ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230603

Argument Attribution Explanations in Quantitative Bipolar Argumentation Frameworks

Xiang Yin[a;*], **Nico Potyka**[b,a] **and Francesca Toni**[a]

[a]Imperial College London, UK
[b]Cardiff University, UK
ORCiD ID: Xiang Yin https://orcid.org/0000-0002-6096-9943, Nico Potyka https://orcid.org/0000-0003-1749-5233, Francesca Toni https://orcid.org/0000-0001-8194-1459

Abstract. Argumentative explainable AI has been advocated by several in recent years, with an increasing interest on explaining the reasoning outcomes of Argumentation Frameworks (AFs). While there is a considerable body of research on qualitatively explaining the reasoning outcomes of AFs with debates/disputes/dialogues in the spirit of *extension-based semantics*, explaining the quantitative reasoning outcomes of AFs under *gradual semantics* has not received much attention, despite widespread use in applications. In this paper, we contribute to filling this gap by proposing a novel theory of *Argument Attribution Explanations (AAEs)* by incorporating the spirit of feature attribution from machine learning in the context of Quantitative Bipolar Argumentation Frameworks (QBAFs): whereas feature attribution is used to determine the influence of features towards outputs of machine learning models, AAEs are used to determine the influence of arguments towards *topic argument*s of interest. We study desirable properties of AAEs, including some new ones and some partially adapted from the literature to our setting. To demonstrate the applicability of our AAEs in practice, we conclude by carrying out two case studies in the scenarios of fake news detection and movie recommender systems.

1 Introduction

Explainable AI (XAI) is playing an increasingly important role in AI towards safety, reliability and trustworthiness [1]. Various methods have been proposed in this field for providing explanations for several AI algorithms, models, and systems (e.g. see recent overviews [30, 1]).

A popular category of explanation methods is *feature attribution*, aiming at assigning a "feature importance score" to each input feature fed to the AI of interest (notably machine learning models), denoting its contribution to the output decision by the AI. Feature attribution methods include, amongst others, LIME [39], SHAP [29], SILO [12] and gradient-scoring [8]. Explanations returned by feature attribution methods are intuitive in that they focus on explaining the outputs in terms of the inputs alone, making it unnecessary to go into the details of the inner mechanism of the underlying AI. Furthermore, feature attribution explanations are easy for people to understand by just checking the positive or negative influence of the input features towards the outputs and the ranking of the magnitude of the scores.

Alongside feature attribution methods, in recent years *argumentative XAI* is increasingly showing benefits for various forms of AI (see [21, 44] for overviews). Basically, argumentative XAI applies computational argumentation [7] to extract *argumentation frameworks (AFs)* as skeletons for explanations. For example, [20] uses abstract AFs [23] to explain the outputs of schedulers, [34] proposes to use weighted bipolar AFs to explain multi-layer perceptrons and [17] propose to use *Quantitative Bipolar AFs (QBAFs)* [10] to explain movie review aggregations. Whereas feature attribution methods focus on the input-output behaviour of the underlying AI, AFs as explanations point to the dialectical relationships among arguments, abstractly representing interactions among the inner components of the underlying AI encoded by the AFs. These AFs provide a natural mechanism for users to interact with the AI [37] and may help find irrationalities in the underlying AI to aid debugging and improving the AI [26].

Existing forms of argumentative XAI are predominantly *qualitative* in that they focus on explaining the reasoning outcomes of AFs with debates/disputes/dialogues in the spirit of *extension-based semantics* (e.g. as in [23]). These qualitative explanations mirror interactions within the inner mechanism of the underpinning AI as dialectical exchanges between arguments. For example, [20] use 'explanation via (non-)attacks' and [17] define explanations as template-driven dialogues using attacks and supports in the AFs. Instead, explaining the *quantitative* reasoning outcomes of AFs under *gradual semantics* (e.g. those proposed in [10, 34]) has not received much attention, in spite of the widespread use of this form of semantics in several applications (e.g. fake news detection [28], movie recommendations [17] and fraud detection [16]). However, in many application settings, it is important to see how arguments in AFs topically influence one another, and how much positive/negative influence is transmitted from one argument to another. This is especially the case when explanations are needed for a *topic argument* of interest (e.g. an argument corresponding to the output of a classifier as in [2, 34]) and it is essential to assess which arguments have more importance towards the topic argument.

In this paper, we contribute to filling this gap by proposing a novel theory of *Argument Attribution Explanations (AAEs)* by incorporating the spirit of feature attribution from machine learning in the context of QBAFs. With respect to qualitative explanations alone, AAEs allow to measure and compare the contribution of different arguments towards topic arguments in QBAFs under the Discontinuity Free Quantitative Argumentation Debate (*DF-QuAD*) gradual semantics [38], even when the comparison is difficult with qualitative explanations alone. This is the case with large QBAFs as visualized in Figure 1 from [16], where it is hard to see how quantitative explanations can be intuitively

* Corresponding Author. Email: x.yin20@imperial.ac.uk

delivered for the children of the root as topic arguments.[1] Additionally, AAEs take the *base scores* of arguments in QBAFs into account. Different base scores should give rise to different explanations, but qualitative explanations disregard base scores, as they only consider the QBAFs' structure regardless of quantitative information.

Figure 1. Fraud detection in e-commerce [16]. (Our emphasis is on the QBAF's size/complexity rather than contents, so readability is not a concern).

