation. *Artificial Intelligence* 171:642–674.

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

Egilmez, S.; Martins, J.; and Leite, J. 2013. Extending social abstract argumentation with votes on attacks. In *Theory and Applications of Formal Argumentation - Second International Workshop, TAFA-2013*, 16–31.

Fan, X., and Toni, F. 2015a. On computing explanations in argumentation. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence, AAAI*, 1496–1502. AAAI Press.

Fan, X., and Toni, F. 2015b. On explanations for non-acceptable arguments. In *Third International Workshop on Theory and Applications of Formal Argumentation, TAFA*, volume 9524 of *Lecture Notes in Computer Science*, 112–127. Springer.

Gabbay, D. M. 2012. Equational approach to argumentation networks. *Argument & Computation* 3(2-3):87–142.

Kampik, T.; Cyras, K.; and Alarcón, J. R. 2024. Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning* 164:109066.

Leite, J., and Martins, J. 2011. Social abstract argumentation. In *IJCAI'11*, 2287–2292.

Liao, B., and van der Torre, L. 2020. Explanation semantics for abstract argumentation. In *Computational Models of Argument - Proceedings of COMMA*, volume 326, 271–282. IOS Press.

Libman, A.; Oren, N.; and Yun, B. 2024. Abstract weighted based gradual semantics in argumentation theory. *arXiv*.

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

Pollock, J. L. 2001. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence* 133(1):233–282.

Potyka, N. 2019. Open-mindedness of gradual argumentation semantics. In *In Proceedings of Scalable Uncertainty Management - 13th International Conference, SUM*, volume 11940 of *Lecture Notes in Computer Science*, 236–249. Springer.

Pu, F.; andY. Zhang, J. L.; and Luo, G. 2014. Argument ranking with categoriser function. In *Proceedings of the 7th International Knowledge Science, Engineering and Management Conference, KSEM'14*, 290–301.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *KR'16*, 63–73.

Saribatur, Z. G.; Wallner, J. P.; and Woltran, S. 2020. Explaining non-acceptability in abstract argumentation. In *24th European Conference on Artificial Intelligence, ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 881–888.

Simari, G.; Giacomin, M.; Gabbay, D.; and Thimm, M., eds. 2021. *Handbook of Formal Argumentation, Vol. 2*. College Publications.

Zeng, Z.; Miao, C.; Leung, C.; Shen, Z.; and Chin, J. J. 2019. Computing argumentative explanations in bipolar argumentation frameworks. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):10079–10080.