Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis

Timotheus Kampik [a,b,*], Nico Potyka [c], Xiang Yin [d], Kristijonas Čyras [e], Francesca Toni [d]

[a] *Umeå University, Sweden*
[b] *SAP Signavio, Germany*
[c] *Cardiff University, UK*
[d] *Imperial College London, UK*
[e] *Ericsson, USA*

A B S T R A C T

We present a principle-based analysis of *contribution functions* for quantitative bipolar argumentation graphs that quantify the contribution of one argument to another. The introduced principles formalise the intuitions underlying different contribution functions as well as expectations one would have regarding the behaviour of contribution functions in general. As none of the covered contribution functions satisfies all principles, our analysis can serve as a tool that enables the selection of the most suitable function based on the requirements of a given use case.

1. Introduction

Formal argumentation has emerged as a vibrant research area within the field of artificial intelligence. In particular, the dialectical nature of formal argumentation is considered a promising facilitator of joint human-machine reasoning and decision-making, as well as a potential bridge between subsymbolic and symbolic AI [1]. In formal argumentation, arguments and their relations are represented as directed graphs, in which nodes are arguments and edges are argument relationships (typically: attack or support). From these *argumentation graphs*, inferences about the acceptability statuses or strengths of arguments are drawn. One formal argumentation approach that is gaining increased research attention is Quantitative Bipolar Argumentation (QBA). In QBA, (typically numerical) weights—so-called *initial strengths*—are assigned to arguments, and arguments are connected by a support and an attack relation. Hence, arguments directly connected to a node through the node's incoming edges can be referred to as *attackers* and *supporters* (depending on the relation). Given a Quantitative Bipolar Argumentation Graph (QBAG), an *argumentation semantics* then infers the arguments' *final strengths*; intuitively, an argument's attackers tend to decrease its final strength, whereas supporters tend to increase it. Nascent applications of formal argumentation in general and QBA in particular are often related to *explainability* [2,3], e.g., in the context of explainable recommender systems [4], review aggregations [5] or machine learning models like random forests [6] or neural networks [7]. In order to utilise QBA as a facilitator of explainability, it is crucial to develop a rigorous understanding of the influence of one argument on the final strength of another one. This follows the above-mentioned intuition of argument influence

---

1.1: G.

1.2: Final strength of a, given initial strength of e.

Fig. 1. QBAG G (Example 1.1). A node labelled $\begin{smallmatrix} \mathsf{x}:(i) \\ \mathbf{f} \end{smallmatrix}$ represents argument x with initial strength $\tau(\mathsf{x}) = i$ and final strength $\sigma(\mathsf{x}) = \mathbf{f}$. Edges labelled + and −, respectively, represent supports and attacks.

that (direct or indirect) attackers or supporters exercise and reflects the ideas of feature attributions in machine learning (see [8–12] as examples), as well as of contributions of agents towards a joint objective in cooperative game theory. Initial steps towards this objective have been undertaken by nascent/preliminary research [13,14] that introduced so-called *contribution functions* quantifying the influence of one argument in a QBAG on another one.

Yet, a comprehensive theoretical study of contribution functions is missing. Our paper aims to close this gap in the literature by introducing a set of principles that a contribution function should satisfy, and by conducting a principle-based analysis for *acyclic* QBAGs. We claim that this limitation is well-motivated: in many applications of QBAGs there is a clear hierarchical or temporal structure to the arguments [15,5,16,17], which helps avoid cycles or allows for removing them on the meta-level. The introduced principles reflect the ideas that underlie the different contribution functions, as well as some intuitive, common-sense assumptions (that are, however, not always satisfied).

Let us introduce a simple example that provides an intuitive understanding of the paper's core contribution.

Example 1.1. Consider the QBAG G in Fig. 1.1 (the interpretation of the figure is explained in the caption). The initial strengths of G's arguments are determined by DFQuAD semantics [18], according to the following intuition:

- Final strengths are propagated through the graph following a topological ordering of the nodes.
- Given an argument a and its direct attackers A and supporters S, the final strengths of A and S are aggregated as follows:

$$f(A, S) := \Pi_{\mathsf{b} \in A}(1 - (\sigma(\mathsf{b})) - \Pi_{\mathsf{c} \in S}(1 - (\sigma(\mathsf{c})),$$

where $\sigma(\mathsf{b})$ and $\sigma(\mathsf{c})$ denote the final strengths of b and c, respectively.
- Given the aggregated score $f(A, S)$, a is then updated as follows, given its initial strength $\tau(\mathsf{a})$:

$$g(\tau(\mathsf{a}), f(A, S)) := \tau(\mathsf{a}) - \tau(\mathsf{a}) \times max\{0, -f(A, S)\} + (1 - \tau(\mathsf{a})) \times max\{0, f(A, S)\}.$$

We are interested in explaining how different arguments ("contributors") contribute to the final strength of a. As we are primarily interested in a and its final strength, we refer to a as the *topic argument*. Obviously, the contribution of supporter b should be positive (or possibly zero) and the contributions of attackers c and d should be negative (or zero): in these cases, the direction (with respect to 0) of the contribution is clear. In contrast, it is not at all clear how e contributes to a. Considering this more intricate case, we may specify our expectation regarding the behaviour of the function that determines the contribution (in this case: of e to the final strength of a) as a *contribution function principle*. For example, one may stipulate that:

1. The contribution should be negative if (and only if) removing e from G increases the final strength of a and positive if (and only if) the removal decreases it; we call this desideratum *counterfactuality*.
2. If the contribution is negative/positive, then marginally increasing the initial strength of e should decrease/increase the strength of a; we call this desideratum *faithfulness*.

The QBAG in Fig. 1.2 illustrates that these two ideas can be in conflict. To explain this, we plotted the final strength of a (y-axis) against the initial strength of e (x-axis) in the function graph on the right. Since e's initial strength is 0.5 in the original QBAG, a's final strength is about 0.37. When removing e, the final strength of b, c and d will just be their initial strength, which is 0. Since arguments with strength 0 cannot affect other arguments (given DFQuAD semantics), a's final strength will also be its initial strength, which is 0.5. Hence, since a's strength increases from 0.37 to 0.5 when removing e, e should have a negative contribution according to *counterfactuality*. Now, if e's contribution is negative, then *faithfulness* demands that marginally increasing the strength of e in the

original example must decrease the strength of a. However, as we can see in the strength plot in Fig. 1.2, increasing the initial strength of e will actually increase the final strength of a. Hence, it is impossible to satisfy both *counterfactuality* and *faithfulness* simultaneously in this example.

The previous example highlights that a principle-based analysis is required to shed light on contribution functions in a rigorous manner. After introducing some preliminaries in Section 2, we present four contribution functions for QBA (Section 3)[1]:

Removal. The contribution of one argument to another is determined by the effect that the removal of the former argument has on the final strength of the latter.

Intrinsic removal. The contribution of one argument to another is determined by the effect that the removal of the former argument, *from a QBAG from which all attack and support relationships that target the contributing argument have already been removed*, has on the final strength of the latter.

Shapley values. The contribution of one argument to another is determined by the contribution of the former argument to the final strength of the latter when interpreting the problem as a coalitional game and applying Shapely values [20] to it.

Gradient. The contribution of one argument to another is determined by the effect that a marginal change to the initial strength of the former argument has on the final strength of the latter.

We then introduce a set of principles (Section 4), following intuitions underlying some of the contribution functions, as well as common sense-based desiderata:

Directionality. Many quantitative argumentation semantics satisfy a *directionality* principle [21]: argument a can influence argument b only if there is a directed path from a to b. Consequently, under these semantics, if there is no path from a to b, then a's contribution to b should be 0.

(Quantitative) contribution existence. If our topic argument's initial strength does not equal its final strength then there exists another argument whose contribution to the topic argument is not 0. In the quantitative case, the sum of the contributions of all other arguments (to the topic argument) should account for the delta between the topic's final and initial strength (where negative contributions give rise to a decrease in the strength of the topic argument).

(Quantitative) local faithfulness. If an argument's contribution is negative/positive, then marginally increasing the initial strength of the contributor should increase/decrease the strength of the topic argument. In the quantitative case, the contributor's contribution to the topic argument should be proportional to the change in strength.

(Quantitative) counterfactuality. The contribution should be negative if (and only if) removing the contributor from the QBAG increases the final strength of the topic argument and positive if (and only if) the removal decreases it. In the quantitative case, the delta between the topic argument's final strength in the initial QBAG and the topic's argument final strength with the contributor removed should be equal to the contribution of the contributor to the topic argument.

Subsequently, we analyse principle satisfaction of the four contribution functions with respect to five argumentation semantics (Section 5). The results are summarised in Table 1. These results are restricted to acyclic QBAGs (the focus of our study, as pointed out earlier). Note that this special case is worth studying because in many nascent applications of QBAGs there is a clear hierarchical or temporal structure to the arguments that prevents the existence of cycles [15,5,16,17,6,7]. Note also that many ideas that we consider here can be extended to general QBAGs with cycles, but the existence of cycles causes some additional technical difficulties (given that strength values may not converge in particular situations) whose discussion would obscure our main points. Hence, we leave the exploration of contribution functions for general QBAGs to future work.

We also provide some minor results and conjectures that we produced as a side-effect of our work (Section 6). To highlight practical perspectives, we position our results in the context of an application example (Section 7) Finally, we conclude the paper by discussing our results in the light of related research (Section 8).

2. Preliminaries

This section introduces the formal preliminaries of our work. Let $\mathbb{I}$ be a real interval of *strength values*. Here, we will focus on the unit interval $\mathbb{I} = [0, 1]$. A *quantitative bipolar argumentation graph* contains a set of arguments related by binary *attack* and *support* relations, and assigns an *initial strength* in $\mathbb{I}$ to the arguments. The (initial) strength can be thought of as initial credence in, or importance of, arguments. Typically, the greater the strength in $\mathbb{I}$, the more credible or important the argument is.

Definition 2.1 (*Quantitative Bipolar Argumentation Graph (QBAG)* [22,23]). A *Quantitative Bipolar Argumentation Graph (QBAG)* is a quadruple $(Args, \tau, Att, Supp)$ consisting of a set of arguments $Args$, an *attack* relation $Att \subseteq Args \times Args$, a *support* relation $Supp \subseteq Args \times Args$ such that $Att \cap Supp = \emptyset$, and a total function $\tau : Args \rightarrow \mathbb{I}$ that assigns an *initial strength* $\tau(x)$ to every $x \in Args$.

---

[1] We first introduced these contribution functions in a workshop paper [13]; here, we fix some technicalities for the Shapley value-based contribution function. The first two contribution functions have been introduced first to abstract gradual argumentation in a roughly analogous manner in [19].

Table 1

Satisfaction and violation of contribution principles (defined in Section 5), given a semantics σ (defined in Section 2) and a contribution function Ctrb (defined in Section 3).

Ctrb / σ	QE	DFQuAD	SD-DFQuAD	EB	EBT
Contribution Existence					
$Ctrb^{\mathcal{R}}$	✓	✗	✗	✓	✗
$Ctrb^{\mathcal{R}'}$	✓	✗	✗	✓	✗
$Ctrb^{\mathcal{S}}$	✓	✓	✓	✓	✓
$Ctrb^{\partial}$	✓	✗	✗	✓	✗
Quantitative Contribution Existence					
$Ctrb^{\mathcal{R}}$	✗	✗	✗	✗	✗
$Ctrb^{\mathcal{R}'}$	✗	✗	✗	✗	✗
$Ctrb^{\mathcal{S}}$	✓	✓	✓	✓	✓
$Ctrb^{\partial}$	✗	✗	✗	✗	✗
Directionality					
$Ctrb^{\mathcal{R}}$	✓	✓	✓	✓	✓
$Ctrb^{\mathcal{R}'}$	✓	✓	✓	✓	✓
$Ctrb^{\mathcal{S}}$	✓	✓	✓	✓	✓
$Ctrb^{\partial}$	✓	✓	✓	✓	✓
(Quantiative) Local Faithfulness					
$Ctrb^{\mathcal{R}}$	✗	✗	✗	✗	✗
$Ctrb^{\mathcal{R}'}$	✗	✗	✗	✗	✗
$Ctrb^{\mathcal{S}}$	✗	✗	✗	✗	✗
$Ctrb^{\partial}$	✓	✓	✓	✓	✓
(Quantitative) Counterfactuality					
$Ctrb^{\mathcal{R}}$	✓	✓	✓	✓	✓
$Ctrb^{\mathcal{R}'}$	✗	✗	✗	✗	✗
$Ctrb^{\mathcal{S}}$	✗	✗	✗	✗	✗
$Ctrb^{\partial}$	✗	✗	✗	✗	✗

Given a QBAG $(Args, \tau, Att, Supp)$ and $a, b \in Args$, a path from a to b is a path from a to b in the directed graph $(Args, Att \cup Supp)$. We say that a QBAG $(Args, \tau, Att, Supp)$ is acyclic iff there exists no $a \in Args$ for which there exists a path from a to a. Henceforth, we assume as given a fixed acyclic QBAG $G = (Args, \tau, Att, Supp)$ with a finite number of arguments.

Given $a, x \in Args$, the set $Att_G(x) := \{z \in Args \mid (z, x) \in Att\}$ is the set of attackers of x and each $z \in Att_G(x)$ is a (direct) *attacker* of x (z *attacks* x); the set $Supp_G(x) := \{y \in Args \mid (y, x) \in Supp\}$ is the set of supporters of x and each $y \in Supp_G(x)$ is a (direct) *supporter* of x (y *supports* x). We may drop the subscript G when the context is clear. x is an indirect attacker of a iff x is a supporter of some direct or indirect attacker of a or x attacks some direct or indirect supporter of a. x is an indirect supporter of a iff x attacks some direct or indirect attacker of a or x supports some direct or indirect supporter of a. Given $G = (Args, \tau, Att, Supp)$ and $A \subseteq Args$, we define the QBAG *restriction* of G to A, denoted by $G \downarrow_A$, as follows: $G \downarrow_A := (A, \tau \cap (A \times \mathbb{I}), Att \cap (A \times A), Supp \cap (A \times A))$.

Reasoning in QBAGs amounts to updating the initial strengths of arguments to their final strengths, taking into account the strengths of attackers and supporters. Specifically, given a QBAG, a strength function assigns final strengths to arguments in the QBAG. Different ways of defining a strength function are called gradual semantics.

Definition 2.2 (*Gradual Semantics and Strength Functions [23,22]*). A *gradual semantics* σ defines for $G = (Args, \tau, Att, Supp)$ a (possibly partial) *strength function* $\sigma_G : Args \to \mathbb{I} \cup \{\bot\}$ that assigns the *final strength* $\sigma_G(x)$ to each $x \in Args$, where $\bot$ is a reserved symbol meaning 'undefined'.

Table 2

Common aggregation α and influence ι functions [22, pp. 1724 Table 1; with a fixed typo for p-Max(k)]. Intuitively, $s \in [0,1]^n$ is a strength vector (associating each argument with its current strength), $v \in \{-1,0,1\}^n$ is a relationship vector indicating which arguments attack (-1), support (1) or are in no relationship to (0) the argument of interest, and w is an initial strength.

Aggregation Functions		
Sum	$\alpha_v^\Sigma : [0,1]^n \to \mathbb{R}$	$\alpha_v^\Sigma(s) = \sum_{i=1}^n v_i \times s_i$
Product	$\alpha_v^\Pi : [0,1]^n \to [-1,1]$	$\alpha_v^\Pi(s) = \prod_{i : v_i = -1}(1 - s_i) - \prod_{i : v_i = 1}(1 - s_i)$
Top	$\alpha_v^{max} : [0,1]^n \to [-1,1]$	$\alpha_v^{max}(s) = M_v(s) - M_{-v}(s),$
		where $M_v(s) = max\{0, v_1 \times s_1, \ldots, v_n \times s_n\}$

Influence Functions		
Linear(k)	$\iota_w^l : [-k,k] \to [0,1]$	$\iota_w^l(s) = w - \frac{w}{k} \times max\{0,-s\} + \frac{1-w}{k} \times max\{0,s\}$
Euler-based	$\iota_w^e : \mathbb{R} \to [w^2, 1]$	$\iota_w^e(s) = 1 - \frac{1-w^2}{1 + w \times e^s}$
p-Max(k)	$\iota_w^p : \mathbb{R} \to [0,1]$	$\iota_w^p(s) = w - w \times h(-\frac{s}{k}) + (1-w) \times h(\frac{s}{k}),$
for $p \in \mathbb{N}$		where $h(x) = \frac{max\{0,x\}^p}{1 + max\{0,x\}^p}$

Table 3

Examples of gradual semantics.

Semantics	Aggregation	Influence
QuadraticEnergyModel (QE)	Sum	2-Max(1)
DFQuADModel (DFQuAD)	Product	Linear(1)
SquaredDFQuADModel (SD-DFQuAD)	Product	1-Max(1)
EulerBasedModel (EB)	Sum	EulerBased
EulerBasedTopModel (EBT)	Top	EulerBased

If G is clear from the context, we will drop the subscript $_G$ and denote the strength function by σ.

Many QBAG semantics from the literature including [24,18,21,25,26] can be seen as *modular semantics* [27]. Modular semantics define the final strength values of arguments by an iterative procedure that works as follows:

Initialization: Let $(x_1, \ldots, x_n)$ be an arbitrary but fixed ordering of the arguments in *Args*. Define a vector $s^{(0)} \in \mathbb{R}^n$ of initial strength values of the arguments by letting $s_i^{(0)} = \tau(x_i)$.