In this paper, we formalize AAEs and analyze some qualitative and quantitative guarantees for different types of *connectivity* in QBAFs. We then study several desirable properties of AAEs as explanations, including some adapted from the literature and some novel ones. Finally, we show the applicability of AAEs in two scenarios: fake news detection [28] and movie recommender systems [17]. Overall, the contribution of this paper is threefold.

- We propose the novel theory of AAEs (Section 4).
- We study (new and adapted) desirable properties of AAEs as explanations (Section 5).
- We show applicability of AAEs in practice (Section 6).

The proofs of all results are in https://arxiv.org/abs/2307.13582.

2 Related Work

Argumentative XAI Following [21], we can distinguish two types of argumentative explanations. One is *intrinsic* argumentative explanations, whereby the explained underlying models themselves are already AFs. Examples include recommender systems [13, 40] built with suitable AFs specified in DeLP [25] and generating recommendations by the reasoning outcomes of the AFs. The other is *post-hoc* argumentative explanations, for underlying models that are not based on AFs and are not argumentative in spirit. In order to extract argumentative explanations from these models, it is crucial to first extract AFs, that is, extract arguments and dialectical relations while identifying an appropriate argumentation semantics matching the models' behaviour. Depending on the representational extent of the AFs for the underlying models, post-hoc explanations can be further divided into two sub-types: *complete* and *approximate* argumentative explanations [21]. Complete argumentative explanations suit many settings, including decision-making systems [45], knowledge-based systems [6] and scheduling [20]. In the case of approximate argumentative explanations, the mapping process from the underlying models to AFs is incomplete, in the sense that AFs are extracted from parts of the underlying models rather than the whole models. For example, [41] first extract rules from trained neural networks, and construct AFs based on these rules, and [2] abstract away trained neural networks as QBAFs by treating groups of neurons as arguments and understanding feature attribution methods as a gradual semantics. Our argumentative explanations, in the form of AAEs, assume an AF as a starting point, in the form of a QBAF under the DF-QuAD gradual semantics, and are usable alongside any existing form of argumentative XAI based on the same AFs and semantics.

Feature Attribution Methods The intuition of feature attribution methods is to measure the contribution of the features to the output by feature attribution scores. ***LIME*** [39] is a model-agnostic explanation method, which can locally explain any instance (input-output pair) by linear approximation, learning a linear surrogate model based on sampling around the instance of interest. ***SILO*** [12] shares the same idea of LIME, but differs from it in two aspects. First, LIME uses synthetic data generated by the approximation method to fit the linear surrogate model, while SILO directly uses the data from the training set. Second, for LIME, the weight of each synthetic data point is decided by the distance, while for SILO, the weight is decided by calling a random forest classifier. ***SHAP*** [29] is another popular model-agnostic feature attribution method, which is theoretically guaranteed by game theory [24] and satisfies several desirable properties. SHAP computes the marginal contribution of the features as their attribution scores. However, SHAP is computationally inefficient because of the combinatorial combination of features. ***Gradient-scoring*** [8] is also an attribution method especially suitable for differentiable classifiers with continuous data, which demonstrates a greater accuracy and efficiency than LIME, SHAP and other popular attribution methods in such settings [35]. Our AAEs are in the spirit of gradient scoring, but take a QBAF under the DF-QuAD semantics as a starting point.

Properties of Explanations [14] summarize popular properties of explanations in the literature into four categories. Two of them are related to this paper. The first category is *robustness/sensitivity*, amounting to the robustness of explanation methods under small perturbations of the inputs. The second category is *faithfulness/fidelity*, concerning the loss between the explanation model and the underlying model. Many properties in the literature are proposed to measure faithfulness, which is core for attribution explanation methods [14].

For argumentative explanations, with a few exceptions (notably [4]), properties are not well-studied so far [21]. Some of the existing properties borrow ideas behind general explanation properties. For instance, *fidelity* [20] is used to measure whether the extracted AF is faithful to the underlying model. Also, *transparency* and *trust* are studied as cognition-related properties of explanations [37]. We contribute some novel properties of argumentative explanations while also adapting some existing ones for argumentative explanations (e.g. *explainability* from [4]) and for standard feature attribution (e.g. faithfulness).

3 Background

We recall the definition of QBAF [10], a form of quantitative bipolar AFs [5], and the DF-QuAD gradual semantics [38].

Definition 1 (QBAF). *A* QBAF *is a quadruple* $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$*, where* $\mathcal{A}$ *is the set of* arguments*;* $\mathcal{R}^-$ *is the* attack *relation (*$\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$*);* $\mathcal{R}^+$ *is the* support *relation (*$\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$*);* $\mathcal{R}^-$ *and* $\mathcal{R}^+$ *are disjoint;* τ *is the* base score *(*$\tau : \mathcal{A} \to [0, 1]$*).*

We often denote the structure of QBAFs graphically, where nodes represent the arguments and edges the support and attack relations. We label edges in attack and support relations with $-$ and $+$, respectively. Figure 2 shows an example. A QBAF is called *acyclic* if the corresponding graph is acyclic. As in [38], we restrict attention to *acyclic QBAFs*.

[1] AAEs for this example can be found in https://arxiv.org/abs/2307.13582.

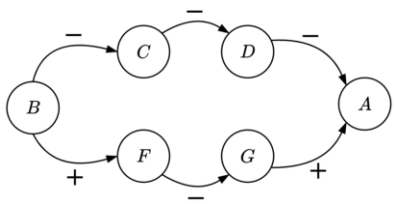

Figure 2. An example QBAF structure.

The *dialectical strength* of arguments in QBAFs can be evaluated by several gradual semantics $\sigma : \mathcal{A} \to [0,1]$, e.g. as defined in [10, 3, 32, 34]. The explainability of some gradual semantics like the h-categorizer semantics [11] and the counting semantics [36] has been studied in [22]. In this paper, we focus on explaining the DF-QuAD gradual semantics [38]. This problem has been neglected in the literature so far even though DF-QuAD has broad applicability [38, 28, 17, 16, 15] in settings where explainability is important.

In DF-QuAD, for any argument A, $\sigma(A)$ is defined as follows:

$$\sigma(A) = \begin{cases} \tau(A) - \tau(A) \cdot (v_{Aa} - v_{As}) & if\ v_{Aa} \geq v_{As} \\ \tau(A) + (1 - \tau(A)) \cdot (v_{As} - v_{Aa}) & if\ v_{Aa} < v_{As} \end{cases}$$

where v_{Aa} is the *aggregation strength* of all the attackers against A, while v_{As} is the *aggregation strength* of all the supporters for A. v_{Aa} and v_{As} are defined as follows:

$$v_{Aa} = 1 - \prod_{\{X \in \mathcal{A} | (X,A) \in \mathcal{R}^-\}} (1 - \sigma(X));$$

$$v_{As} = 1 - \prod_{\{X \in \mathcal{A} | (X,A) \in \mathcal{R}^+\}} (1 - \sigma(X)).$$

An example of applying DF-QuAD is as follows.

Example 1. *Consider the acyclic QBAF $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ with arguments and relations as in Figure 2 and $\tau(X) = 0.5$ for all $X \in \mathcal{A}$. It is easy to see that $v_{Aa} = 0.375$, $v_{As} = 0.125$, $v_{Ba} = v_{Bs} = 0$, $v_{Ca} = 0.5$, $v_{Cs} = 0$, $v_{Da} = 0.25$, $v_{Ds} = 0$, $v_{Fa} = 0$, $v_{Fs} = 0.5$, $v_{Ga} = 0.75$, $v_{Gs} = 0$. Then, $\sigma(A) = 0.375$, $\sigma(B) = 0.5$, $\sigma(C) = 0.25$, $\sigma(D) = 0.375$, $\sigma(F) = 0.75$, $\sigma(G) = 0.125$.*

In the remainder, unless specified otherwise, we will assume an acyclic QBAF $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ and σ given by DF-QuAD.[2]

4 Argument Attribution in QBAFs

We formally propose the theory of AAEs in acyclic QBAFs evaluated by DF-QuAD gradual semantics. Then, we define *connectivity* in QBAFs and study the properties of *direct* and *indirect* connectivity *qualitatively* and *quantitatively*, which will play a role in studying the properties of AAEs as explanations in Section 5.

Our AAEs are inspired by gradient-based feature attribution explanations in machine learning (e.g. see [43, 42]) and gradient-based *contribution function* in [19], which, in our setting, capture the *sensitivity* of the dialectical strength of a *topic argument* with respect to the base score of other arguments, as follows.

Definition 2 (Argument Attribution Explanations (AAEs))**.** *Let $A, B \in \mathcal{A}$, where A is called the* topic argument. *For a perturbation $\varepsilon \in [-\tau(B), 0) \cup (0, 1 - \tau(B)]$, let $\mathcal{Q}_\varepsilon$ be the QBAF resulting from $\mathcal{Q}$ by perturbing the base score $\tau(B)$ to $\tau_\varepsilon(B) = \tau(B) + \varepsilon$. Let $\sigma, \sigma_\varepsilon$ denote DF-QuAD for $\mathcal{Q}, \mathcal{Q}_\varepsilon$, respectively. The* AAE[3] *from B to A is*

$$\nabla|_{B \mapsto A} = \lim_{\varepsilon \to 0} \frac{\sigma_\varepsilon(A) - \sigma(A)}{\varepsilon}.$$

As an illustration, consider $\mathcal{Q}$ in Figure 2. The perturbation of $\tau(B)$ will give rise to a perturbation of $\sigma(A)$. Then, we can use the limit of the ratio of the perturbation, as defined above, to represent the AAE from B to A (see https://arxiv.org/abs/2307.13582 for details).

We next distinguish three possible types of influence.

Definition 3 (Attribution Influence)**.** *We say that the* attribution influence *from B to A is* positive *if $\nabla|_{B \mapsto A} > 0$,* negative *if $\nabla|_{B \mapsto A} < 0$ and* neutral *if $\nabla|_{B \mapsto A} = 0$.*

In order to analyze the influence of an argument on a topic argument, we have to consider the different paths connecting them. For this purpose, we define some terminology next.

Definition 4 (Path and Path Set)**.** *Let $X, Y \in \mathcal{A}$ and $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+$. A* path *$\phi$ between X and Y is a sequence $X_1, \ldots, X_n$ of arguments in $\mathcal{A}$ such that $n \geq 2$, $X_1 = X, X_n = Y$ and $(X_i, X_{i+1}) \in \mathcal{R}$ for $1 \leq i \leq n-1$. We refer to $X_2, \ldots, X_{n-1}$ as the* middle arguments *and to $m_\phi = n - 2$ as the number of middle arguments in path ϕ. We let $\Phi_{X \mapsto Y}$ denote the set of all paths ϕ from X to Y, and $|\Phi_{X \mapsto Y}|$ denote the number of paths in $\Phi_{X \mapsto Y}$.*

We distinguish four types of connectivity based on the number of paths between two arguments.