Update: Given the vector $s^{(i)}$ from the previous iteration, apply an update function $u_G : \mathbb{R}^n \to \mathbb{R}^n$ to obtain $s^{(i+1)} = u_G(s^{(i)})$.

$u_G(s^{(i)})$ will adapt the strength of every argument x_i in two steps.

Aggregation. First an aggregation function α combines the current strength values of x_i's supporters and attackers with respect to $s^{(i)}$ to a single value that reflects how x_i should be adapted. For representing attack, support, and lack of both attack and support, we may utilise a vector $v \in \{-1,0,1\}^n$ in which each attacker of x_i is represented by -1, each supporter by 1, and each argument neither attacking nor supporting x_i by 0.

Influence. Then an influence function ι will determine a new strength value for x_i based on the aggregate and x_i's initial strength $w = \tau(x_i)$.

Table 2 gives a list of common influence and aggregation functions. Table 3 shows some examples of gradual semantics. With abuse of notation, we may denote the evaluation of the aggregation function of an argument a given a QBAG G $= (Args, \tau, Att, Supp)$ (with a $\in Args$) by $\alpha_G(a)$, i.e., $\alpha_G(a) := \alpha(a_0, \ldots, a_n)$, where $a_0, \ldots, a_n$ are the supporters and attackers of a in G. Analogously, we may denote the evaluation of a's influence function ι given by G by $\iota_G(a)$.

Formally, the final strength $\sigma_G(x_i)$ of the i-th argument is the i-th component s_i of the limit vector $s = \lim_{i \to \infty} s^{(i)}$. As demonstrated in [27], the limit may not exist in cyclic graphs. However, as we focus on acyclic graphs here, the limit is guaranteed to exist and can, in fact, be computed in linear time by a simple forward propagation procedure [22]. Roughly speaking, it works by first computing a topological ordering of the arguments. The strength values are then computed following this order. The final strength of the arguments without parents (i.e. either attackers or supporters) is just their initial strength. When the next argument in the topological order is selected, all strength values of its parents are already fixed, so its final strength can be computed by a single application of the update function.

As we can easily see, the semantics in Table 3 are modular semantics. For example, using the Sum aggregation function and the 2-Max(1) influence function yields the Quadratic Energy (QE) semantics. For a step-by-step walk-through of final strength computation using a modular semantics, we refer to Example 1.1.

To facilitate the analysis and comparison of gradual semantics, formal principles have been defined. In the context of this paper, the *directionality* and *stability* principles [21] are of particular relevance. Directionality stipulates that removing arguments that cannot reach an argument of interest from the QBAG must not affect the latter argument's final strength. When defining the principles we

assume an arbitrary (acyclic and finite) QBAG $G = (Args, \tau, Att, Supp)$ and $a \in Args$. Directionality states, intuitively, that an argument a can only be influenced by another argument x if there is a directed path from x to a.

Principle 2.1 *(Directionality (Gradual Semantics))*. A gradual semantics σ satisfies the *directionality principle* iff for all $G = (Args, \tau, Att, Supp)$ and $G' = (Args, \tau, Att', Supp')$ such that $Att \cup Supp = Att' \cup Supp' \cup (x, y)$ for some $x \in Args$, $y \in Args \setminus \{a\}$ such that there is no directed path from y to a, we have $\sigma_G(a) = \sigma_{G'}(a)$.

Stability stipulates that an argument's final strength equals its initial strength in the absence of supporters and attackers.

Principle 2.2 *(Stability)*. A QBAG semantics σ satisfies the *stability principle* iff $Att_G(a) = Supp_G(a) = \emptyset$ implies that $\sigma(a) = \tau(a)$.

All semantics in Table 3 satisfy directionality and stability [21,27,28].

3. Argument contributions

To determine how much an argument $x \in Args$ *contributes* to the final strength $\sigma_G(a)$ of a *topic* argument $a \in Args$ in a given QBAG G (given a semantics σ), we make use of contribution functions $Ctrb_{\sigma,G,a} : Args \to \mathbb{R} \cup \{\bot\}$. We drop the subscripts σ and G where the context is clear. Intuitively, $Ctrb_a(x)$ measures the influence of x on the topic argument a. The exact meaning of $Ctrb_a(x)$ depends on the contribution function. We will investigate the following contribution functions from [13].

Removal-based contribution: Intuitively, we ask: how does the removal of the contributor affect the topic argument's final strength? For this, we determine the difference between the final strength of the topic argument in G and its final strength in G with the contributor removed, i.e. we compute $\sigma_G(a) - \sigma_{G'}(a)$, where G' is obtained from G by removing x:

$$Ctrb_a^{\mathscr{R}}(x) = \sigma_G(a) - \sigma_{G\downarrow_{Args \setminus \{x\}}}(a) \tag{1}$$

Removal-based contribution without indirection: Continuing with the intuition above, we may require that the contribution of one argument to another one should be based on the former's *intrinsic* (initial) strength and not on boosts received by (direct or indirect) attackers and supporters [19]. To reflect this criticism of the removal-based contribution function, we may first remove all attacks and supports that have the contributor as a target before determining the difference (assuming that our semantics is *directional*, i.e. that strength updates are propagated only in the direction represented by the support and attack relations); i.e., we compute $\sigma_{G^-}(a) - \sigma_{G'}(a)$ where G^- is obtained from G by removing direct relations to x and G' is obtained from G by removing x entirely:

$$Ctrb_a^{\mathscr{R}'}(x) = \sigma_{(Args, \tau, Att \setminus \{(y,x) | (y,x) \in Att\}), Supp \setminus \{(y,x) | (y,x) \in Supp\})}(a) - \sigma_{G\downarrow_{Args \setminus \{x\}}}(a) \tag{2}$$

Shapley-based contribution: Yet more elaborately, we can see the difference between a topic argument's final and initial strengths as a value achieved by a coalition of all other arguments in our QBAG, which we can then consider *players* in the game theoretical sense. Then, we can use Shapley values [29] as a well-established approach to quantify the argument contributions.[2] Intuitively, we determine a weighted average of the topic argument's final strength difference that is achieved by adding the contributor to all possible subgraphs of our QBAG that already contain the topic argument:

$$Ctrb_a^{\mathscr{S}}(x) =$$
$$\sum_{X \subseteq Args \setminus \{x,a\}} \frac{|X|! \cdot (|Args \setminus \{a\}| - |X| - 1)!}{|Args \setminus \{a\}|!} \left(\sigma_{G\downarrow_{Args \setminus X}}(a) - \sigma_{G\downarrow_{Args \setminus (X \cup \{x\})}}(a) \right) \tag{3}$$

Gradient-based contribution: The last contribution function is based on the *gradient* from real analysis [30]: intuitively, the gradient is a vector (whose components are partial derivatives) that measures how a small change in an input parameter of a function affects the output. In our case, the function of interest maps the initial strength values of arguments to the final strength of the topic argument. Formally, we consider a function $f_a : \mathbb{R}^n \to \mathbb{R}$ that maps a vector $s^{(0)} \in \mathbb{R}^n$ of initial strength values to the final strength $f_a(s^{(0)})$ of a.

Since we assume that G is acyclic, we can construct an explicit representation of f_a recursively. Intuitively, the representation of f_a corresponds to a long composition of influence and aggregation functions of the semantics applied to the initial strength values in $s^{(0)}$. Let i be the index of the topic argument a in the topological ordering $\{x_1, \ldots, x_n\}$ associated with G. If a does not have any parents, then its final strength is just $s_i^{(0)}$ by *stability* (Principle 2.2) of the semantics. In this case, f_a is just the constant function that always returns $s_i^{(0)}$. If, on the other hand, a has the set of parents $S = \{x_{i_1}, \ldots, x_{i_{k_a}}\}$, we create placeholders $T_{i_1}, \ldots, T_{i_{k_a}}$ for the explicit representation of the strength value computation of the parents and let

---

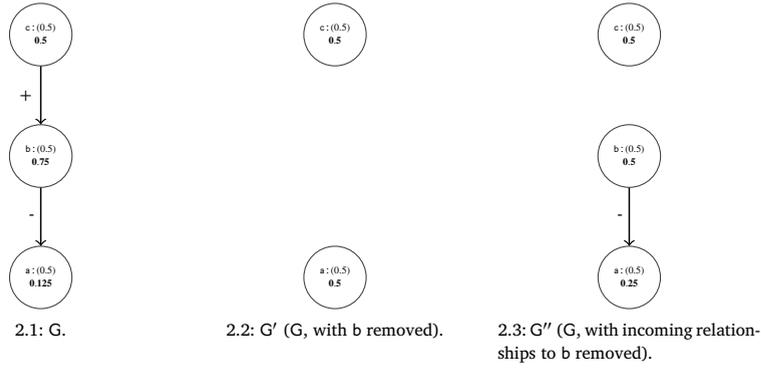[2] Note that we fixed some technicalities in the Shapley value-based contribution function initially presented in [13].

2.1: G. 2.2: G′ (G, with b removed). 2.3: G″ (G, with incoming relation-ships to b removed).

Fig. 2. A QBAG G and its updates, with b (G′) and incoming relationships to b removed (G″).

$$f_{\mathsf{a}} = \iota_{w_{\mathsf{a}}}(\alpha_{\upsilon_{\mathsf{a}}}(T_{i_1}, \dots, T_{i_{k_{\mathsf{a}}}})), \tag{4}$$

where $w_{\mathsf{a}} = \tau(\mathsf{a})$ is a's initial strength and υ_{a} is the parent vector that contains $-1/0/1$ at the i-th component if x_i attacks/is unrelated to/supports a. We now have to replace $T_{i_1}, \dots, T_{i_{k_{\mathsf{a}}}}$ with explicit function representations of the strength of $\mathsf{x}_{i_1}, \dots, \mathsf{x}_{i_{k_{\mathsf{a}}}}$ as well. This can be done exactly as for a. In the process, we may create new placeholders, but since G is acyclic, the recursion will stop after a finite number of steps. A naive implementation of the recursive scheme may compute the same template multiple times. This can be avoided by storing templates (there can be at most $|Args|$) or using a dynamic programming approach.

Let us note that, in the process of constructing the explicit function representation of f_{a}, we can only add strength values $s_i^{(0)}$ of arguments x_i such that there is a directed path from x_i to a in G. This is in line with the directionality property (Principle 2.1) of gradual semantics. We emphasise this observation as it will be useful in the following.

Observation 3.1. $f_{\mathsf{a}}(s^{(0)})$ depends on $s_i^{(0)}$ if and only if there is a directed path from x_i to a.

The gradient-based contribution function is now defined as

$$\mathsf{Ctrb}_{\mathsf{a}}^{\partial}(\mathsf{x}) = \frac{\partial f_{\mathsf{a}}}{\partial \tau(\mathsf{x})}\left(\tau(\mathsf{x}_1), \dots, \tau(\mathsf{x}_n)\right). \tag{5}$$

With abuse of notation, we may also denote the partial derivative $\frac{\partial f_{\mathsf{a}}}{\partial \tau(\mathsf{x})}$ as $\frac{\partial \sigma(\mathsf{a})}{\partial \tau(\mathsf{x})}$.

Irrespective of the specific method for determining an argument's contribution to the final strength of some topic argument in a given QBAG G, we use a generic *contribution function* $\mathsf{Ctrb} : Args \times Args \to \mathbb{R}$ defined by $\mathsf{Ctrb}(\mathsf{x}, \mathsf{a}) = \mathsf{Ctrb}_{\mathsf{a}}(\mathsf{x})$. For reference, we use $\mathsf{Ctrb}^{\mathscr{R}}, \mathsf{Ctrb}^{\mathscr{R}'}, \mathsf{Ctrb}^{\mathscr{S}}, \mathsf{Ctrb}^{\partial}$ to denote the contribution functions given, respectively, in Equations (1)–(3) and (5).

Let us illustrate the intuitions behind the contribution functions with an example.

Example 3.1. Consider the QBAF G in Fig. 2.1. We use DFQuAD semantics and are interested in quantifying the contribution of b to a. The removal-based contribution $\mathsf{Ctrb}_{\mathsf{a}}^{\mathscr{R}}(\mathsf{b}) = -0.375$ is obtained by subtracting the final strength of a in G′ from the final strength of a in G. However, one may argue that b's contribution to a is not entirely intrinsic to b, as b benefits from its supporter c, following a contribution quantification approach first introduced to abstract gradual argumentation [19]. To determine the intrinsic contribution, we may first remove all attack and support relationships to b to then determine the difference of a's final strengths with and without b, as seen in Figs. 2.3 and 2.2, respectively: $\mathsf{Ctrb}_{\mathsf{a}}^{\mathscr{R}'}(\mathsf{b}) = -0.25$. Still, one may argue that this comes at the cost of the simpler intuition behind $\mathsf{Ctrb}^{\mathscr{R}}$: $\mathsf{Ctrb}^{\mathscr{R}'}$ is not entirely aligned with the topological reality of G and hence neither with the intuition of the counterfactuality principle as sketched in the introduction. Yet more nuanced, $\mathsf{Ctrb}^{\mathscr{S}}$ determines a weighted average contribution of b to a as provided by $\mathsf{Ctrb}^{\mathscr{R}}$ in each restriction of G to some subset of its nodes that contains a and b: $\mathsf{Ctrb}_{\mathsf{a}}^{\mathscr{S}}(\mathsf{b}) = -0.3125$. Interestingly, we have $\mathsf{Ctrb}_{\mathsf{a}}^{\mathscr{S}}(\mathsf{b}) + \mathsf{Ctrb}_{\mathsf{a}}^{\mathscr{S}}(\mathsf{c}) = \tau(\mathsf{a}) - \sigma(\mathsf{a})$, i.e. the sum of the contributions of all arguments to a (note that the contribution of a to itself is not defined) exactly accounts for the difference between a's initial and final strength. This is aligned with the intuition of (quantitative) contribution existence and can formally be traced back to the efficiency principle of Shapley values (see Subsection 4.1).

Note that the three contribution functions we have covered so far all use some form of argument removal (and, indeed, always remove at least the contributing argument) to determine the contribution. In contrast, the gradient contribution function measures the effect of marginally changing the contributing argument's initial strength on the final strength of the topic argument. With the strength function expanded as $\sigma(\mathsf{a}) = \tau(\mathsf{a}) - \tau(\mathsf{a})\sigma(\mathsf{b}) = \tau(\mathsf{a})(1 - (\tau(\mathsf{b}) + (1 - \tau(\mathsf{b}))\sigma(\mathsf{c}))) = \tau(\mathsf{a}) - \tau(\mathsf{a})\tau(\mathsf{b}) + \tau(\mathsf{a})\tau(\mathsf{b})\tau(\mathsf{c}) - \tau(\mathsf{a})\tau(\mathsf{c})$, we for instance find $\frac{\partial}{\partial \tau(\mathsf{b})}\sigma(\mathsf{a}) = -\tau(\mathsf{a}) + \tau(\mathsf{a})\tau(\mathsf{c})$, which at the vector $(0.5, 0.5, 0.5)$ of initial strengths evaluates to $\mathsf{Ctrb}_{\mathsf{a}}^{\partial}(\mathsf{b}) = -0.25$. This also means that we can quantify the strength of an argument to itself: $\frac{\partial}{\partial \tau(\mathsf{a})}\sigma(\mathsf{a}) = 1 - \tau(\mathsf{b}) + \tau(\mathsf{b})\tau(\mathsf{c}) - \tau(\mathsf{c})$ evaluates at the initial

Table 4

Argument contributions, given DFQuAD semantics, in G (Fig. 2.1), given different contribution functions (contributors are row headers, topic arguments column headers; ⊥ denotes undefined contributions).

$Ctrb^{\mathscr{R}}$	a	b	c
a	⊥	0	0
b	−0.375	⊥	0
c	−0.125	0.25	⊥

$Ctrb^{\mathscr{R}'}$	a	b	c
a	⊥	0	0
b	−0.25	⊥	0
c	−0.125	0.25	⊥

$Ctrb^{\mathscr{S}}$	a	b	c
a	⊥	0	0
b	−0.3125	⊥	0
c	−0.0625	0.25	⊥

$Ctrb^{\partial}$	a	b	c
a	0.25	0	0
b	−0.25	0.5	0
c	−0.25	0.5	1

strengths to $Ctrb^{\partial}_a(a) = 0.25$. We may claim that, in contrast to the other contribution functions, the gradient contribution function answers a fundamentally different question, which is not about the argument's existence, but about the local effect of changing its initial strength. Table 4 shows all contributions that can be computed given G. This overview already gives us an intuition of some of the principles, such as the quantitative contribution existence principle of $Ctrb^{\mathscr{S}}$; also, note that contribution functions $Ctrb^{\mathscr{R}}, Ctrb^{\mathscr{R}'}, Ctrb^{\mathscr{S}}$ given, respectively, in Equations (1), (2), (3) are partial: they do not define an argument's contribution to itself $Ctrb_a(a)$, since $\sigma_{G'}(a)$ is undefined when a is not in (the set of arguments of) G'. In contrast, the gradient contribution function $Ctrb^{\partial}$ (Equation (5)) is total.[3]

4. Principles

We expect some intuitive behaviour from a contribution function and can formalise these expectations as contribution function principles, similar to principles that have been defined for gradual semantics [23]. In the case of contribution functions, principle satisfaction obviously depends on the properties of the semantics used. A few contribution function principles were proposed in [13]. We recap those next and propose new ones in the following subsections. In the remainder of this section, as well as of the next one, we restrict attention to acyclic QBAGs and fix otherwise arbitrary gradual semantics σ, QBAG G = $(Args, \tau, Att, Supp)$ and topic argument $a \in Args$, denoting the contributor by x (with $x \in Args$). A contribution function Ctrb will be said to satisfy a principle w.r.t. σ if (and only if) a given condition holds for arbitrary (acyclic) G and a, if not stated otherwise.