Definition 5 (Connectivity)**.** *For any $A, B \in \mathcal{A}$:*

- *B is* disconnected *from A iff $|\Phi_{B \mapsto A}| = 0$;*
- *B is* directly *connected to A iff $|\Phi_{B \mapsto A}| = 1$ and $m_\phi = 0$ for $\phi \in \Phi_{B \mapsto A}$;*
- *B is* indirectly *connected to A iff $|\Phi_{B \mapsto A}| = 1$ and $m_\phi \geq 1$ for $\phi \in \Phi_{B \mapsto A}$;*
- *B is* multifold *connected to A iff $|\Phi_{B \mapsto A}| > 1$.*

Here, we show an example to explain connectivity in QBAFs.

Example 2. *In Figure 2, C is disconnected from F; D is directly connected to A because there is only one path from D to A, and no middle arguments in between. C is indirectly connected to A because there is only one single path connecting C and A, and D is the only middle argument in the path. B is multifold connected to A because two paths connecting B to A exist.*

Next, we give some qualitative and quantitative guarantees for AAEs that will be useful to prove properties of AAEs later.

We start by showing that AAEs correctly capture the qualitative effect (positive or negative) of direct connectivity, in that arguments attacking (supporting) the topic argument always have negative (positive, respectively) or zero attribution scores.

Proposition 1 (Direct Qualitative Attribution Influence)**.** *If $B, A \in \mathcal{A}$ are directly connected, then*

1. *If $(B, A) \in \mathcal{R}^-$, then $\nabla|_{B \mapsto A} \leq 0$;*
2. *If $(B, A) \in \mathcal{R}^+$, then $\nabla|_{B \mapsto A} \geq 0$.*

The next proposition gives an exact quantification of the influence of arguments B directly attacking or supporting the topic argument A, in terms of three parameters: the base score of A, the aggregation strength of B and the strength of other arguments Z that are in the same relation with A as B.

[2] Note that acyclic QBAFs are not restricted to trees, as shown in Figure 2.

[3] The well-definedness of AAE is shown in Proposition 5.

Proposition 2 (Direct Quantitative Attribution Influence). *If $B, A \in \mathcal{A}$ are directly connected and $(B, A) \in \mathcal{R}^*$, for $\mathcal{R}^* = \mathcal{R}^-$ or $\mathcal{R}^* = \mathcal{R}^+$, then*

$$\nabla|_{B \mapsto A} = \xi_B(1 - |v_{Ba} - v_{Bs}|) \prod_{\{Z \in \mathcal{A} \setminus B | (Z,A) \in \mathcal{R}^*\}} [1 - \sigma(Z)],$$

where

$$\xi_B = \begin{cases} -\tau(A) & if\ \mathcal{R}^* = \mathcal{R}^- \wedge v_{Aa} \geq v_{As}; \\ (\tau(A) - 1) & if\ \mathcal{R}^* = \mathcal{R}^- \wedge v_{Aa} < v_{As}; \\ \tau(A) & if\ \mathcal{R}^* = \mathcal{R}^+ \wedge v_{Aa} > v_{As}; \\ (1 - \tau(A)) & if\ \mathcal{R}^* = \mathcal{R}^+ \wedge v_{Aa} \leq v_{As}. \end{cases}$$

Inspired by the chain rule of gradients, we further study some qualitative and quantitative guarantees of indirect connectivity in the next two propositions. First, we find that one argument indirectly connected to a topic argument always has positive (negative) or neutral influence if the number of attacks in the path between the argument and the topic argument is even (odd, respectively).

Proposition 3 (Indirect Qualitative Attribution Influence). *Let $X_1, \ldots, X_n \in \mathcal{A}$ $(n \geq 3)$ and $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+$. Suppose $S = \{(X_1, X_2), (X_2, X_3), \ldots, (X_{n-1}, X_n)\} \subseteq \mathcal{R}$. Let $|S \cap \mathcal{R}^-| = \Theta$. If $X_1, X_n \in \mathcal{A}$ are indirectly connected through path $\phi = X_1, \ldots, X_n$, then*

1. *If Θ is odd, then $\nabla|_{X_1 \mapsto X_n} \leq 0$;*
2. *If Θ is even, then $\nabla|_{X_1 \mapsto X_n} \geq 0$.*

Then we show that the AAE from one argument X_1 to an indirectly connected topic argument X_n can be precisely characterized in terms of the attribution scores from the arguments in the path from X_1 to X_n and the strengths of these arguments.

Proposition 4 (Indirect Quantitative Attribution Influence). *Let $X_1, \ldots, X_n \in \mathcal{A}$. If $X_1, X_n \in \mathcal{A}$ are indirectly connected through path $\phi = X_1, \ldots, X_n$, then*

$$\nabla|_{X_1 \mapsto X_n} = (1 - |v_{X_1 a} - v_{X_1 s}|) \cdot \prod_{i=1}^{n-1} \frac{\nabla|_{X_i \mapsto X_{i+1}}}{(1 - |v_{X_i a} - v_{X_i s}|)}.$$

We illustrate next the application of the propositions in this section (see https://arxiv.org/abs/2307.13582 for more examples).

Example 3. *The settings are the same as in Example 1. According to Proposition 1, we have $\nabla|_{D \mapsto A} \leq 0$ because $(D, A) \in \mathcal{R}^-$. According to Proposition 2, we have*

$$\nabla|_{D \mapsto A} = -\tau(A) \cdot (1 - |v_{Da} - v_{Ds}|) \cdot \prod_{\{Z \in \mathcal{A} \setminus D | (Z,A) \in \mathcal{R}^-\}} (1 - \sigma(Z)) = -0.375.$$

Similarly, $\nabla|_{C \mapsto D} = -0.25$. According to Proposition 3, we have $\nabla|_{C \mapsto A} \geq 0$ because the number of attacks is two (even). According to Proposition 4, we have

$$\nabla|_{C \mapsto A} = \nabla|_{C \mapsto D} \times \frac{\nabla|_{D \mapsto A}}{1 - |v_{Da} - v_{Ds}|} = 0.125.$$

5 Properties

We now study some desirable properties [4] of our AAEs as explanations for outcomes under the DF-QuAD gradual semantics. Although the properties that AAEs can and should satisfy depend, of course, on the underlying semantics, our properties make guarantees about the explanations, not about the underlying semantics.

We start with the *explainability* property from [4], which guarantees that an explanation always exist. This property, for AAEs, amounts to well-definedness, and can be formulated as follows.

Proposition 5 (Explainability). *$\forall A, B \in \mathcal{A}$, $\nabla|_{B \mapsto A} \in \mathbb{R}$ is well-defined.*

Missingness is an important property for attribution methods, guaranteeing the *faithfulness* of the explanation for non-relevant features [29]. Here, it states that if one argument is not connected to the topic argument, then the AAE is zero.

Proposition 6 (Missingness). *$\forall A, B \in \mathcal{A}$, if B is disconnected from A, then*

$$\nabla|_{B \mapsto A} = 0.$$

The next four properties guarantee the *faithfulness* of AAEs to the underlying QBAF. The former two consider the faithfulness of one particular argument while the latter two consider the faithfulness between arguments.

We propose *completeness*, inspired by [42] and by *quantitative faithfulness* in [35]. In our setting, it states that the change of the strength of the topic argument should be proportional to its AAE.

Property 1 (Completeness). *Let $A, B \in \mathcal{A}$ and let $\sigma'_B(A)$ denote the strength of A when setting $\tau(B)$ to 0. Then*

$$-\tau(B) \cdot \nabla|_{B \mapsto A} = \sigma'_B(A) - \sigma(A).$$

Proposition 7. *Completeness is satisfied if B is directly or indirectly connected to A.*

Proposition 8. *Completeness can be violated if B is multifold connected to A.*

Counterfactuality is inspired by [18], which considers the situation of removing arguments. This property states that if one argument B has a positive (negative) influence on A, then removing B will decrease (increase, respectively) the strength of A.

Property 2 (Counterfactuality). *Let $A, B \in \mathcal{A}$ and let $\sigma'_B(A)$ denote the strength of A when setting $\tau(B)$ to 0. Then*

1. *If $\nabla|_{B \mapsto A} \leq 0$, then $\sigma'_B(A) \geq \sigma(A)$;*
2. *If $\nabla|_{B \mapsto A} \geq 0$, then $\sigma'_B(A) \leq \sigma(A)$.*

Proposition 9. *Counterfactuality is satisfied if B is directly or indirectly connected to A.*

Proposition 10. *Counterfactuality can be violated if B is multifold connected to A.*

We propose *agreement* as a property for comparing AAEs across any two arguments. It states that two arguments have the same influence on the strength of a topic argument whenever they have the same base scores and the same AAEs to the topic argument.

Property 3 (Agreement). *Let $A, B, C \in \mathcal{A}$ and $\sigma'_B(A), \sigma'_C(A)$ denote the strength of A when setting $\tau(B), \tau(C)$ to 0 respectively. If*

$$\left|\tau(B) \cdot \nabla|_{B \mapsto A}\right| = \left|\tau(C) \cdot \nabla|_{C \mapsto A}\right|$$

then

$$\left|\sigma'_B(A) - \sigma(A)\right| = \left|\sigma'_C(A) - \sigma(A)\right|.$$

[4] With 'desirable' we mean properties that a user may demand from an argumentative explanation method.

Proposition 11. *Agreement is satisfied if B and C are directly or indirectly connected to A.*

Proposition 12. *Agreement can be violated if B or C is multifold connected to A.*

Monotonicity states that features with larger attribution scores have a larger influence on the output [14]. In our setting, monotonicity means that the larger the AAE from an argument, the more influence it should have on the strength of the topic argument.

Property 4 (Monotonicity)**.** *Let $A, B, C \in \mathcal{A}$, $\sigma'_B(A), \sigma'_C(A)$ denote the strength of A when setting $\tau(B), \tau(C)$ to 0 respectively. If*

$$\left|\tau(B) \cdot \nabla|_{B \mapsto A}\right| < \left|\tau(C) \cdot \nabla|_{C \mapsto A}\right|$$

then

$$\left|\sigma'_B(A) - \sigma(A)\right| < \left|\sigma'_C(A) - \sigma(A)\right|.$$

Proposition 13. *Monotonicity is satisfied if B and C are directly or indirectly connected to A.*

Proposition 14. *Monotonicity can be violated if B or C is multifold connected to A.*

Although gradient-based attributions such as AAEs are *local* rather than *global* explanations in general, we find some interesting qualitative and quantitative global guarantees for AAEs under direct and indirect connectivity. *Qualitative invariability* states that one argument will always have a positive (or negative) influence on the topic arguments, regardless of the change of its base score, while *quantitative invariability* shows that the attribution score of one argument will keep invariant with the change of its base score.