4.1. Contribution existence

One would expect that whenever an argument's final strength differs from its initial strength, there is another argument whose contribution explains this difference. Accordingly, the contribution existence principle stipulates that some contribution from another argument exists whenever the final strength of the topic argument changes.

Principle 4.1 (*Contribution Existence [13]*[4]). Ctrb satisfies the *contribution existence principle* w.r.t. a gradual semantics σ iff $\sigma(a) \neq \tau(a)$ implies that $\exists x \in Args \setminus \{a\}$ with $Ctrb_a(x) \neq 0$.

Here and henceforth, where the context is clear we may drop "w.r.t. a gradual semantics σ" for any principle. We can further strengthen the principle and stipulate that the sum of all contributions must equal the difference between the topic argument's initial strength and its final strength.

Principle 4.2 (*Quantitative Contribution Existence*). Ctrb satisfies the *quantitative contribution existence principle* w.r.t. a gradual semantics σ iff it holds that $\sum_{x \in Args \setminus \{a\}} Ctrb_a(x) = \sigma(a) - \tau(a)$.

Intuitively, quantitative contribution existence implies contribution existence.

Proposition 4.1. *If* Ctrb *satisfies quantitative contribution existence w.r.t. a gradual semantics* σ *then* Ctrb *satisfies contribution existence w.r.t.* σ.

Proof. Assume Ctrb satisfies quantitative contribution existence w.r.t. σ. Observe that by definition of quantitative contribution existence (Principle 4.2), if $\sigma(a) \neq \tau(a)$ then $\sum_{x \in Args \setminus \{a\}} Ctrb_a(x) \neq 0$ and hence it must hold that $\exists x \in Args \setminus \{a\}$ with $Ctrb_a(x) \neq 0$. This means contribution existence (Principle 4.1) must be satisfied, which proves the proposition. □

---

[3] For the sake of conciseness, we abstain from formalising this trivial observation.

[4] Note that we slightly adjusted the principle: the non-zero contribution needs to be explained by an argument other than the topic argument.

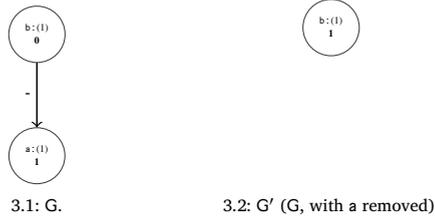3.1: G.   3.2: G′ (G, with a removed).

Fig. 3. While our contribution functions can be intuitively expected to satisfy directionality with respect to modular semantics, this may not be the case with respect to other semantics: consider Ctrb$^{\mathcal{R}}$ and $\sigma(x) = \tau(x) - \sum_{\{y|x \in Att(y)\}} \tau(y)$.

4.2. Directionality

Assuming that the gradual semantics satisfies *directionality* (Principle 2.1 for gradual semantics), contribution functions should respect the direction of influence as represented by the topology of the QBAG. Accordingly, the directionality principle for contribution functions stipulates that an argument's contribution to a topic argument can only be non-zero if there is a directed path from the former to the latter.

Principle 4.3 (*Directionality (Contribution Function) [13]*). Ctrb satisfies the *directionality principle* w.r.t. a gradual semantics σ iff whenever there is no directed path in G from $x \in Args$ to a, then Ctrb$_a$(x) = 0.

Intuitively, we can expect that our contribution functions satisfy directionality with respect to modular semantics that traverse a QBAG from leave nodes (without incoming supports or attacks). However, directionality is not necessarily satisfied if we cannot make any assumptions about semantics behaviour. In particular, if a gradual semantics makes an argument's final strength dependent on the arguments it attacks or supports (somewhat reflecting the idea of *range-based* semantics in abstract argumentation [31]), then it does not respect *directionality*. For such a semantics, all of the contribution functions introduced above would violate the directionality principle for contribution functions. For instance, consider the strength function $\sigma(x) = \tau(x) - \sum_{\{y|x \in Att(y)\}} \tau(y)$, the QBAG G = ({a, b}, {(a, 1), (b, 1)}, {(b, a)}, {}) with b attacking a, and Ctrb$^{\mathcal{R}}$ defined using Equation (1) (removal-based contribution). There is no directed path from a to b, but removing a changes the final strength of b from 0 to 1, whence the contribution of a to b is $0 - 1 \neq 0$ (see Fig. 3).

4.3. Faithfulness

One may want to demand that a contribution function *faithfully* represents the effect of one argument on another one. The effect can be measured in different ways. One natural way in QBAGs is to measure it based on the effect of the initial strength of an argument. If increasing the initial strength of an argument increases (decreases) the final strength of the topic argument, then the contribution score should be positive (negative). We can also imagine that there is no effect at all. To formalise this intuition, we first define an initial strength modification of a QBAG.

Definition 4.1 (*QBAG Initial Strength Modification*). Given a QBAG G = $(Args, \tau, Att, Supp)$ and an argument $x \in Args$, we define the *initial strength modification of x in G* as the QBAG G $\downarrow_{\tau(x) \leftarrow \epsilon} := (Args, \tau', Att, Supp)$, where $\epsilon \in \mathbb{I}$, $\tau'(x) = \epsilon$ and $\tau'(y) = \tau(y)$ for all other $y \in Args \setminus \{x\}$.

One natural first faithfulness property could then be stated as follows: if an argument's contribution is positive/zero/negative, then increasing the argument's base score should increase/not affect/decrease the strength of the topic argument.

Principle 4.4 (*Strong Faithfulness*). Ctrb satisfies the *strong faithfulness principle* w.r.t. a gradual semantics σ iff for every QBAG G = $(Args, \tau, Att, Supp)$, for all a, x $\in Args$, $\epsilon \in \mathbb{I}$ and $G_\epsilon = G \downarrow_{\tau(x) \leftarrow \epsilon}$ the following statements hold:

- If Ctrb$_a$(x) < 0, then $\sigma_G(a) < \sigma_{G_\epsilon}(a)$ whenever $\epsilon < \tau(x)$ and $\sigma_G(a) > \sigma_{G_\epsilon}(a)$ whenever $\epsilon > \tau(x)$.
- If Ctrb$_a$(x) = 0, then $\sigma_G(a) = \sigma_{G_\epsilon}(a)$ for all $\epsilon \in \mathbb{I}$.
- If Ctrb$_a$(x) > 0, then $\sigma_G(a) > \sigma_{G_\epsilon}(a)$ whenever $\epsilon < \tau(x)$ and $\sigma_G(a) < \sigma_{G_\epsilon}(a)$ whenever $\epsilon > \tau(x)$.

Intuitively, a negative (positive) score guarantees that increasing the initial strength of the evaluated argument will decrease (increase) the final strength of the topic argument. However, this property can only be satisfied if we can guarantee a monotonic effect of arguments. This is not necessarily the case as we illustrate in Fig. 1. In the QBAG on the left (Fig. 1.1), argument e has a non-monotonic influence on the topic argument a. Fig. 1.2 plots the final strength of a (y-axis) as a function of the initial strength of e under DFQuAD semantics. e's initial strength has a negative influence up to 0.5. Then the influence becomes positive. The plot illustrates how the initial strength of e influences the final strength of a. As we increase $\tau(e)$ from 0 to 0.5, a becomes weaker. However, at this point, the effect reverses, and increasing $\tau(e)$ further will make a stronger. In particular, if we let $\tau(e) = 0.2$ in the QBAG on the

left in Fig. 1, then the effect is negative for $\epsilon \in [0, 0.8)$, neutral for $\epsilon = 0.8$ and positive for $\epsilon \in (0.8, 1]$. We provide analogous examples for QE semantics (Fig. 6), as well as for EB semantics (Fig. 9); for SD-DFQuAD and EBT semantics, we provide counterexamples for the case $\mathsf{Ctrb}_a(x) = 0$ ($\sigma_{G_\epsilon}(a) \neq \sigma_{G_\epsilon}(a)$ for at least some $\epsilon \in \mathbb{I}$) in Figs. 8 and 10, respectively. Hence, we assume that strong faithfulness cannot be reasonably satisfied. The property may be desirable for explaining "monotonic" QBAGs (see Section 6), but is too strong for general acyclic QBAGs.

If an argument has a non-monotonic effect on the topic argument, we may still be able to capture its effect faithfully in a small environment of the initial strength, where it behaves monotonically. However, even the local effect cannot always be categorised as positive, negative, or neutral. For example, when we let $\tau(\mathsf{e}) = 0.5$ in Fig. 1.1, the local effect switches from positive to negative at this point and cannot really be classified as any of the three. In our definition of local faithfulness, we therefore only demand that positive and negative effects are faithfully captured. 0 may mean that there is no local effect or that the local effect changes at this point. These considerations motivate the following definition.

Principle 4.5 *(Local Faithfulness)*. Ctrb satisfies the *local faithfulness principle* w.r.t. a gradual semantics σ iff there exists a $\delta > 0$ such that for all $\epsilon \in [\tau(x) - \delta, \tau(x) + \delta] \cap \mathbb{I}$ and $G_\epsilon = G \downarrow_{\tau(x) \leftarrow \epsilon}$ the following statements hold:

- If $\mathsf{Ctrb}_a(x) < 0$, then $\sigma_G(a) < \sigma_{G_\epsilon}(a)$ whenever $\epsilon < \tau(x)$ and $\sigma_G(a) > \sigma_{G_\epsilon}(a)$ whenever $\epsilon > \tau(x)$.
- If $\mathsf{Ctrb}_a(x) > 0$, then $\sigma_G(a) > \sigma_{G_\epsilon}(a)$ whenever $\epsilon < \tau(x)$ and $\sigma_G(a) < \sigma_{G_\epsilon}(a)$ whenever $\epsilon > \tau(x)$.

As opposed to strong faithfulness, local faithfulness does not consider all strength modifications, but only those in a small δ-environment of the initial strength of the contributing argument. Furthermore, it only demands that positive (negative) scores guarantee a positive (negative) effect of the argument. As discussed before, the score 0 may mean no effect or a changing effect.

We can strengthen the local faithfulness principle by requiring that the change in the topic argument must be approximately equal to the contribution of an argument times its change, that is, we want that $\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) \approx \sigma_G(a) + \epsilon \cdot \mathsf{Ctrb}_a(x)$ if ϵ is small (locality). To formalize this, we consider the error term $e(\epsilon) = \sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) - (\sigma_G(a) + \epsilon \cdot \mathsf{Ctrb}_a(x))$ and demand that it goes to 0 as ϵ goes to 0. The condition is satisfied trivially for all continuous semantics.[5] Therefore, we add a condition on the convergence speed, namely $\lim_{\epsilon \to 0} \frac{e(\epsilon)}{\epsilon} = 0$. This means that the error term goes asymptotically significantly faster to 0 than ϵ does.

Principle 4.6 *(Quantitative Local Faithfulness)*. Ctrb satisfies the *quantitative local faithfulness principle* w.r.t. a gradual semantics σ iff we have

$$\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) = \sigma_G(a) + \epsilon \cdot \mathsf{Ctrb}_a(x) - e(\epsilon), \tag{6}$$

where $e(\epsilon)$ is an error term with the property $\lim_{\epsilon \to 0} \frac{e(\epsilon)}{\epsilon} = 0$.

Let us note that both strong and quantitative local faithfulness imply local faithfulness.

Proposition 4.2. *Given a contribution function* Ctrb *and a gradual semantics* σ, *if* Ctrb *satisfies strong faithfulness w.r.t.* σ *then* Ctrb *satisfies local faithfulness w.r.t.* σ *and if* Ctrb *satisfies quantitative local faithfulness w.r.t.* σ *then* Ctrb *satisfies local faithfulness w.r.t.* σ.

Proof. Strong faithfulness obviously implies local faithfulness because the first and second condition of local faithfulness are relaxations of the first and third condition of strong faithfulness.

To see that quantitative local faithfulness implies local faithfulness, first note that the local change is

$$\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) - \sigma_G(a) = \epsilon \cdot \mathsf{Ctrb}_a(x) - e(\epsilon).$$

We can assume $\mathsf{Ctrb}_a(x) \neq 0$ because local faithfulness excludes this case. Since $\lim_{\epsilon \to 0} \frac{e(\epsilon)}{\epsilon} = 0$, we can find a $\delta > 0$ such that $\epsilon < \delta$ implies $|\frac{e(\epsilon)}{\epsilon}| < |\mathsf{Ctrb}_a(x)|$. Hence, for all such ϵ, we have

$$\frac{\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) - \sigma_G(a)}{\epsilon} = \mathsf{Ctrb}_a(x) - \frac{e(\epsilon)}{\epsilon}.$$

If $\mathsf{Ctrb}_a(x) < 0$, then the difference must be negative (because $|\frac{e(\epsilon)}{\epsilon}| < |\mathsf{Ctrb}_a(x)|$) and if $\mathsf{Ctrb}_a(x) > 0$, then the difference must be positive (again, because $|\frac{e(\epsilon)}{\epsilon}| < |\mathsf{Ctrb}_a(x)|$). Since $\epsilon > 0$, we can conclude that $\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) - \sigma_G(a) < 0$ if $\mathsf{Ctrb}_a(x) < 0$, that is, $\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) < \sigma_G(a)$ as desired. Symmetrically, $\mathsf{Ctrb}_a(x) > 0$ implies $\sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) > \sigma_G(a)$, which completes the proof. □

Strong and quantitative local faithfulness are incomparable. The former is stronger than the latter because it is global rather than local and the latter is stronger than the former because it makes quantitative rather than just qualitative assumptions (assumptions

---

[5] This is because continuity implies that $\lim_{\epsilon \to 0} \sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) = \sigma_G(a)$ and $\lim_{\epsilon \to 0} \epsilon \cdot \mathsf{Ctrb}_a(x) = 0$.

about the magnitude of contribution values rather than only assumptions about the sign). One could define a strong quantitative notion but since even the qualitative version is too strong for our purposes (as discussed above), we refrain from doing so.

4.4. Counterfactuality

The counterfactuality principle formalises the intuition that an argument should have a positive/negative contribution value only if its removal from G leads to a decrease/increase in the topic argument's strength.

Principle 4.7 (*Counterfactuality*). Ctrb satisfies the *counterfactuality principle* w.r.t. a gradual semantics σ iff for any $x \in Args$ the following statements hold:

- If $\mathrm{Ctrb}_a(x) < 0$, then $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$.
- If $\mathrm{Ctrb}_a(x) = 0$, then $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$.
- If $\mathrm{Ctrb}_a(x) > 0$, then $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$.

We can further strengthen this principle by introducing the *quantitative counterfactuality principle*, requiring that the argument's contribution to the topic argument equals the effect of the former's removal to the latter's final strength.

Principle 4.8 (*Quantitative Counterfactuality*). Ctrb satisfies the *quantitative counterfactuality principle* w.r.t. a gradual semantics σ iff for any $x \in Args$ it holds that $\mathrm{Ctrb}_a(x) = \sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$.

As one may expect intuitively, quantitative counterfactuality implies counterfactuality.

Proposition 4.3. *If* Ctrb *satisfies quantitative counterfactuality w.r.t. a gradual semantics* σ *then* Ctrb *satisfies counterfactuality w.r.t.* σ.

Proof. Considering the definition of the counterfactuality (Principle 4.7), we have three cases:

$\mathrm{Ctrb}_a(x) < 0$. Then, $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$ must hold. This is the case, given that Ctrb satisfies quantitative counterfactuality w.r.t. σ because this principle stipulates that we have $\mathrm{Ctrb}_a(x) = \sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{x\}}}(a) < 0$.

$\mathrm{Ctrb}_a(x) = 0$. Then, $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$ must hold. This is the case, given that Ctrb satisfies quantitative counterfactuality w.r.t. σ because this principle stipulates that we have $\mathrm{Ctrb}_a(x) = \sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{x\}}}(a) = 0$.

$\mathrm{Ctrb}_a(x) > 0$. Then, $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$ must hold. This is the case, given that Ctrb satisfies quantitative counterfactuality w.r.t. σ because this principle stipulates that we have $\mathrm{Ctrb}_a(x) = \sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{x\}}}(a) > 0$. $\square$

5. Principle-based analysis

We now provide a theoretical analysis of the principles introduced in the previous section, considering the semantics listed in Table 3. The results are summarised in Table 1.

To make our counterexamples more robust against numerical inaccuracies, e.g., when approximating gradients or executing floating point operations, we computed every example with two different implementations: once using *QBAF-Py*, an extended version of a C/Python (C with Python bindings) library first introduced in [32], which is available at https://github.com/TimKam/Quantitative-Bipolar-Argumentation and once using *Uncertainpy*, a Python implementation based on ideas from [33], which is available at https://github.com/nicopotyka/Uncertainpy.

5.1. Contribution existence

We first observe that $\mathrm{Ctrb}^{\mathscr{R}}$ satisfies contribution existence with respect to QE and EB semantics.

Proposition 5.1. $\mathrm{Ctrb}^{\mathscr{R}}$ *satisfies the contribution existence principle w.r.t. QE and EB semantics* σ.