Property 5 (Qualitative Invariability)**.** *$\forall A, B \in \mathcal{A}$, let ∇_δ denote the AAE from B to A when setting $\tau(B)$ to some $\delta \in [0, 1]$. Then*

1. *If $\nabla|_{B \mapsto A} \leq 0$, then $\forall \delta \in [0, 1]$, $\nabla_\delta \leq 0$;*
2. *If $\nabla|_{B \mapsto A} \geq 0$, then $\forall \delta \in [0, 1]$, $\nabla_\delta \geq 0$.*

Proposition 15. *Qualitative invariability is satisfied if B is directly or indirectly connected to A.*

Proposition 16. *Qualitative invariability can be violated if B is multifold connected to A.*

Property 6 (Quantitative Invariability)**.** *$\forall A, B \in \mathcal{A}$, let ∇_δ denote the AAE from B to A when setting $\tau(B)$ to some $\delta \in [0, 1]$. For all δ, there always exists a constant $C \in [-1, 1]$ such that*

$$\nabla_\delta \equiv C.$$

Proposition 17. *Quantitative invariability is satisfied if B is directly or indirectly connected to A.*

Proposition 18. *Quantitative invariability can be violated if B is multifold connected to A.*

The final property is *tractability*, measuring the computational complexity of AAEs. This property shows that our explanation method is efficient in the sense that the AAEs can be generated in linear time in the number of arguments.

Proposition 19 (Tractability)**.** *If $|\mathcal{A}| = n$, then AAEs can be generated in linear time $\mathcal{O}(n)$.*

Let us note that many of the aforementioned properties do not hold for the multifold connectivity case. We would like to point out that this is a feature, not a bug, for an explanation method that is supposed to be faithful to the underlying argumentation semantics. This is because many properties satisfied by DF-QuAD, like *balance* and *monotonicity*[5] [9], make only guarantees about the effect of direct attackers or supporters on an argument. The effect in the multifold connectivity case depends on various other factors (including base scores of all directly and indirectly connected arguments) and there is an infinite (finite for a fixed number of involved arguments) number of special cases that may occur.

6 Case Studies

We carry out two case studies to show the applicability and usefulness of AAEs in practice.

Case Study 1: Fake News Detection.
Fake news detection plays an important role in avoiding the spread of rumors on social media. In [28], QBAFs are used to detect whether a source tweet is a rumor by aggregating the weight of replies to the source tweet. Concretely, Figure 3 from [28] shows an intuitive example of using QBAFs to detect a rumor tweet. The QBAF in the figure is built up by extracting the dialectical relation between the source tweet and the replies with the help of deep learning techniques. The thread of tweets w.r.t Figure 3 is as follows.

***A** [u1/source tweet] Up to 20 held hostage in Sydney Lindt Cafe siege.*
***B** [u2/reply1] @u1 pretty sure it was*
***C** [u3/reply2] @u1 yeah terrible*
***D** [u4/reply3] @u1 all reports say 13*
***E** [u5/reply4] @u2 nonsense*
***F** [u6/reply5] @u4 this number is ridiculous*
***G** [u7/reply6] @u6 not convincing at all*
***H** [u8/reply7] @u6 you are an insensitive idiot!*

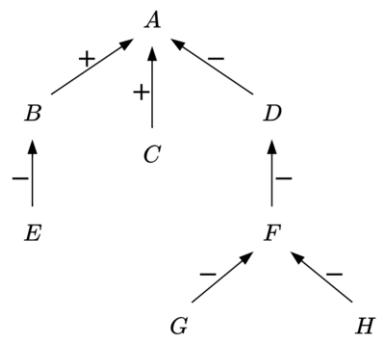

Figure 3. Structure of a QBAF for fake news detection (taken from [28]).

In Figure 3, the task is to detect whether A is a rumor, hence A is the topic argument. B, C and D are the direct replies for A. E is the reply for B; F is the reply for D; G and H are replies for F. Therefore, E, F, G and H are indirect replies for A. Since the authors of [28] mainly focus on the overall performance for fake news detection, they do not provide explanations with respect to this QBAF. Here, we show possible conversation threads as qualitative explanations which are consistent with the intuition of [28], before showing how AAEs can additional provide quantitative explanations.

[5] Here monotonicity refers to guarantees about DF-QuAD, and is different from Property 4, which gives guarantees about the explanations.

The base scores for all the arguments are initially set to 0.5. According to the DF-QuAD gradual semantics, the strength for A is $\sigma(A) = 0.59375$, which is seen as "True News" because $\sigma(A) > 0.5$. Instead of just providing this prediction, qualitative conversational explanations can be obtained from the QBAF structure. For instance, an explanation for the output decision might be:

User: *Why the source tweet A is not a rumor?*
QBAF: *Because replies B and C support A, despite D is against A.*
User: *Why A is still true in spite of the attack of D?*
QBAF: *Although D attacks A, D is also attacked by F, which decreases D's strength when attacking A. Although G and H attack F, F still supports A, just with a weaker strength.*

However, using these conversational explanations alone, it is unclear how much each argument contributes to the final strength of the topic argument. Next, we apply AAEs to explain the outcome for A.

We compute the AAEs in Table 1 (last column), presented in descending order to obtain a ranking. These attributions can also be computed by applying Propositions 2 and 4, because all arguments (except A) are either directly or indirectly connected to A. Based on the ranking, we can obtain qualitative and quantitative analyses that qualitative explanations alone are unable to provide, as follows.

Table 1. AAEs in descending order (last column) for the QBAF in Figure 3.

ARGUMENT	τ	$\sigma'_X(A)$	$\sigma'_X(A) - \sigma(A)$	∇
REPLY2: C	0.5	0.40625	-0.18750	0.3750
REPLY1: B	0.5	0.53125	-0.06250	0.1250
REPLY5: F	0.5	0.56250	-0.03125	0.0625
REPLY6: G	0.5	0.62500	0.03125	-0.0625
REPLY7: H	0.5	0.62500	0.03125	-0.0625
REPLY4: E	0.5	0.65625	0.06250	-0.1250
REPLY3: D	0.5	0.81250	0.21875	-0.4375

Qualitative Analysis B, C and F have a positive influence on A. Indeed B and C directly support A, and a path with two attackers link F to A, so F has a positive influence on A as well. D, E, G and H have a negative influence on A. D directly attacks A, while there is an odd number of attackers from E, G and H to A. This qualitative analysis complements well the qualitative, conversational explanation we saw earlier.

Quantitative Analysis Among arguments with a positive attribution influence, C has the largest influence on A, while the influence of B is less than C because argument E attacks B, thus weakening B's influence. F has the smallest positive influence on A because the influence is indirect, and, at the same time, two attackers of F weaken its positive influence. Of all the arguments with a negative attribution influence, D has the highest influence on A. This is because D not only directly attacks A, but also has indirect supporters from G and H as they attack F, the attacker of D. E has an indirect negative influence on A by attacking its supporter B. Then, G and H have a smaller negative influence than B since they are farther away than B.

Additionally, note that when the base scores of arguments change, the attribution ranking changes accordingly, giving rise to different explanations. However, qualitative explanations always remain identical regardless of the numerical information conveyed by AAEs, because they only focus on the structure of the QBAFs.

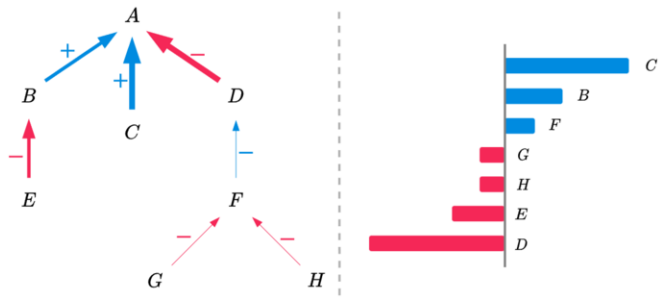

Figure 4. Visualization of AAEs for fake news detection.

Visualization In order to be provide intuitive explanations, we can visualize the AAEs for instance as shown in Figure 4. The left-hand side shows the AAEs to the topic argument. Here, blue arrows show positive attribution while red arrows show negative attribution, and the thickness of arrows shows the magnitude of the attribution. The right-hand side shows ranking and magnitude of the AAEs.

Property Analysis We discuss the satisfaction of the proposed properties to show that AAEs are a good explanation method. In Table 1, we can see that each argument is assigned a real number as attribution to the topic argument, which satisfies *explainability*. If we change the base score of B in the range of $[0, 1]$, the attribution score will not change ($\nabla|_{B\mapsto A} = 0.125 > 0$), which satisfies *invariability* qualitatively and quantitatively. Next, we demonstrate the *faithfulness* of AAEs by analyzing *missingness*, *counterfactuality*, *completeness*, *agreement* and *monotonicity*. For any two arguments without any connections like C and E, the attribution score is 0 ($\nabla|_{C\mapsto E} = 0$), which is intuitive and satisfies *missingness*. For any argument whose attribution score is positive like C, then setting the base score to 0 will decrease the strength of A from 0.59375 to 0.40625, and $\sigma'_C(A) - \sigma(A) = -\tau(C) \cdot \nabla|_{C\mapsto A} = -0.1875$, which satisfies both *counterfactuality* and *completeness*. Any two arguments with the same contributions, like G and H, have the same influence on A. Given that $|\tau(G) \cdot \nabla|_{G\mapsto A}| = |\tau(H) \cdot \nabla|_{H\mapsto A}| = 0.03125$, we have $|\sigma'_G(A) - \sigma(A)| = |\sigma'_H(A) - \sigma(A)|$, which satisfies *agreement*. For those with different contributions, like B and C, due to $|\tau(B) \cdot \nabla|_{B\mapsto A}| = 0.0625 < |\tau(C) \cdot \nabla|_{C\mapsto A}| = 0.1875$, we get $|\sigma'_B(A) - \sigma(A)| = 0.0625 < |\sigma'_C(A) - \sigma(A)| = 0.1875$, guaranteeing *monotonicity*. From the computational angle, all AAEs can be computed in linear time in the number of arguments.

Case Study 2: Movie Recommender Systems.
Online review aggregation has become an increasingly effective quality control method in the era of information explosion. [17] propose to build up movie recommender systems based on QBAFs by aggregating online movie reviews. Basically, they build Argumentative Dialogical Agents (ADAs) to aggregate movie reviews in natural language and then generate ratings of movies from the aggregations, recommending highly rated movies to users. ADAs can also engage in interactive explanations (conversations) with users explaining why recommended movies are highly rated, leveraging on the underlying QBAFs. Let us take movie *The Post* as an example in their paper. Figure 5 shows the QBAF underpinning an ADA for this example, where the topic argument m stands for the movie in question and three features support or attack m (here f_A stands for *acting*, f_D stands for *directing*, and f_W stands for *writing*). Some of these features have sub-features (f_A has sub-features f'_{A1} and f'_{A2}). The base scores in the QBAF result from the review aggregation method employed by

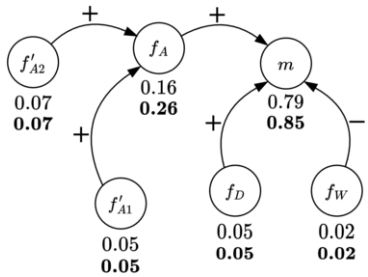

Figure 5. QBAF for movie recommendations (taken from [17]). Values in regular font are base scores and values in bold are DF-QuAD strengths.