Proof. Consider a modular semantics that uses the aggregation function $\alpha = \alpha_v^{\Sigma}$ and any of the influence functions ι_w^e or ι_w^p (Table 2). We observe that QE and EB semantics are such semantics. From the definition of α_v^{Σ} it follows that if $\sigma_G(a) \neq \tau_G(a)$ then there must exist an argument $x \in Args$ s.t. $\sigma_G(x) \neq 0$ and x is a direct attacker or supporter of a; because our QBAG is acyclic, there must exist such x that is *not* also an indirect attacker or supporter of a. Hence, it holds that $\alpha_G(a) \neq \alpha_{G\downarrow_{Args\setminus\{x\}}}(a)$. Consequently, by definition of $\iota \in \{\iota_w^e, \iota_w^p\}$ it must hold that $\iota_G(a) \neq \iota_{G\downarrow_{Args\setminus\{x\}}}(a)$, from which it follows by definition of $\mathrm{Ctrb}^{\mathscr{R}}$ that $\mathrm{Ctrb}_a^{\mathscr{R}}(x) \neq 0$, i.e., the contribution existence principle is satisfied. $\square$

Similarly, $\mathrm{Ctrb}^{\mathscr{R}'}$ satisfies contribution existence with respect to QE and EB semantics.
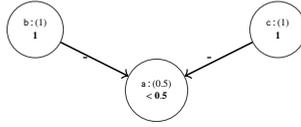
Fig. 4. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ violate the contribution existence and quantitative contribution existence principles w.r.t. DFQuAD, SD-DFQuAD, and EBT semantics.

Proposition 5.2. $\mathrm{Ctrb}^{\mathscr{R}'}$ *satisfies the contribution existence principle w.r.t. QE and EB semantics* σ.

Proof. Consider a modular semantics that uses the aggregation function $\alpha = \alpha_v^{\Sigma}$ and any of the influence functions ι_w^e, or ι_w^p (Table 2). We observe that QE and EB semantics are such semantics. From the definition of α_v^{Σ} it follows that if $\sigma_G(a) \neq \tau_G(a)$, there must exist an argument x s.t. $\tau_G(x) \neq 0$ (for this observation, we also rely on the definition of $\iota \in \{\iota_w^e, \iota_w^p\}$: if $\sigma_G(a) \neq 0$ then $\tau_G(a) \neq 0$ must hold, because $\tau_G(a) = 0$ implies that $\sigma_G(a) = 0$), x is a direct attacker or supporter of a; because our QBAG is acyclic, there must exist such x that is *not* also an indirect attacker or supporter of a. Hence it holds that $\alpha_{(Args,\tau,Att\setminus\{(y,x)|(y,x)\in Att\}, Supp\setminus\{(y,x)|(y,x)\in Supp\})}(a) - \alpha_{G\downarrow_{Args\setminus\{x\}}}(a) \neq 0$. Consequently, for $\iota \in \{\iota_w^e, \iota_w^p\}$ it holds that $\iota_{(Args,\tau,Att\setminus\{(y,x)|(y,x)\in Att\}, Supp\setminus\{(y,x)|(y,x)\in Supp\})}(a) \neq \iota_{G\downarrow_{Args\setminus\{x\}}}(a)$, from which it follows by definition of $\mathrm{Ctrb}^{\mathscr{R}'}$ that $\mathrm{Ctrb}_a^{\mathscr{R}'}(x) \neq 0$, i.e., the contribution existence principle is satisfied. $\quad\square$

In contrast, both $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ can violate contribution existence—and hence also quantitative contribution existence—with respect to DFQuAD, SD-DFQuAD, and EBT semantics.

Proposition 5.3. $\mathrm{Ctrb}^{\mathscr{R}}$ *and* $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate the contribution existence and quantitative contribution existence principles w.r.t. DFQuAD, SD-DFQuAD, and EBT semantics.*

Proof. Consider the QBAG depicted in Fig. 4, which we denote by $G = (Args, \tau, Att, Supp)$. Given DFQuAD, SD-DFQuAD, and EBT semantics σ, we observe that $\sigma_G(a) < 0.5 = \tau(a)$ (0.5 is the initial strength of a). Also, $\sigma_{G\downarrow_{\{a,b\}}}(a) = \sigma_{G\downarrow_{\{a,c\}}}(a) = \sigma_G(a)$. Consequently, by definition of $\mathrm{Ctrb}^{\mathscr{R}}$ it must hold that $\mathrm{Ctrb}_a^{\mathscr{R}}(b) = \mathrm{Ctrb}_a^{\mathscr{R}}(c) = 0$, which proves the violation of contribution existence for $\mathrm{Ctrb}^{\mathscr{R}}$. Similarly, it holds for $G' = (Args, \tau, Att \setminus \{(y,b)|(y,b)\in Att\}, Supp \setminus \{(y,b)|(y,b)\in Supp\})$ and $G'' = (Args, \tau, Att \setminus \{(y,c)|(y,c)\in Att\}, Supp \setminus \{(y,c)|(y,c)\in Supp\})$ that $\sigma_{G'}(a) = \sigma_{G''}(a) = \sigma_G(a)$. Hence, by definition of $\mathrm{Ctrb}^{\mathscr{R}'}$ it must hold that $\mathrm{Ctrb}_a^{\mathscr{R}'}(b) = \mathrm{Ctrb}_a^{\mathscr{R}'}(c) = 0$, which proves the violation of contribution existence for $\mathrm{Ctrb}^{\mathscr{R}'}$. $\quad\square$

The counterexample in Fig. 4 applies to other semantics that use α_v^{max} or α_v^{Π} as the aggregation and ι_w^l, ι_w^e, or ι_w^p as the influence function because a single attacker (supporter) with strength 1 under these semantics has the maximum effect, so that removing another attacker (supporter) cannot affect the topic argument.

The Shapley value-based contribution function satisfies even the stronger quantitative contribution existence principle w.r.t. all of the surveyed argumentation semantics. Intuitively, it does not suffer from the previously observed problem because it does not only remove the argument in the existing situation, but looks at the effect of the argument in all possible situations.

Proposition 5.4. $\mathrm{Ctrb}^{\mathscr{S}}$ *satisfies the quantitative contribution existence principle w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics* σ.

Proof. We provide the proof by characterising $\mathrm{Ctrb}^{\mathscr{S}}$ relative to the contributor's Shapley value of a coalition game where the set of players is $Args \setminus \{a\}$ (where a is our topic argument) and the utility function $v : 2^{Args\setminus\{a\}} \to \mathbb{R}$ is defined as follows (given $S \subseteq Args \setminus \{a\}$):

$$v(S) := \sigma(a) - \sigma_{G\downarrow_{Args\setminus S}}(a)$$

Note that by subtracting $\sigma_{G\downarrow_{Args\setminus S}}(a)$ from $\sigma(a)$, we ensure that $v(\emptyset) = 0$. Then, we obtain the following Shapley value (given player/contributor $x \in Args \setminus \{a\}$):

$$\phi_x(v) =$$
$$\sum_{X \subseteq Args\setminus\{x,a\}} \frac{|X|! \cdot (|Args \setminus \{a\}| - |X| - 1)!}{|Args \setminus \{a\}|!} ((v(X \cup \{x\}) - v(X))) =$$
$$\sum_{X \subseteq Args\setminus\{x,a\}} \frac{|X|! \cdot (|Args \setminus \{a\}| - |X| - 1)!}{|Args \setminus \{a\}|!} \left(\sigma(a) - \sigma_{G\downarrow_{Args\setminus(X\cup\{x\})}}(a) - (\sigma(a) - \sigma_{G\downarrow_{Args\setminus X}}(a)) \right) =$$
$$\sum_{X \subseteq Args\setminus\{x,a\}} \frac{|X|! \cdot (|Args \setminus \{a\}| - |X| - 1)!}{|Args \setminus \{a\}|!} \left(\sigma_{G\downarrow_{Args\setminus X}}(a) - \sigma_{G\downarrow_{Args\setminus(X\cup\{x\})}}(a) \right) =$$
$$\mathrm{Ctrb}_a^{\mathscr{S}}(x).$$

Now, from the efficiency principle of Shapley values, it follows that $\sum_{x \in Args \setminus \{a\}} \phi_x(v) = v(Args \setminus \{a\})$ [20]. Hence, because $v(Args \setminus \{a\}) = \sigma(a) - \tau(a)$ and $\sum_{x \in Args \setminus \{a\}} \phi_x(v) = \sum_{x \in Args \setminus \{a\}} \text{Ctrb}_a^{\mathscr{S}}(x)$, it must hold that $\sum_{x \in Args \setminus \{a\}} \text{Ctrb}_a^{\mathscr{S}}(x) = \sigma(a) - \tau(a)$ and consequently, quantitative contribution existence must be satisfied. $\square$

Corollary 5.5. $\text{Ctrb}^{\mathscr{S}}$ *satisfies the contribution existence principle w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics σ.*

Proof. Because $\text{Ctrb}^{\mathscr{S}}$ satisfies quantitative contribution existence w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics (Proposition 5.4), the proof follows from Proposition 4.1 (if a semantics satisfies quantitative contribution existence then it satisfies contribution existence). $\square$

The gradient-based contribution function Ctrb^{∂} violates contribution existence (and hence quantitative contribution existence) with respect to some of the surveyed argumentation semantics, i.e., with respect to DFQuAD, SD-DFQuAD, and EBT semantics.

Proposition 5.6. Ctrb^{∂} *violates the contribution existence and quantitative contribution existence principles w.r.t. DFQuAD, SD-DFQuAD, and EBT semantics σ.*

Proof. Consider again the QBAG depicted in Fig. 4. Our topic argument is a. Given DFQuAD, SD-DFQuAD, and EBT semantics σ, we observe that $\sigma(a) - \tau(a) \neq 0$ but $\text{Ctrb}_a^{\partial}(b) = \text{Ctrb}_a^{\partial}(c) = 0$ and hence contribution existence and quantitative contribution existence must be violated. $\square$

In contrast, we can show that Ctrb^{∂} satisfies (non-quantitative) contribution existence with respect to QE and EB semantics. The proof is roughly based on the same intuition as the one for $\text{Ctrb}^{\mathscr{R}}$ and $\text{Ctrb}^{\mathscr{R}'}$ with respect to QE and EB semantics: when our topic argument's final strength does not equal its initial strength, then there must exist a direct supporter or attacker of this argument whose marginal change in initial strength has an effect on the final strength of the topic argument.

Proposition 5.7. Ctrb^{∂} *satisfies contribution existence w.r.t. QE and EB semantics.*

Proof. Consider a modular semantics that uses the aggregation function $\alpha = \alpha_v^{\Sigma}$ and any of the influence functions ι_w^e or ι_w^p (Table 2). We observe that QE and EB semantics are such semantics. QE and EB satisfy the stability property [25,21] (Principle 2.2), which states that the final strength of an argument without parents is its base score. Since the final strength of the topic argument a is unequal to its base score by assumption, it must have predecessors. Since G is acyclic, we can assume that there exists a topological ordering of the arguments. Consider an arbitrary topological ordering and let x be the predecessor of a with highest index in the ordering. Then x must be a parent of a and there can be no other paths from a to x (for otherwise, x could not have the highest index among a's predecessors). By definition of the aggregation and influence functions, if $(x, a) \in Att$, then (marginally) increasing $\tau(x)$ will decrease $\sigma_G(a)$ and (marginally) decreasing $\tau(x)$ will increase $\sigma_G(a)$. Hence, the partial derivative w.r.t. x must be strictly negative (non-zero). If $(a, x) \in Supp$, it follows symmetrically that the partial derivative w.r.t. x must be strictly positive (non-zero). $\square$

However, only $\text{Ctrb}^{\mathscr{S}}$ satisfies *quantitative* contribution existence; the other contribution functions cannot guarantee that the contributions of a topic argument's contributor "add up" to difference between the topic argument's final and initial strength. We provide counterexamples for $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, and Ctrb^{∂} and QE and EB semantics in the proof of the following proposition.

Proposition 5.8. $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, *and Ctrb^{∂} violate the quantitative contribution existence principle w.r.t. QE and EB semantics σ.*

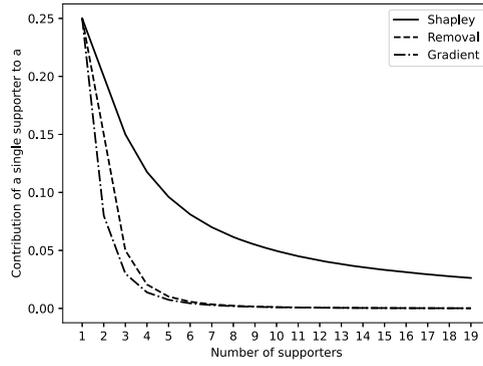**Proof.** Consider yet again the QBAG depicted in Fig. 4, which we denote by $G = (Args, \tau, Att, Supp)$. Our topic argument is a. For QE semantics, we have $\sigma(a) - \tau(a) = -0.4$, but $\text{Ctrb}_a^{\mathscr{R}}(b) + \text{Ctrb}_a^{\mathscr{R}}(c) = \text{Ctrb}_a^{\mathscr{R}'}(b) + \text{Ctrb}_a^{\mathscr{R}'}(c) = -0.3$ and $\text{Ctrb}_a^{\partial}(b) + \text{Ctrb}_a^{\partial}(c) \approx -0.16$. For EB semantics, we have $\sigma(a) - \tau(a) \approx -0.2025$, but $\text{Ctrb}_a^{\mathscr{R}}(b) + \text{Ctrb}_a^{\mathscr{R}}(c) = \text{Ctrb}_a^{\mathscr{R}'}(b) + \text{Ctrb}_a^{\mathscr{R}'}(c) \approx -0.138$ and $\text{Ctrb}_a^{\partial}(b) + \text{Ctrb}_a^{\partial}(c) \approx -0.089$. In these cases, we can easily see that $\sum_{x \in Args \setminus \{a\}} \text{Ctrb}_a(x) = \sigma(a) - \tau(a)$ does not hold and hence, quantitative contribution existence must be violated. $\square$

Finally, let us introduce a simple example that highlights the advantage of the Shapley value-based contribution functions, following the intuition of quantitative contribution existence.

Example 5.1. Consider a set of simple QBAGs, where each QBAG consists of a topic argument a and a number of supporters $b_1, \ldots, b_n$, where $n = 1$ for the smallest QBAG and $n = 20$ for the largest (Fig. 5.1). The arguments are not related otherwise. The initial strength of a is 0.5, whereas each argument of $b_1, \ldots, b_n$ has an initial strength of 1. We apply QE semantics, which satisfies contribution existence with respect to $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, and Ctrb^{∂}. Obviously, in a given QBAG, the contribution of every contributor b_i, $1 \leq i \leq n$ is the same, given a specific contribution function. We plot this contribution for $\text{Ctrb}^{\mathscr{R}}$ (which equals the one of $\text{Ctrb}^{\mathscr{R}'}$ in this case), $\text{Ctrb}^{\mathscr{S}}$, and Ctrb^{∂}; i.e., we plot $\text{Ctrb}_{b_i}^{\mathscr{R}}(a)$, $\text{Ctrb}_{b_i}^{\mathscr{S}}(a)$, and $\text{Ctrb}_{b_i}^{\partial}(a)$ for our set of QBAGs. What we can see is that the contributions
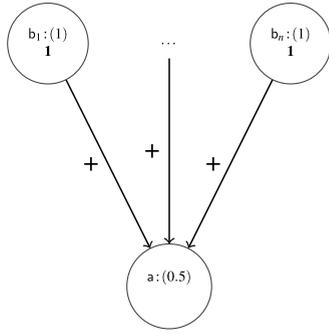
5.1: Argument a with n supporters.   5.2: Contribution of b_i, $1 \leq i \leq n$ to a, given n.

Fig. 5. The contributions of $\text{Ctrb}^{\mathcal{R}}$ and Ctrb^{∂} converge to zero faster than the contributions of $\text{Ctrb}^{\mathcal{S}}$. Here, the contribution of b_i, $1 \leq i \leq n$ to a depends on the number of a's supporters and we apply QE semantics.

of $\text{Ctrb}^{\mathcal{R}}$ and Ctrb^{∂} converge to zero faster with increasing n than the contributions of $\text{Ctrb}^{\mathcal{S}}$. For $n = 10$, the Shapley value-based contribution is still at around 0.05, whereas the removal- and gradient-based contributions are, visually, already indistinguishable from 0. The example highlights that contribution existence alone may not always be a strong enough principle: for larger n we see in the plot (approx. for $6 \leq n \leq 20$), $\text{Ctrb}_{b_i}^{\mathcal{R}}(a)$ and $\text{Ctrb}_{b_i}^{\partial}(a)$ are negligibly small, although the proportional effect of a single supporter is still substantial, which can be considered misleading.

5.2. Directionality

Since modular semantics generally satisfy the *directionality* principle for gradual semantics [21], we should expect that a faithful contribution function satisfies the corresponding principle for contribution functions. As we show next, this is indeed the case for all contribution functions considered here.

Proposition 5.9. $\text{Ctrb}^{\mathcal{R}}$, $\text{Ctrb}^{\mathcal{R}'}$, $\text{Ctrb}^{\mathcal{S}}$, and Ctrb^{∂} satisfy directionality w.r.t. all modular argumentation semantics σ.

Proof. Let $G = (Args, \tau, Att, Supp)$ be a QBAG.

The proof for $\text{Ctrb}^{\mathcal{R}}$. If there is no directed path from x to a, then $\sigma_G(a) = \sigma_{G \downarrow_{Args \setminus \{x\}}}(a)$ by definition of a modular semantics. Consequently, it holds that $\text{Ctrb}_a^{\mathcal{R}}(x) = 0$, as required.

The proof for $\text{Ctrb}^{\mathcal{R}'}$. If there is no directed path from x to a, then for $G' = (Args, \tau, Att \setminus \{(y,x)|(y,x) \in Att\}, Supp \setminus \{(y,x)|(y,x) \in Supp\})$ it holds that $\sigma_{G'}(a) = \sigma_{G \downarrow_{Args \setminus \{x\}}}(a)$ by definition of a modular semantics. Consequently, $\text{Ctrb}_a^{\mathcal{R}'}(x) = 0$, as required.

The proof for $\text{Ctrb}^{\mathcal{S}}$. If there is no directed path from x to a, then for $X \subseteq Args$ it holds that $\sigma_{G \downarrow_{Args \setminus (X \cup \{x\})}}(a) = \sigma_{G \downarrow_{Args \setminus X}}(a)$ by definition of a modular semantics. Consequently, $\text{Ctrb}_a^{\mathcal{S}}(x) = 0$, as required.

The proof for Ctrb^{∂}. If there is no directed path from x to a, then $\sigma(a) = f_a(\ldots, \tau(a), \tau(x), \ldots)$ is independent of $\tau(x)$, by definition of a modular semantics. In that case, trivially, the partial derivative of f_a w.r.t. $\tau(x)$ is null, so that $\text{Ctrb}_a^{\partial}(x) = 0$, as required. $\square$

5.3. (Quantitative) local faithfulness

Intuitively, (quantitative) local faithfulness is best aligned with the gradient-based contribution function, which indeed satisfies the principle with respect to every differentiable modular argumentation semantics. Of course, we require differentiability of the semantics for the gradient to be defined. For acyclic graphs, the condition is satisfied for all commonly considered semantics. In general, a modular semantics is differentiable whenever both the aggregation and influence functions are differentiable. Note that all aggregation functions and the linear and Euler-based influence functions in Table 2 are differentiable. The p-Max aggregation function is differentiable for $p = 2$ [25, Proposition 1] but can be pointwise non-differentiable for some other choices of p. As we explain next, Ctrb^{∂} satisfies quantitative local faithfulness for every differentiable modular argumentation semantics and, thus, also the weaker principle of local faithfulness.
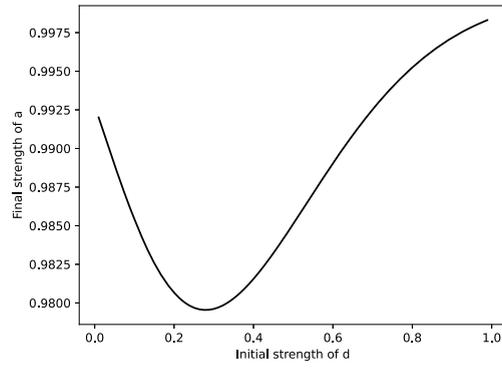
Proposition 5.10. Ctrb^{∂} satisfies quantitative local faithfulness w.r.t. every differentiable modular argumentation semantics.

Proof. Note that (6) is trivially satisfied when letting

$$e(\epsilon) = \sigma_{G \downarrow_{\tau(x) \leftarrow \epsilon}}(a) - (\sigma_G(a) + \epsilon \cdot \text{Ctrb}_a^{\partial}(x)). \qquad (7)$$
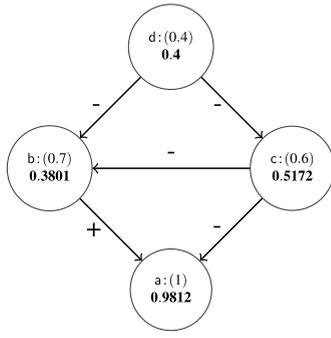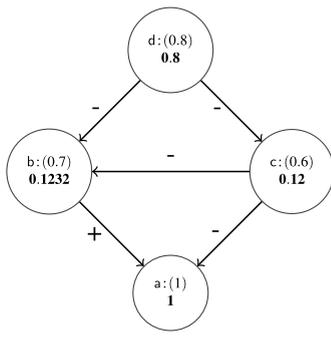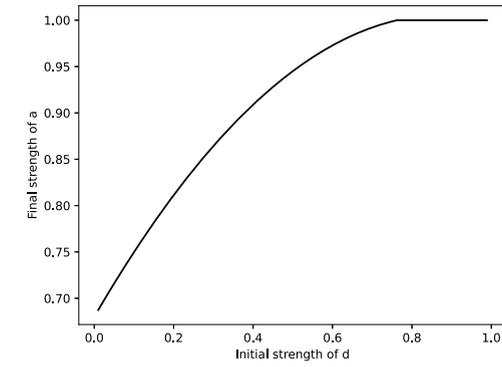
6.1: G.

6.2: Final strength of a, given initial strength of d.

Fig. 6. Ctrb$^{\mathscr{R}}$ and Ctrb$^{\mathscr{R}'}$ violate local faithfulness w.r.t. QE semantics.

7.1: G.

7.2: Final strength of a, given initial strength of d.

Fig. 7. Ctrb$^{\mathscr{R}}$ and Ctrb$^{\mathscr{R}'}$ violate local faithfulness w.r.t. DFQuAD semantics.

It remains to show that $\lim_{\epsilon \to 0} \frac{e(\epsilon)}{\epsilon} = 0$. We have

$$
\begin{aligned}
\lim_{\epsilon \to 0} \frac{e(\epsilon)}{\epsilon} &= \lim_{\epsilon \to 0} \frac{\sigma_{G\downarrow_{\tau(x)\leftarrow\epsilon}}(a) - (\sigma_G(a) + \epsilon \cdot \mathsf{Ctrb}_a^{\partial}(x))}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{\sigma_{G\downarrow_{\tau(x)\leftarrow\epsilon}}(a) - \sigma_G(a)}{\epsilon} - \mathsf{Ctrb}_a^{\partial}(x) \\
&= \mathsf{Ctrb}_a^{\partial}(x) - \mathsf{Ctrb}_a^{\partial}(x) = 0.
\end{aligned}
$$

For the third equality, note that the first term in the second equality is just the partial derivative of the strength of a with respect to the initial strength of x, which is $\mathsf{Ctrb}_a^{\partial}(x)$ by definition. □

Proposition 5.11. *Ctrb$^{\partial}$ satisfies local faithfulness w.r.t. every differentiable modular argumentation semantics.*

Proof. The proof follows immediately from Proposition 4.2 (quantitative local faithfulness implies local faithfulness). □

As we show next by a sequence of counterexamples, the other introduced contribution functions do not satisfy local faithfulness and, thus, by Proposition 4.2, they also do not satisfy quantitative local faithfulness.[6] For each counter-example, we provide a plot of the topic argument's final strength dependent on the contributor's initial strength in order to facilitate a more intuitive grasp of the underlying issues (saddle points or plateaus). Since Ctrb$^{\partial}$ satisfies local faithfulness, we can demonstrate violation of the principle by showing that the sign under a given contribution function differs from the sign under Ctrb$^{\partial}$.

---

[6] According to Proposition 4.2, quantitative local faithfulness implies local faithfulness. Hence, by contraposition, not local faithfulness implies not quantitative local faithfulness.
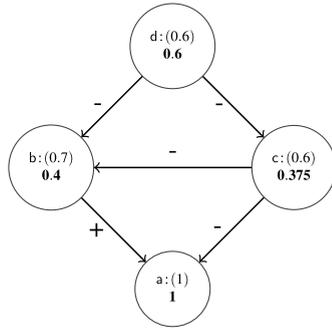
8.1: G.
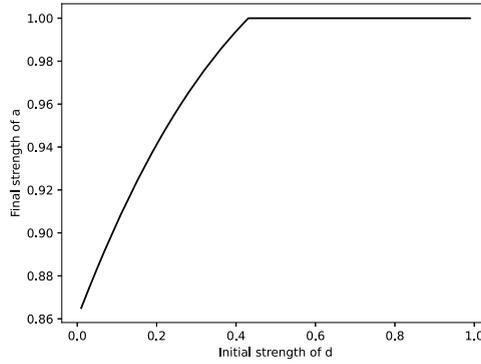
8.2: Final strength of a, given initial strength of d.

Fig. 8. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ violate local faithfulness w.r.t. SD-DFQuAD semantics.

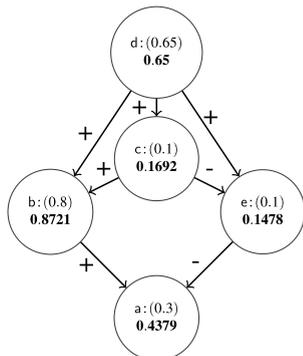**Proposition 5.12.** $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate local faithfulness w.r.t. QE semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 6.1, with the topic argument a and the contributor d. Given QE semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) \approx -0.01122$ and $\mathrm{Ctrb}_a^{\partial}(d) \approx 0.02987$; local faithfulness must be violated because $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) < 0$ but $\mathrm{Ctrb}_a^{\partial}(d) > 0$. ☐

Proposition 5.13. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate local faithfulness w.r.t. DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 7.1, with the topic argument a and the contributor d. Given DFQuAD semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) \approx 0.32 > 0$, but as can be seen in Fig. 7.2, increasing d's initial strength does not increase the topic argument's strength. ☐

Proposition 5.14. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate local faithfulness w.r.t. SD-DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 8.1, with the topic argument a and the contributor d. Given SD-DFQuAD semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) \approx 0.1398 > 0$. However, as can be seen in Fig. 8.2, marginally increasing d's initial strength (0.6) does not increase the topic argument's strength. ☐

Proposition 5.15. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate local faithfulness w.r.t. EB semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 9.1, with the topic argument a and the contributor d. Given EB semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) \approx 0.0016$ and $\mathrm{Ctrb}_a^{\partial}(d) \approx -0.002$; local faithfulness must be violated because $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) > 0$ but $\mathrm{Ctrb}_a^{\partial}(d) < 0$. ☐

Proposition 5.16. $\mathrm{Ctrb}^{\mathscr{R}}$ and $\mathrm{Ctrb}^{\mathscr{R}'}$ *violate local faithfulness w.r.t. EBT semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 10.1, with the topic argument a and the contributor d. Given EBT semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{R}}(d) = \mathrm{Ctrb}_a^{\mathscr{R}'}(d) \approx 0.013 > 0$, but as can be seen in Fig. 10.2, increasing d's initial strength does not increase the topic argument's strength. ☐

Proposition 5.17. $\mathrm{Ctrb}^{\mathscr{S}}$ *violates local faithfulness w.r.t. QE semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 11.1, with the topic argument a and the contributor d. Given QE semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{S}}(d) = \approx -0.0016$ and $\mathrm{Ctrb}_a^{\partial}(d) \approx 0.0302$; local faithfulness must be violated because $\mathrm{Ctrb}_a^{\mathscr{S}}(d) < 0$ but $\mathrm{Ctrb}_a^{\partial}(d) > 0$. ☐

Proposition 5.18. $\mathrm{Ctrb}^{\mathscr{S}}$ *violates local faithfulness w.r.t. DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 1.1, with the topic argument a and the contributor e. Given DFQuAD semantics σ, we have $\mathrm{Ctrb}_a^{\mathscr{S}}(e) \approx -0.0833 < 0$, but as can be seen in Fig. 1.2, increasing e's initial strength does actually increase the topic argument's strength. ☐
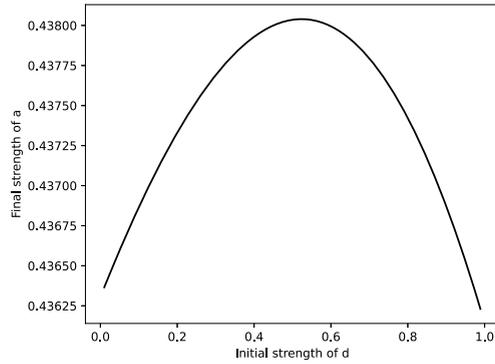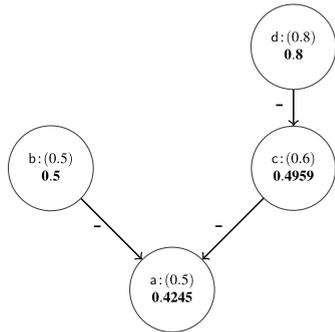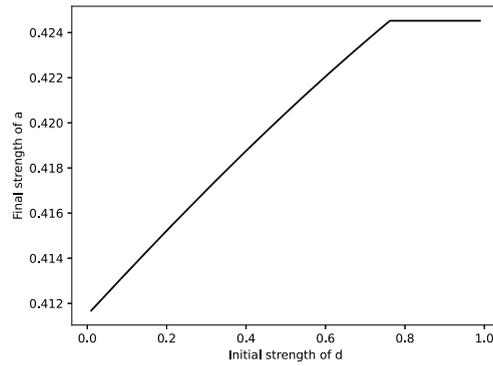
9.1: G.   9.2: Final strength of a, given initial strength of d.

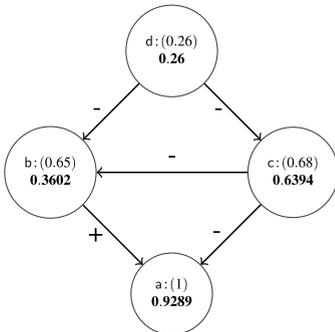**Fig. 9.** Ctrb$^{\mathcal{R}}$ and Ctrb$^{\mathcal{R}'}$ violate local faithfulness w.r.t. EB semantics.
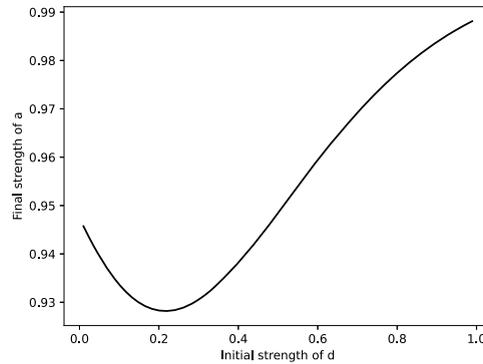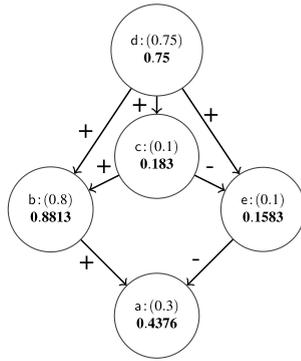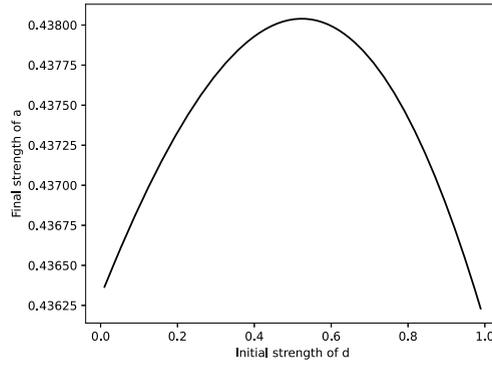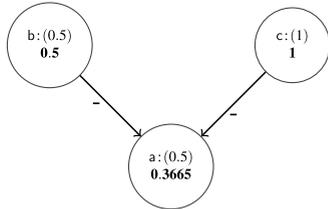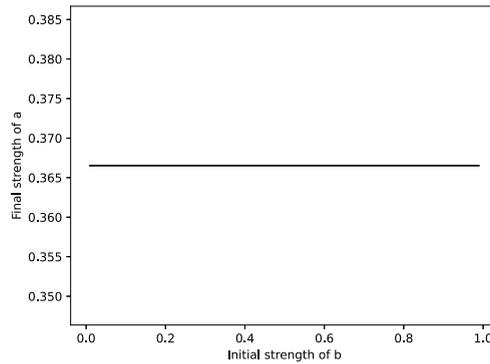

10.1: G.   10.2: Final strength of a, given initial strength of d.

Fig. 10. Ctrb$^{\mathcal{R}}$ and Ctrb$^{\mathcal{R}'}$ violate local faithfulness w.r.t. EBT semantics.

11.1: G.   11.2: Final strength of a, given initial strength of d.

Fig. 11. Ctrb$^{\mathcal{S}}$ violates local faithfulness w.r.t. QE semantics.

Proposition 5.19. Ctrb$^{\mathcal{S}}$ *violates local faithfulness w.r.t. SD-DFQuAD semantics.*

Proof. Consider the QBAG G = $(Args, \tau, Att, Supp)$ depicted in Fig. 8.1, with the topic argument a and the contributor d. Given SD-DFQuAD semantics σ, we have Ctrb$_a^{\mathcal{S}}$(d) $\approx 0.0636 > 0$, but as can be seen in Fig. 8.2, increasing d's initial strength does not increase the topic argument's strength. ◻

Proposition 5.20. Ctrb$^{\mathcal{S}}$ *violates local faithfulness w.r.t. EB semantics.*

12.1: G.

12.2: Final strength of a, given initial strength of d.

Fig. 12. Ctrb$^{\mathscr{S}}$ violates local faithfulness w.r.t. EB semantics.

13.1: G.

13.2: Final strength of a, given initial strength of b.

Fig. 13. Ctrb$^{\mathscr{S}}$ violates local faithfulness w.r.t. EBT semantics.

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 12.1, with the topic argument a and the contributor d. Given EB semantics σ, we have $\text{Ctrb}_{a}^{\mathscr{S}}(d) \approx 0.0007$ and $\text{Ctrb}_{a}^{\partial}(d) \approx -0.0037$; local faithfulness must be violated because $\text{Ctrb}_{a}^{\mathscr{S}}(d) > 0$ but $\text{Ctrb}_{a}^{\partial}(d) < 0$. □

Proposition 5.21. Ctrb$^{\mathscr{S}}$ *violates local faithfulness w.r.t. EBT semantics.*

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 13.1, with the topic argument a and the contributor b. Given EBT semantics σ, we have $\text{Ctrb}_{a}^{\mathscr{S}}(b) \approx -0.0377 < 0$, but changing b's initial strength actually does not affect the strength of the topic argument. □

5.4. (Quantitative) counterfactuality

Let us first show that the contribution function Ctrb$^{\mathscr{R}}$ satisfies quantitative counterfactuality w.r.t. all argumentation semantics.

Proposition 5.22. Ctrb$^{\mathscr{R}}$ *satisfies quantitative counterfactuality w.r.t. all argumentation semantics σ.*

Proof. By definition of Ctrb$^{\mathscr{R}}$, it holds for every G $= (Args, \tau, Att, Supp)$, for every arguments a, x $\in Args$ that $\text{Ctrb}_{a}^{\mathscr{R}}(x) = \sigma_{G}(a) - \sigma_{G\downarrow_{Args\setminus\{x\}}}(a)$. Hence, quantitative counterfactuality (Definition 4.8) is trivially satisfied. □

Corollary 5.23. Ctrb$^{\mathscr{R}}$ *satisfies counterfactuality w.r.t. all argumentation semantics σ.*

Proof. Counterfactuality is satisfied because quantitative counterfactuality is satisfied (Proposition 5.22) and the latter implies the former (Proposition 4.3). □

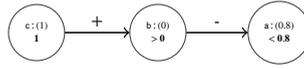**Fig. 14.** Ctrb$^{\mathscr{R}'}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. QE, SD-DFQuAD, and DFQuAD semantics.
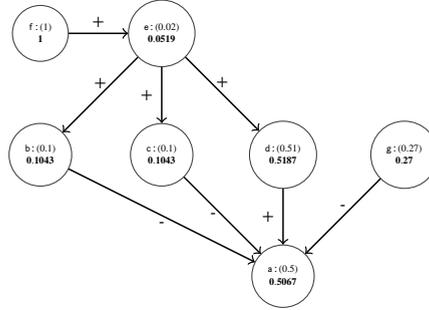

Fig. 15. Ctrb$^{\mathscr{R}'}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. EB semantics.
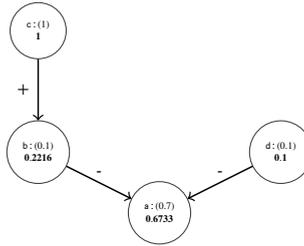

Fig. 16. Ctrb$^{\mathscr{R}'}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. EBT semantics.

All other contribution functions can violate counterfactuality and quantitative counterfactuality w.r.t. all of the surveyed semantics.

Proposition 5.24. Ctrb$^{\mathscr{R}'}$ *violates counterfactuality and quantitative counterfactuality w.r.t. QE, DFQuAD, SD-DFQuAD and semantics.*

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 14, with the topic argument a and the contributor b. Given QE, DFQuAD, or SD-DFQuAD semantics σ, it trivially holds that $\sigma_G(b) > 0$ and hence, $\sigma_G(a) < \tau_G(a) = 0.8$. Intuitively reflecting the influence of b on a, $\sigma_{G\downarrow_{Args\setminus\{b\}}}(a) = 0.8$, i.e., we have $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{b\}}}(a)$. However, we have Ctrb$_a^{\mathscr{R}'}(b) = 0$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{b\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.25. Ctrb$^{\mathscr{R}'}$ *violates counterfactuality and quantitative counterfactuality w.r.t. EB semantics.*

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 15, with the topic argument a and the contributor e. Given EB semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{e\}}}(a) \approx -2.5 \times 10^{-6}$). However, we have Ctrb$_a^{\mathscr{R}'}(e) \approx 3.5431 \times 10^{-6}$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.26. Ctrb$^{\mathscr{R}'}$ *violates counterfactuality and quantitative counterfactuality w.r.t. EBT semantics.*

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 16, with the topic argument a and the contributor b. Given EBT semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{b\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{b\}}}(a) \approx -0.0145$). However, we have Ctrb$_a^{\mathscr{R}'}(b) = 0$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{b\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.27. Ctrb$^{\mathscr{S}}$ *violates counterfactuality and quantitative counterfactuality w.r.t. QE semantics.*

Proof. Consider the QBAG G $= (Args, \tau, Att, Supp)$ depicted in Fig. 17, with the topic argument a and the contributor e. Given QE semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{e\}}}(a) \approx -0.0149$). However, we have Ctrb$_a^{\mathscr{S}}(e) \approx 4.9326 \times$
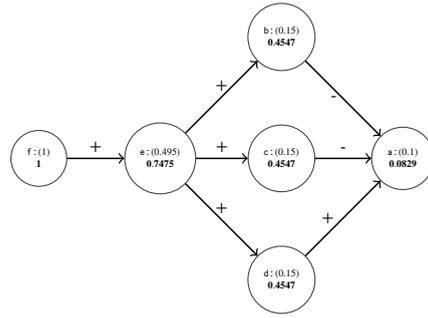
Fig. 17. Ctrb$^{\mathscr{S}}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. QE semantics.
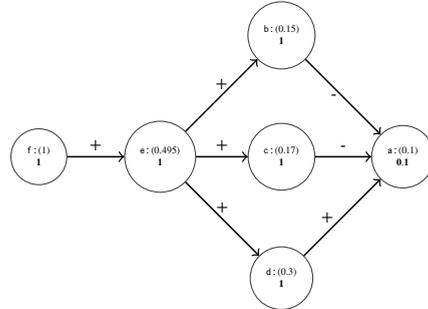

Fig. 18. Ctrb$^{\mathscr{S}}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. DFQuAD semantics.
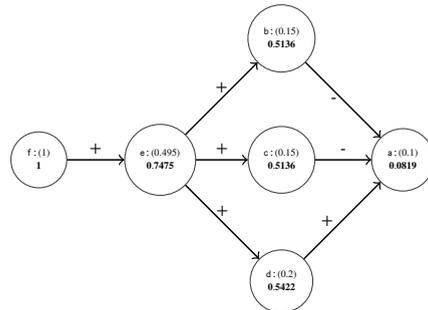

Fig. 19. Ctrb$^{\mathscr{S}}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. SD-DFQuAD semantics.

10^{-5} and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.28. Ctrb$^{\mathscr{S}}$ *violates counterfactuality and quantitative counterfactuality w.r.t. DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 18, with the topic argument a and the contributor e. Given DFQuAD semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{e\}}}(a) \approx -0.0109$). However, we have Ctrb$^{\mathscr{S}}_a$(e) ≈ 0.0021 and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.29. Ctrb$^{\mathscr{S}}$ *violates counterfactuality and quantitative counterfactuality w.r.t. SD-DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 19, with the topic argument a and the contributor e. Given SD-DFQuAD semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{e\}}}(a) \approx -0.0049$). However, we have Ctrb$^{\mathscr{S}}_a$(e) ≈ 0.0027 and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.30. Ctrb$^{\mathscr{S}}$ *violates counterfactuality and quantitative counterfactuality w.r.t. EB semantics.*
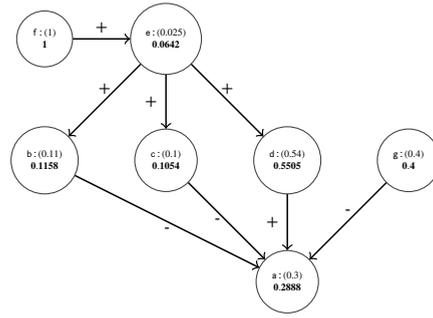
Fig. 20. $\mathrm{Ctrb}^{\mathscr{S}}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. EB semantics.
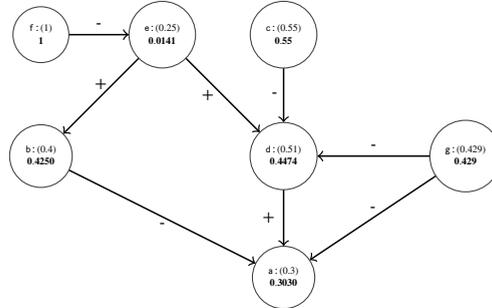

Fig. 21. $\mathrm{Ctrb}^{\mathscr{S}}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. EBT semantics.

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 20, with the topic argument a and the contributor f. Given EB semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{f\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{f\}}}(a) \approx -7.8369 \times 10^{-5}$). However, we have $\mathrm{Ctrb}_a^{\mathscr{S}}(f) \approx 3.438 \times 10^{-6}$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{f\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.31. $\mathrm{Ctrb}^{\mathscr{S}}$ *violates counterfactuality and quantitative counterfactuality w.r.t. EBT semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 21, with the topic argument a and the contributor f. Given EBT semantics σ, it holds that $\sigma_G(a) > \sigma_{G\downarrow_{Args\setminus\{f\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{f\}}}(a) = 7.3331 \times 10^{-5}$). However, we have $\mathrm{Ctrb}_a^{\mathscr{S}}(f) \approx -2.7043 \times 10^{-5}$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{f\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.32. Ctrb^{∂} *violates counterfactuality and quantitative counterfactuality w.r.t. QE semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 22, with the topic argument a and the contributor d. Given QE semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{d\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{d\}}}(a) \approx -0.0038$). However, we have $\mathrm{Ctrb}_a^{\partial}(f) = 0$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{d\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.33. Ctrb^{∂} *violates counterfactuality and quantitative counterfactuality w.r.t. DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 1, with the topic argument a and the contributor e. Given DFQuAD semantics σ, it holds that $\sigma_G(a) < \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$ (observe that $\sigma_G(a) - \sigma_{G\downarrow_{Args\setminus\{e\}}}(a) = -0.125$). However, we have $\mathrm{Ctrb}_a^{\partial}(e) = 0$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) = \sigma_{G\downarrow_{Args\setminus\{e\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.34. Ctrb^{∂} *violates counterfactuality and quantitative counterfactuality w.r.t. SD-DFQuAD semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 23, with the topic argument a and the contributor b. Given SD-DFQuAD semantics σ, it trivially holds that $\sigma_G(b) = 0$; and hence, $\sigma_G(a) = \tau_G(a) = 0.5$. Intuitively reflecting the influence of b on a,
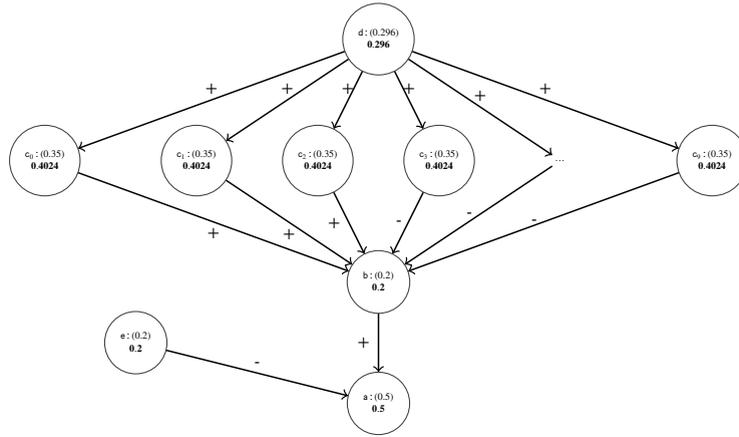
Fig. 22. $Ctrb^{\partial}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. QE semantics. Note that "…" (with incoming support and outgoing attack) represents arguments c_4 - c_8 with same initial strength and incoming support and outgoing attack as c_3 and c_9.
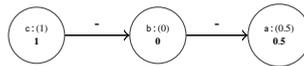

Fig. 23. $Ctrb^{\partial}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. SD-DFQuAD semantics.
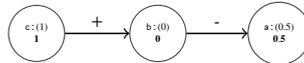
Fig. 24. $Ctrb^{\partial}$ violates the counterfactuality and quantitative counterfactuality principles w.r.t. EB and EBT semantics.

$\sigma_{G\downarrow_{Args\backslash\{b\}}}(a) = 0.5$, i.e., we have $\sigma_G(a) = \sigma_{G\downarrow_{Args\backslash\{b\}}}(a)$. However, we have $Ctrb_a^{\partial}(b) = -0.25$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) < \sigma_{G\downarrow_{Args\backslash\{b\}}}(a)$, which proves the violation of the principles. $\square$

Proposition 5.35. $Ctrb^{\partial}$ *violates counterfactuality and quantitative counterfactuality w.r.t. EB and EBT semantics.*

Proof. Consider the QBAG $G = (Args, \tau, Att, Supp)$ depicted in Fig. 24, with the topic argument a and the contributor b. Given EB or EBT semantics σ, it trivially holds that $\sigma_G(b) = 0$; and hence, $\sigma_G(a) = \tau_G(a) = 0.5$. Intuitively reflecting the influence of b on a, $\sigma_{G\downarrow_{Args\backslash\{b\}}}(a) = 0.5$, i.e., we have $\sigma_G(a) = \sigma_{G\downarrow_{Args\backslash\{b\}}}(a)$. However, we have $Ctrb_a^{\mathscr{R}'}(b) \approx -0.4530$ and hence, according to counterfactuality and quantitative counterfactuality, we must have $\sigma_G(a) < \sigma_{G\downarrow_{Args\backslash\{b\}}}(a)$, which proves the violation of the principles. $\square$

6. Minor results and conjectures

In this section, we describe minor results obtained as a by-product of our search for and analysis of contribution function principles. We also speculate about additional results that future research may obtain.

6.1. Proximity

Intuitively, one may expect that arguments that are *strictly closer* to a topic argument, considering the direction of the attack/support relations, contribute more. In order to turn this intuition into a principle, let us first define what we mean by *strictly closer*.

Definition 6.1 (*Strictly Closer*). Given $G = (Args, \tau, Att, Supp)$ and $a, x, y \in Args$, we say that y is strictly closer to a than x is iff y is on every directed path (in G) from $x \in Args$ to a.

Consider the QBAG in Fig. 17 and the topic argument a. Here, e is strictly closer to a than f is: we can reach a from f only through e. In contrast, b is not strictly closer to a than e is, because we can reach a from e through c (and b is not on the corresponding path). Now, we can formalise the intuition above by defining the *proximity principle*.

Principle 6.1 (*Proximity*). Ctrb satisfies the *proximity principle* w.r.t. a gradual semantics σ iff whenever $y \in Args$ is strictly closer to a than x is then $|Ctrb_a(y)| \geq |Ctrb_a(x)|$.
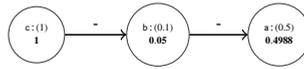
Fig. 25. $\text{Ctrb}^{\mathscr{R}}$ violates the proximity principle w.r.t. QE semantics.

Interestingly, proximity is violated for very simple cases. For example, consider $\text{Ctrb}^{\mathscr{R}}$, QE semantics, and the QBAG in Fig. 25, with topic argument a and contributors b and c. Although b is obviously strictly closer to a than c is, we have $|\text{Ctrb}_a^{\mathscr{R}}(b)| \approx 0.0012 <$ $|\text{Ctrb}_a^{\mathscr{R}}(c)| \approx 0.0037$. Intuitively, b is so substantially weakened by c that the removal of b only marginally affects the final strength of a, whereas the removal of c substantially strengthens b and thus substantially weakens a. Indeed, we could not find any contribution function that satisfies proximity with respect to any semantics (leaving the case for Ctrb^∂ and EBT open). Hence, we opted to exclude proximity from our main list of principles.

Proposition 6.1. $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, and $\text{Ctrb}^{\mathscr{S}}$ *violate proximity w.r.t. QE, DFQuAD, SD-DFQuAD, EB and EBT semantics;* Ctrb^∂ *violates proximity w.r.t. QE, DFQuAD, SD-DFQuAD, and EB semantics.*

The proofs are provided in the appendix. We leave questions whether Ctrb^∂ satisfies or violates proximity with respect to EBT semantics for future work.

We speculate that proximity may be satisfied by the Shapley values-based contribution function $\text{Ctrb}^{\mathscr{S}}$ with respect to some semantics in special cases. By the assumptions of proximity, an argument c that is closer to the topic argument a is on every path from b to c. This means that if we have a "coalition" of arguments that does not contain c, then adding b will not result in a marginal contribution. For the special case of an argumentation framework in which a, b, and c are connected only via the support relation of the QBAG one would expect that if the coalition of arguments does contain c and adding b results in a marginal contribution, then adding c to the coalition where we replace c with b will result in an absolute marginal contribution that is at least as large. As a prerequisite for our conjecture, let us define the notion of a *pure support path*.

Definition 6.2 (*Pure Support Paths*). Given a QBAG $(Args, \tau, Att, Supp)$ and $a, b \in Args$, a path a to b is a pure support path iff it is a path in the directed graph $(Args, Supp)$.

Note that by definition of a QBAG support and attack relations are disjoint, so no edge occurring in a pure support path a to b is in Att.

Let us now provide the conjecture, assuming that we can define principle satisfaction with respect to arguments with specific graph-topological properties.

Conjecture 1. $\text{Ctrb}^{\mathscr{S}}$ *satisfies proximity w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics, for contributors* $x \in Args$ *and topic arguments* $a \in Args$ *for which it holds that all paths from* x *to* a *are pure support paths.*

6.2. Violation of strong faithfulness

In Subsection 4.3, we argued that *strong faithfulness* is too strong as a property because it is incompatible with the fact that the effect of arguments can be non-monotonic (they may be positive in one region and negative in another) or locally but not globally neutral. Here, we elaborate on this idea and give concrete counterexamples for various semantics.

Proposition 6.2. $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, $\text{Ctrb}^{\mathscr{S}}$, *and* Ctrb^∂ *violate strong faithfulness w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics.*

Proof. Observe that if strong faithfulness is satisfied then one of the following statements must be true for all QBAGs $(Args, \tau, Att, Supp)$, all $a, x \in Args$ and all $\epsilon, \epsilon' \in \mathbb{I}$ s.t. $\epsilon > \tau(x) > \epsilon'$, and $G_\epsilon = G \downarrow_{\tau(x) \leftarrow \epsilon}$ and $G'_{\epsilon'} = G \downarrow_{\tau(x) \leftarrow \epsilon'}$:

1. $\sigma_{G_\epsilon}(a) < \sigma_G(a) < \sigma_{G'_{\epsilon'}}(a)$;
2. $\sigma_{G_\epsilon}(a) > \sigma_G(a) > \sigma_{G'_{\epsilon'}}(a)$;
3. $\sigma_{G_\epsilon}(a) = \sigma_G(a) = \sigma_{G'_{\epsilon'}}(a)$.

This is not the case for any of QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics:

QE semantics. Consider the QBAG in Fig. 6.1, contributor d and topic argument a, and the plot in Fig. 6.2 for values of ϵ and ϵ' that provide a counter-example.

DFQuAD semantics. Consider the QBAG in Fig. 1.1, contributor e and topic argument a, and the plot in Fig. 1.2 for values of ϵ and ϵ' that provide a counter-example.

SD-DFQuAD semantics. Consider the QBAG in Fig. 8.1, contributor d and topic argument a, and the plot in Fig. 8.2 for values of ϵ and ϵ' that provide a counter-example.

EB semantics. Consider the QBAG in Fig. 9.1, contributor d and topic argument a, and the plot in Fig. 9.2 for values of ϵ and ϵ' that provide a counter-example.

EBT semantics. Consider the QBAG in Fig. 10.1, contributor d and topic argument a, and the plot in Fig. 10.2 for values of ϵ and ϵ' that provide a counter-example.

Let us provide a step-by-step walk-through for the case of DFQuAD semantics (and the QBAG/plot in Fig. 1); the other cases follow analogously from the examples referenced above. In the QBAG on the left (Fig. 1.1), argument e has a non-monotonic influence on the topic argument a. Fig. 1.2 plots the final strength of a (y-axis) as a function of the initial strength of e under DFQuAD semantics. e's initial strength has a negative influence up to 0.5. Then the influence becomes positive. The plot illustrates how the initial strength of e influences the final strength of a. As we increase $\tau(\mathsf{e})$ from 0 to 0.5, a becomes weaker. However, at this point, the effect reverses, and increasing $\tau(\mathsf{e})$ further will make e stronger. In particular, if we let $\tau(\mathsf{e}) = 0.2$ in the QBAG on the left in Fig. 1.1, then the effect is negative for $\epsilon \in [0, 0.8)$, neutral for $\epsilon = 0.8$ and positive for $\epsilon \in (0.8, 1]$. We provide analogous examples for QE semantics (Fig. 6), as well as for EB semantics (Fig. 9); for SD-DFQuAD and EBT semantics, we provide counterexamples for the case $\mathsf{Ctrb}_\mathsf{a}(x) = 0$ ($\sigma_\mathsf{G}(\mathsf{a}) \neq \sigma_{\mathsf{G}_\epsilon}(\mathsf{a})$ for at least some $\epsilon \in \mathbb{I}$) in Figs. 8 and 10, respectively. $\square$

6.3. Strong faithfulness given monotonic effects

In Subsection 6.2, we show that strong faithfulness is violated by the gradient-based contribution function with respect to all of the surveyed argumentation semantics. In the counter-examples we provide, we can see that intuitively, the contributor we are interested in does not have a *monotonic effect* on the topic argument: the effect that marginally changing the initial strength of the contributor has on the topic argument may be positive, zero, or negative, depending on the contributor's initial strength (see Fig. 1.2). This means that strong faithfulness must be violated; indeed, from a common sense perspective, the principle *should not be satisfied* in such situations. However, we may consider the satisfaction of strong faithfulness desirable in situations in which the effect of the contributor on the topic argument is monotonic, an intuition which we formalise below.

Definition 6.3 (*Monotonic Effect*). Given a semantics σ, a QBAG $\mathsf{G} = (Args, \tau, Att, Supp)$, and arguments $\mathsf{x}, \mathsf{a} \in Args$, we say that "x has a monotonic effect on a" w.r.t. σ iff for every QBAG $\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon}$, for every $\epsilon, \epsilon' \in \mathbb{I}$ the following holds:

- either, for every QBAG $\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon'}$, if $\epsilon < \epsilon'$ then $\sigma_{\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon}}(\mathsf{a}) \leq \sigma_{\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon'}}(\mathsf{a})$;
- or, for every QBAG $\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon'}$, if $\epsilon < \epsilon'$ then $\sigma_{\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon}}(\mathsf{a}) \geq \sigma_{\mathsf{G}\downarrow_{\tau(\mathsf{x}) \leftarrow \epsilon'}}(\mathsf{a})$.

Now, we can conjecture that Ctrb^∂ satisfies strong faithfulness given contributors that have a monotonic effect on our topic argument, i.e., given the class of QBAGs with respect to a semantics σ in which the effect of any argument on another one in any QBAG is monotonic. As a prerequisite for this conjecture, let us formalise the notion of monotonic effect QBAGs.

Definition 6.4 (*Monotonic Effect QBAGs*). Given a semantics σ, a QBAG $\mathsf{G} = (Args, \tau, Att, Supp)$ is a monotonic effect QBAG w.r.t. σ iff for all $\mathsf{x}, \mathsf{a} \in Args$ it holds that x has a monotonic effect on a w.r.t. σ. We denote the class of monotonic effect QBAGs w.r.t. σ by $\mathscr{Q}^\sigma_{ME}$.

Now, we can speculate that the gradient-based contribution function satisfies strong faithfulness with respect to all of the surveyed argumentation semantics when assuming monotonic effect QBAGs.

Conjecture 2. *Let σ be QE, DFQuAD, SD-DFQuAD, EB, or EBT semantics and assume the class of QBAGs is Q^σ_{ME}. Ctrb^∂ satisfies strong faithfulness w.r.t. σ.*

We can further speculate about topological properties that can guarantee a monotonic effect. For example, in some cases, the role that a contributor has in terms of (potentially indirect) attack or support of a topic argument may be clear:

- a contributor may either directly attack or support the topic argument (and hence is a direct attacker or supporter);
- a contributor may attack direct or indirect supporters of the topic argument or support its direct or indirect attackers and hence is an indirect attacker;
- a contributor may attack direct or indirect attackers of the topic argument or support its direct or indirect supporters and hence is an indirect supporter;
- finally, a contributor is a clear attacker of a topic argument if it is a direct or indirect attacker but not a direct or indirect supporter and it is a clear supporter if it is a direct or indirect supporter but not a direct or indirect attacker.

Now, we may speculate that contributors that are clear supporters of a topic argument must have an effect on the topic argument that is always ≥ 0, no matter the contributor's initial strength; in the case of clear attackers the effect should be always ≤ 0.
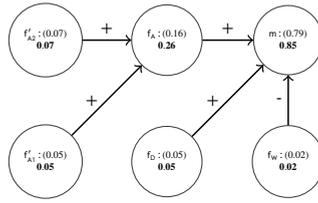
Fig. 26. QBAG for movie rating with initial strengths in normal font and DFQuAD final strength in bold font (taken from [5]).

Table 5
The results of four contribution functions for the QBAG in Fig. 26.

Argument	$\text{Ctrb}^{\mathscr{R}}$	$\text{Ctrb}^{\mathscr{R}'}$	$\text{Ctrb}^{\mathscr{S}}$	Ctrb^{∂}
f_A	0.051443 (1)	0.03192 (1)	0.044726 (1)	0.176258 (1)
f'_{A1}	0.007792 (3,4)	0.007792 (3,4)	0.004065 (4)	0.15585 (3)
f'_{A2}	0.011144 (2)	0.011144 (2)	0.00577 (3)	0.1592 (2)
f_D	0.007792 (3,4)	0.007792 (3,4)	0.011248 (2)	0.15584 (4)
f_W	-0.0042 (5)	-0.0042 (5)	-0.00807 (5)	-0.21 (5)

7. Case study

To give a basic intuition of practical applications, we apply our four contribution functions to an application of QBAGs to explaining the aggregated score of movie ratings on the website Rotten Tomatoes [5]. The QBAGs in this application are based on a hierarchical structure with the overall evaluation at the top. The evaluation can be attacked and supported by evaluation criteria like the quality of acting, directing and writing. The criteria can be further attacked and supported by subcriteria like the performance of individual actors or the writing for particular parts of the movie. The initial strength of the criteria is determined by a natural language processing pipeline, which extracts argumentative phrases from reviews, evaluates their sentiment using sentiment analysis tools, assigns them to criteria using topic classification tools and merges the sentiment of the individual phrases such that the final strength of the evaluation aligns with the Rotten Tomatoe score [5].

Fig. 26 shows a small example QBAG for this scenario. Here, m is the topic argument representing a movie to be rated. The attacker f_W of m stands for the script quality of the movie (writing), while the supporters f_D and f_A represent the quality of directors and actors of the movie, respectively. f_A is further supported by two specific actors f'_{A1} and f'_{A2}. The QBAG is evaluated by DFQuAD semantics and the movie m obtains a final strength of 0.85.

Our four contribution functions can explain in different ways how the criteria affect the final strength of the movie (m's final strength). In Table 5, we provide the contributions of all non-topic arguments determined by $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{R}'}$, $\text{Ctrb}^{\mathscr{S}}$ and Ctrb^{∂}, with the ranking established by the contribution function in parentheses. Note that all contribution functions agree qualitatively: f_W always has a negative contribution, whereas all other arguments have positive contributions. We can also see that f_A always has the largest positive contribution. However, the contribution ranking for the remaining arguments varies across different contribution functions because they measure different things. The removal-based contribution functions $\text{Ctrb}^{\mathscr{R}}$ and $\text{Ctrb}^{\mathscr{R}'}$ measure the difference between the movie's strength when arguments are removed in a particular way. The Shapley values determined by $\text{Ctrb}^{\mathscr{S}}$ have a similar meaning but average over all possible subgraphs and are normalized such that the contribution values sum up to the difference between the movie's final and initial strengths. Finally, Ctrb^{∂} measures the local sensitivity of the movie's strength with respect to changes in the other arguments' initial strength. More precisely, a small change in the arguments' initial strength will change the strength of the movie by a value that is proportional to the argument's contribution value.

Let us use the example to discuss the intuition of some principles. Since $\text{Ctrb}^{\mathscr{S}}$ satisfies (quantitative) contribution existence, Proposition 5.4 guarantees that at least one argument must have a non-zero Shapley-based contribution because the initial strength of m is different from its final strength. Let us note that, in our example, this is also true for all other functions even though they do not satisfy this principle in general (it is possible that all contribution values are 0 even though there was a change). Another interesting property to note for the Shapley values is that their sum equals the difference between the initial and final strengths of m. In this way, it nicely distributes the contributions among the arguments. Ctrb^{∂} satisfies quantitative local faithfulness, as shown in Propositions 5.10, which guarantees that the scores accurately reflect the sensitivity of the final evaluation with respect to their initial strength. Since f_{A1}, f_{A2} and f_D all have very similar scores under Ctrb^{∂}, we learn that the final evaluation is approximately equally sensitive with respect to their initial strengths. More specifically, changing their initial strengths by some small value δ will change the final strength of the movie by approximately $0.15 \cdot \delta$. Similarly, changing f_A and f_W would result in a change of approximately $0.17 \cdot \delta$ and $-0.21 \cdot \delta$, respectively. Since $\text{Ctrb}^{\mathscr{R}}$ satisfies (quantitative) counterfactuality, Proposition 5.22 allows us to infer that removing the criterion f_A would decrease the movie's final strength of m by approximately 0.05. All contribution functions satisfy directionality according to Proposition 5.9. This allows us to derive some contributions without making any computations. For example, if we consider f_A as the topic argument, the contribution from f_D to f_A must be 0 under all four contribution functions due to the absence of a path from f_D to f_A.

8. Discussion and related work

The study presented in this paper can be considered a contribution to argumentative explainability, with a focus on quantitative bipolar argumentation. The former has emerged as a research trend in recent years [2,3,14,34,35]; in applied research where argumentation is utilized, quantitative (bipolar) argumentation is often employed as the formal argumentation approach of choice [4,36,18,15,5,16,17]. However, only a few works focus on the formal study of explaining inference in quantitative bipolar argumentation. Notable exceptions are [32], which introduces several explanation notions for sets of argument explaining changes in inference in QBAGs after updates have been applied to the argumentation graph, and [13], which is the first paper introducing contribution functions to quantitative bipolar argumentation. Let us note that some related contribution functions have been introduced for gradual semantics over non-bipolar abstract argumentation before [34,19]). We refer to [37] for an overview of semantics in this area. [14] studies gradient-based contribution functions under Df-QuAD semantics in more detail. [38] introduces a function that intuitively corresponds to the counterfactual contribution function; however, the study presented in [38] does not focus on single-argument contributions but on the aggregation of contributions of arguments uttered by a specific agent.

Our work can be seen as a consolidation and extension of some of the theoretical contributions provided in [13] and [14], providing a rigorous analysis of previously introduced and new principles that highlights the differences between contribution functions, thus facilitating the well-informed selection of contribution functions in application scenarios. From our analysis, we can see that for each of the three contribution functions $\text{Ctrb}^{\mathscr{R}}$, $\text{Ctrb}^{\mathscr{S}}$, and Ctrb^{∂}, there is a contribution function principle that is satisfied by only this function and with respect to all of the surveyed semantics: (quantitative) counterfactuality for $\text{Ctrb}^{\mathscr{R}}$, quantitative contribution existence for $\text{Ctrb}^{\mathscr{S}}$, and (quantitative) local faithfulness for Ctrb^{∂}. In the case of quantitative contribution existence, a relaxation of the principle—*contribution existence*—is satisfied by other contribution functions as well, albeit merely with respect to some, and not all, of the surveyed semantics. For $\text{Ctrb}^{\mathscr{R}'}$, we did not define a corresponding contribution function principle: let us claim that adjusting the counterfactuality principle to fit the behaviour of $\text{Ctrb}^{\mathscr{R}'}$ is trivially possible but arguably pointless, as the principle would be deliberately designed to be satisfied by $\text{Ctrb}^{\mathscr{R}'}$ (and only by $\text{Ctrb}^{\mathscr{R}'}$). What remains to be studied are the conjectures posed in Section 6, as well as the satisfaction of contribution function principles in the case of cyclic QBAGs.

CRediT authorship contribution statement

Timotheus Kampik: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Nico Potyka:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Xiang Yin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Kristijonas Čyras:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Francesca Toni:** Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

Appendix A. Counter-examples proving the violation of proximity

The appendix contains counterexamples showing that many contribution functions violate the proximity principle with respect to many of the surveyed contribution functions. The only exception is Ctrb^{∂} and EBT semantics, for which we do not provide a proof of violation (or satisfaction).

Proposition A.1. *$\text{Ctrb}^{\mathscr{R}}$ violates proximity w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics.*

Proof. For the counterexample for QE semantics, consider Fig. 25, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\text{Ctrb}^{\mathscr{R}}_a(b)| \approx 0.0012 < |\text{Ctrb}^{\mathscr{R}}_a(c)| \approx 0.0037$. For DFQuAD semantics, consider Fig. 27, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\text{Ctrb}^{\mathscr{R}}_a(b)| = 0 < |\text{Ctrb}^{\mathscr{R}}_a(c)| \approx 0.05$. For SD-DFQuAD
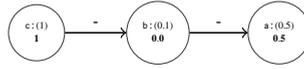
Fig. 27. $\mathrm{Ctrb}^{\mathscr{R}}$ violates the proximity principle w.r.t. DFQuAD semantics.
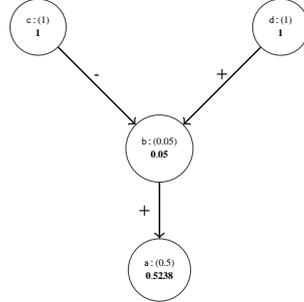

Fig. 28. $\mathrm{Ctrb}^{\mathscr{R}}$ violates the proximity principle w.r.t. SD-DFQuAD semantics.
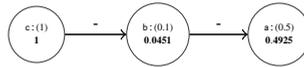

Fig. 29. $\mathrm{Ctrb}^{\mathscr{R}}$ violates the proximity principle w.r.t. EB and EBT semantics.

Fig. 30. $\mathrm{Ctrb}^{\mathscr{R}'}$ violates the proximity principle w.r.t. QE semantics.
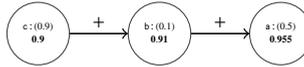

Fig. 31. $\mathrm{Ctrb}^{\mathscr{R}'}$ and Ctrb^{∂} violate the proximity principle w.r.t. DFQuAD semantics.

semantics, consider Fig. 28, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}}(b)| \approx 0.0238 < |\mathrm{Ctrb}_a^{\mathscr{R}}(c)| \approx 0.1483$. For EB and EBT semantics, consider Fig. 29, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}}(b)| \approx 0.0075 < |\mathrm{Ctrb}_a^{\mathscr{R}}(c)| \approx 0.0089$. $\square$

Proposition A.2. $\mathrm{Ctrb}^{\mathscr{R}'}$ *violates proximity w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics.*

Proof. For the counterexample for QE semantics, consider Fig. 30, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}'}(b)| \approx 0.005 < |\mathrm{Ctrb}_a^{\mathscr{R}'}(c)| \approx 0.0959$. For DFQuAD semantics, consider Fig. 31, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}'}(b)| \approx 0.05 < |\mathrm{Ctrb}_a^{\mathscr{R}'}(c)| = 0.405$. For SD-DFQuAD semantics, consider Fig. 32, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}'}(b)| \approx 0.0454 < |\mathrm{Ctrb}_a^{\mathscr{R}'}(c)| \approx 0.127$. For EB and EBT semantics, consider Fig. 33, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^{\mathscr{R}'}(b)| \approx 0.0169 < |\mathrm{Ctrb}_a^{\mathscr{R}'}(c)| \approx 0.0184$. $\square$

Proposition A.3. $\mathrm{Ctrb}^{\mathscr{S}}$ *violates proximity w.r.t. QE, DFQuAD, SD-DFQuAD, EB, and EBT semantics.*

Proof. For the counterexample for QE semantics, consider Fig. 34, topic argument a and contributors e and f. Although e is strictly closer to a than f is, we have $|\mathrm{Ctrb}_a^{\mathscr{S}}(e)| \approx 0.00005 < |\mathrm{Ctrb}_a^{\mathscr{S}}(f)| \approx 0.00056$. For DFQuAD semantics, consider Fig. 35, topic argument a and contributors e and f. Although e is strictly closer to a than f is, we have $|\mathrm{Ctrb}_a^{\mathscr{S}}(e)| \approx 0.0037 < |\mathrm{Ctrb}_a^{\mathscr{S}}(f)| \approx 0.0057$. For SD-DFQuAD semantics, consider Fig. 36, topic argument a and contributors e and f. Although e is strictly closer to a than f is, we have $|\mathrm{Ctrb}_a^{\mathscr{S}}(e)| \approx 0.00065 < |\mathrm{Ctrb}_a^{\mathscr{S}}(f)| \approx 0.00075$. For EB semantics, consider Fig. 37, topic argument a and contributors e and f. Although e is strictly closer to a than f is, we have $|\mathrm{Ctrb}_a^{\mathscr{S}}(e)| \approx 0.00022 < |\mathrm{Ctrb}_a^{\mathscr{S}}(f)| \approx 0.00026$. For EB semantics, consider Fig. 38, topic argument a and contributors e and f. Although e is strictly closer to a than f is, we have $|\mathrm{Ctrb}_a^{\mathscr{S}}(e)| \approx 0.000098 < |\mathrm{Ctrb}_a^{\mathscr{S}}(f)| \approx 0.000108$. $\square$

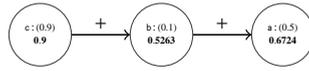**Proposition A.4.** Ctrb^{∂} *violates proximity w.r.t. QE, DFQuAD, SD-DFQuAD, and EB semantics.*

Fig. 32. Ctrb$^{\mathscr{R}'}$ violates the proximity principle w.r.t. SD-DFQuAD semantics.

Fig. 33. Ctrb$^{\mathscr{R}'}$ violates the proximity principle w.r.t. EB and EBT semantics.
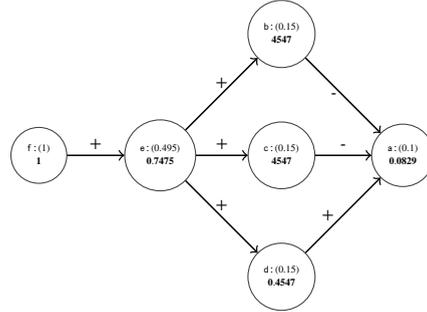

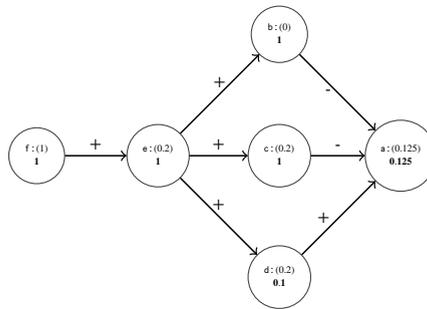**Fig. 34.** Ctrb$^{\mathscr{S}}$ violates the proximity principle w.r.t. QE semantics.

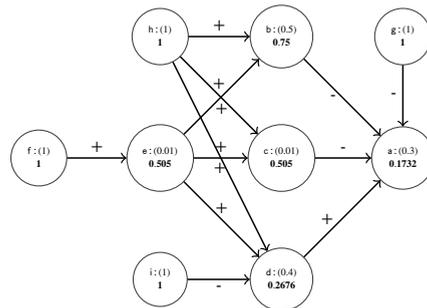**Fig. 35.** Ctrb$^{\mathscr{S}}$ violates the proximity principle w.r.t. DFQuAD semantics.

Fig. 36. Ctrb$^{\mathscr{S}}$ violates the proximity principle w.r.t. SD-DFQuAD semantics.

Proof. For the counterexample for QE semantics, consider Fig. 39, topic argument a and contributors b and d. Although b is strictly closer to a than d is, we have $|\mathrm{Ctrb}_a^\partial(b)| \approx 0.1081 < |\mathrm{Ctrb}_a^\partial(d)| \approx 0.2945$. For DFQuAD semantics, consider Fig. 31, topic argument a and contributors b and c. Although b is strictly closer to a than c is, we have $|\mathrm{Ctrb}_a^\partial(b)| \approx 0.05 < |\mathrm{Ctrb}_a^\partial(c)| = 0.45$. For SD-DFQuAD semantics, consider Fig. 40, topic argument a and contributors b and d. Although b is strictly closer to a than d is, we have $|\mathrm{Ctrb}_a^\partial(b)| \approx 0.121 < |\mathrm{Ctrb}_a^\partial(d)| \approx 0.234$. For EB semantics, consider Fig. 41, topic argument a and contributors b and d. Although b is strictly closer to a than d is, we have $|\mathrm{Ctrb}_a^\partial(b)| \approx 0.0083 < |\mathrm{Ctrb}_a^\partial(d)| \approx 0.0101$. $\square$
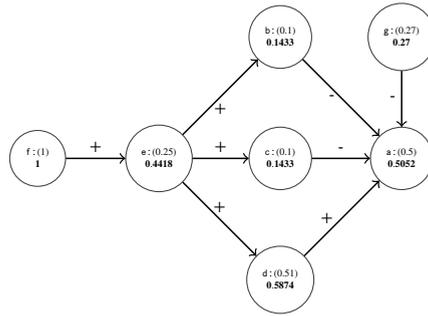
Fig. 37. Ctrb$^{\mathscr{S}}$ violates the proximity principle w.r.t. EB semantics.
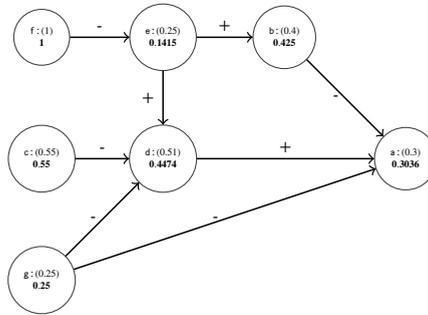

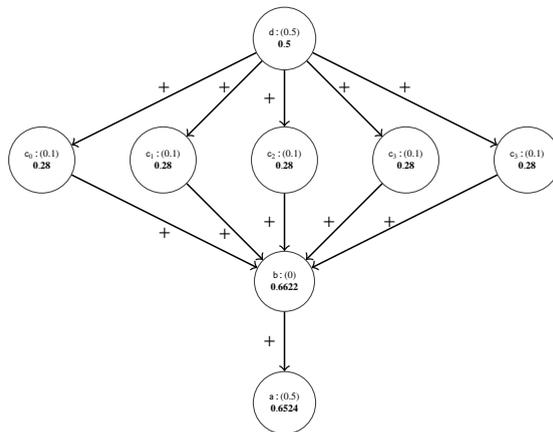**Fig. 38.** Ctrb$^{\mathscr{S}}$ violates the proximity principle w.r.t. EBT semantics.

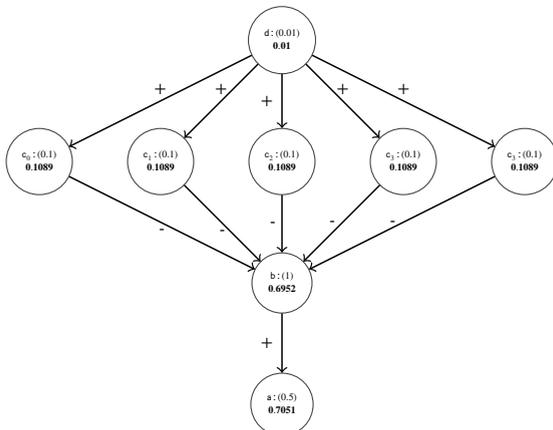**Fig. 39.** Ctrb$^{\partial}$ violates the proximity principle w.r.t. QE semantics.

Fig. 40. Ctrb$^{\partial}$ violates the proximity principle w.r.t. SD-DFQuAD semantics.
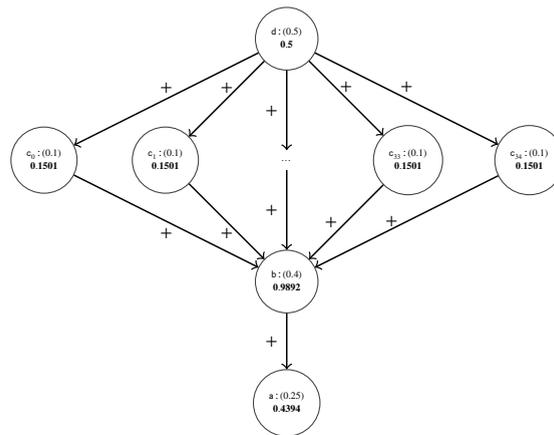
Fig. 41. Ctrb$^\partial$ violates the proximity principle w.r.t. EB semantics.

References

[1] E.I. Sklar, M.Q. Azhar, Explanation through argumentation, in: Proceedings of the 6th International Conference on Human-Agent Interaction, HAI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 277–285.

[2] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: a survey, in: Z.-H. Zhou (Ed.), 30th International Joint Conference on Artificial Intelligence, IJCAI, Montreal, 2021, pp. 4392–4399, arXiv:2105.11266.

[3] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, Knowl. Eng. Rev. 36 (2021) e5, https://doi.org/10.1017/S0269888921000011.

[4] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: fantastic explanations and how to find them, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18, AAAI Press, 2018, pp. 1949–1955.

[5] O. Cocarascu, A. Rago, F. Toni, Extracting dialogical explanations for review aggregations with argumentative dialogical agents, in: E. Elkind, M. Veloso, N. Agmon, M.E. Taylor (Eds.), Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13–17, 2019, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1261–1269, http://dl.acm.org/citation.cfm?id=3331830.

[6] N. Potyka, X. Yin, F. Toni, Explaining random forests using bipolar argumentation and Markov networks, in: B. Williams, Y. Chen, J. Neville (Eds.), Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, 2023, pp. 9453–9460.

[7] H. Ayoobi, N. Potyka, F. Toni, Sparx: sparse argumentative explanations for neural networks, in: K. Gal, A. Nowé, G.J. Nalepa, R. Fairstein, R. Radulescu (Eds.), ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems, PAIS 2023, in: Frontiers in Artificial Intelligence and Applications, vol. 372, IOS Press, 2023, pp. 149–156.

[8] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[9] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[10] A. Bloniarz, A. Talwalkar, B. Yu, C. Wu, Supervised neighborhoods for distributed nonparametric regression, in: Artificial Intelligence and Statistics, PMLR, 2016, pp. 1450–1459.

[11] G. Plumb, D. Molitor, A. Talwalkar, Model agnostic supervised local explanations, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 2520–2529.

[12] N. Potyka, X. Yin, F. Toni, Towards a theory of faithfulness: faithful explanations of differentiable classifiers over continuous data, arXiv preprint, arXiv:2205.09620.

[13] K. Čyras, T. Kampik, Q. Weng, Dispute trees as explanations in quantitative (bipolar) argumentation, in: K. Čyras, T. Kampik, O. Cocarascu, A. Rago (Eds.), 1st International Workshop on Argumentation for EXplainable AI Co-Located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022, in: CEUR Workshop Proceedings, vol. 3209, 2022, pp. 1–12, CEUR-WS.org, https://ceur-ws.org/Vol-3209/0872.pdf.

[14] X. Yin, N. Potyka, F. Toni, Argument attribution explanations in quantitative bipolar argumentation frameworks, in: K. Gal, A. Nowé, G.J. Nalepa, R. Fairstein, R. Radulescu (Eds.), ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023), in: Frontiers in Artificial Intelligence and Applications, vol. 372, IOS Press, 2023, pp. 2898–2905.

[15] N. Kotonya, F. Toni, Gradual argumentation evaluation for stance aggregation in automated fake news detection, in: 6th Workshop on Argument Mining, 2019, pp. 156–166.

[16] P. Li, D. Gabbay, X. Chi, Y. Cheng, Decision making with weighted quantitative argumentation based on regression, in: Logics for New-Generation AI, 2022, pp. 59–71.

[17] H. Chi, Y. Lu, B. Liao, L. Xu, Y. Liu, An optimized quantitative argumentation debate model for fraud detection in e-commerce transactions, IEEE Intell. Syst. 36 (2) (2021) 52–63.

[18] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: C. Baral, J.P. Delgrande, F. Wolter (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016, AAAI Press, 2016, pp. 63–73, http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874.

[19] J. Delobelle, S. Villata, Interpretability of gradual semantics in abstract argumentation, in: G. Kern-Isberner, Z. Ognjanovic (Eds.), 15th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, in: LNAI, vol. 11726, Springer, Belgrade, 2019, pp. 27–38.

[20] L.S. Shapley, Notes on the N-Person Game, Rand Corporation, 1951.

[21] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, Int. J. Approx. Reason. 99 (2018) 39–55.

[22] N. Potyka, Extending modular semantics for bipolar weighted argumentation, in: N. Agmon, E. Elkind, M.E. Taylor, M. Veloso (Eds.), 18th International Conference on Autonomous Agents and MultiAgent Systems, IFAAMAS, Montreal, 2019, pp. 1722–1730.

[23] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: a principled spectrum, Int. J. Approx. Reason. 105 (2019) 252–286, https://doi.org/10.1016/j.ijar.2018.11.019.

[24] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argument Comput. 6 (1) (2015) 24–49.

[25] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: M. Thielscher, F. Toni, F. Wolter (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018, AAAI Press, 2018, pp. 148–157, https://aaai.org/ocs/index.php/KR/KR18/paper/view/17985.

[26] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI), AAAI Press, 2021, pp. 6463–6470.

[27] T. Mossakowski, F. Neuhaus, Modular semantics and characteristics for bipolar weighted argumentation graphs, arXiv preprint, arXiv:1807.06685.

[28] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: Michael Thielscher, F. Toni, F. Wolte (Eds.), Principles of Knowledge Representation and Reasoning: 16th International Conference, 2018, pp. 148–157.

[29] L. Shapley, A value for n-person games, Ann. Math. Stud. 28 (1953) 307–318.

[30] A. Davies, P. Veličković, L. Buesing, S. Blackwell, D. Zheng, N. Tomašev, R. Tanburn, P. Battaglia, C. Blundell, A. Juhász, M. Lackenby, G. Williamson, D. Hassabis, P. Kohli, Advancing mathematics by guiding human intuition with AI, Nature 600 (7887) (2021) 70–74, https://doi.org/10.1038/s41586-021-04086-x.

[31] B. Verheij, Two approaches to dialectical argumentation: admissible sets and argumentation stages, in: Proc. NAIC 96, 1996, pp. 357–368.

[32] T. Kampik, K. Čyras, J. Ruiz Alarcón, Change in quantitative bipolar argumentation: sufficient, necessary, and counterfactual explanations, Int. J. Approx. Reason. (2023) 109066, https://doi.org/10.1016/j.ijar.2023.109066, https://www.sciencedirect.com/science/article/pii/S0888613X23001974.

[33] N. Potyka, A tutorial for weighted bipolar argumentation with continuous dynamical systems and the Java library attractor, in: International Workshop on Non-Monotonic Reasoning (NMR), 2018.

[34] L. Amgoud, J. Ben-Naim, S. Vesic, Measuring the intensity of attacks in argumentation graphs with Shapley value, in: C. Sierra (Ed.), Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, 2017, pp. 63–69, ijcai.org.

[35] X. Yin, N. Potyka, F. Toni, Explaining arguments' strength: unveiling the role of attacks and supports, in: International Joint Conference on Artificial Intelligence (IJCAI), 2024, in press, https://arxiv.org/abs/2404.14304.

[36] A. Rago, O. Cocarascu, C. Bechlivanidis, D. Lagnado, F. Toni, Argumentative explanations for interactive recommendations, Artif. Intell. 296 (2021) 103506, https://doi.org/10.1016/j.artint.2021.103506.

[37] L. Amgoud, D. Doder, S. Vesic, Evaluation of argument strength in attack graphs: foundations and semantics, Artif. Intell. 302 (2022) 103607, https://doi.org/10.1016/J.ARTINT.2021.103607.

[38] A. Himeur, B. Yun, P. Bisquert, M. Croitoru, Assessing the impact of agents in weighted bipolar argumentation frameworks, in: M. Bramer, R. Ellis (Eds.), Artificial Intelligence XXXVIII - 41st SGAI International Conference on Artificial Intelligence, Proceedings, AI 2021, Cambridge, UK, December 14-16, 2021, in: Lecture Notes in Computer Science, vol. 13101, Springer, 2021, pp. 75–88.