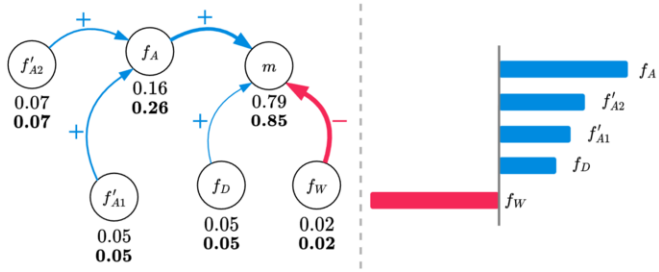

Figure 6. Visualization of AAEs for movie recommendations in ADA.

ADA. ADA then applies DF-QuAD to this QBAF, to obtain the argument strengths indicated in bold in Figure 5. A possible conversational explanation for this example is shown as follows in [17].

***User**: Why was The Post highly rated?*
***ADA**: This is because the acting was really great, although the writing was a little poor.*
***User**: Why was the acting great?*
***ADA**: Because actor Meryl Streep was great.*
***User**: What did critics say about Meryl Streep being great?*
***ADA**: "...Streep's hesitations, ... are soul-deep..."*

As before, we can apply AAEs to further explain m in this example. We give the AAEs in Table 2 and visualizations in Figure 6.

Table 2. AAEs in descending order (last column) for the QBAF in Figure 5.

ARGUMENT	τ	$\sigma'_X(m)$	$\sigma'_X(m) - \sigma(m)$	∇
ACTING	0.16	0.81954	-0.02820	0.17625
ACTOR2	0.07	0.83660	-0.01114	0.15920
ACTOR1	0.05	0.83995	-0.00779	0.15585
DIRECTING	0.05	0.83995	-0.00779	0.15584
WRITING	0.02	0.85194	0.0042	-0.21

Qualitative Analysis According to Propositions 1 and 3, f_A and f_D have a positive influence on m because they are supporters; f_W has a negative influence on m because it is an attacker of m. f'_{A1} and f'_{A2} both have positive influence on m, because they are supporters of a supporter of m, and 0 (even) attackers from them to m.

Quantitative Analysis f_A and its sub-features have the most positive influence on m, which means *acting* is the most important feature for *The Post*. The attribution of f_D is slightly less than for f_A and f_W has the highest negative influence on m. Despite *writing*'s negative influence, *acting* and *directing* still make the movie high-rated.

Property Analysis Here, we mainly focus on the *faithfulness* of AAEs. Any disconnected arguments, like f'_{A2} and f_D, should be assigned attribution scores as 0, which satisfies *faithfulness* of AAEs. From Table 2, removing any arguments with positive (negative) influence on m will give rise to a decrease (increase) of the strength of m, which guarantees the *faithfulness* of AAEs from the perspective of one single argument. For example, if $\tau(f_A)$ is set to 0, then $\sigma'_{f_A}(m)$ will decrease, hence $\sigma'_{f_A}(m) - \sigma(m) < 0$. Arguments f_A and f_D are assigned different contributions, therefore their influence on m is different. For instance, $\left|\tau(f_A) \cdot \nabla|_{f_A \mapsto m}\right| = 0.02820 > \left|\tau(f_D) \cdot \nabla|_{f_D \mapsto m}\right| = 0.00779$, hence $\left|\sigma'_{f_A}(m) - \sigma(m)\right| = 0.02820 > \left|\sigma'_{f_D}(m) - \sigma(m)\right| = 0.00779$, which guarantees the *faithfulness* of AAEs from the perspective of comparing the two arguments.

7 Conclusions and Future Works

We introduced AAEs as quantitative explanations for acyclic QBAFs equipped with DF-QuAD gradual semantics and showed that they satisfy several desirable properties. Inspired by feature attribution methods [8], AAEs quantify the contribution of arguments to the final strength of a topic argument. As our analysis shows, the scores are faithful in the sense that they represent the true effects on the argument (e.g. they satisfy *missingness* and *completeness*). Furthermore, they are computationally efficient as they can be computed in linear time with respect to the number of arguments in the QBAF. Finally, as a first proof of concept, we demonstrated the applicability of AAEs as quantitative explanations in two simple case studies in fake news detection and movie recommender systems, emphasizing their added value against qualitative explanations such as conversations.

We are planning to extend this work in four directions. First, it would be interesting to extend the scope of AAEs from individual influences to the collective influence of a set of arguments on the same topic argument. Second, we would like to study AAEs for QBAFs equipped with other gradual semantics, like Euler-based [3] and quadratic energy semantics [32]. Although our proposed properties are tailored to AAEs under the DF-QuAD gradual semantics, they also have the potential to be evaluated for other quantitative argumentative explanations due to their generality (e.g. *counterfactuality* and *monotonicity*). Third, we would like to extend our approach to cyclic QBAFs. Let us note that this is not straightforward because the strength values in cyclic QBAFs are defined by an iterative procedure that does not necessarily converge [31, 33]. Fourth, we would like to carry out a number of diverse experiments to further improve the human-friendliness of AAEs, which depends on the humans, their expertise and the field of application.

Acknowledgements

This research was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX) and by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors.