Explaining Arguments’ Strength: Unveiling the Role of Attacks and Supports

Xiang Yin¹, Nico Potyka², Francesca Toni¹
¹Department of Computing, Imperial College London, UK
²School of Computer Science and Informatics, Cardiff University, UK
{xy620, ft}@imperial.ac.uk, pottykan@cardiff.ac.uk

Abstract
Quantitatively explaining the strength of arguments under gradual semantics has recently received increasing attention. Specifically, several works in the literature provide quantitative explanations by computing the attribution scores of arguments. These works disregard the importance of attacks and supports, even though they play an essential role when explaining arguments’ strength. In this paper, we propose a novel theory of Relation Attribution Explanations (RAEs), adapting Shapley values from game theory to offer fine-grained insights into the role of attacks and supports in quantitative bipolar argumentation towards obtaining the arguments’ strength. We show that RAEs satisfy several desirable properties. We also propose a probabilistic algorithm to approximate RAEs efficiently. Finally, we show the application value of RAEs in fraud detection and large language models case studies.

1 Introduction
Explainable Artificial Intelligence (XAI) has received increasing attention in fields such as finance and healthcare, which demand a reliable and legitimate reasoning process. Argumentation Frameworks (AFs), e.g. as first studied in [Dung, 1995], are promising tools in the XAI field [Mittelstadt et al., 2019] due to their transparency and interpretability, as well as their ability to support reasoning about conflicting information [ˇCyras et al., 2021; Albini et al., 2020; Potyka, 2021; Potyka et al., 2023; Ayoobi et al., 2023]. In Quantitative Bipolar AFs (QBAFs) [Baroni et al., 2015], each argument has a base score, and its final strength is computed by gradual semantics based on the strength of its attackers and supporters [Baroni et al., 2019]. QBAFs can be deployed to support several applications. For example, [Cocarascu et al., 2019] build QBAFs to rate movies by aggregating movie reviews. The QBAFs have a hierarchical structure, where the goodness of movies is at the top and influenced by arguments about criteria like the quality of acting and directing. These criteria/arguments, in turn, can be affected by subcriteria/subarguments like the performance of particular actors. In this application, the base scores of arguments are obtained from reviews via a natural language processing pipeline; finally, a gradual semantics is applied to determine the final strength of movies as their rating scores.

While the gradual semantics of a QBAF provides an assessment of arguments (e.g., when using QBAFs for aggregating movie reviews, the rating scores of movies), we may also be interested in an intuitive understanding of the underlying reasoning process. This leads to an interesting research question initially raised by [Delobelle and Villata, 2019]: given an argument of interest (topic argument) in a QBAF, how to explain the reasoning outcome (i.e., the strength) of this topic argument?

Most current approaches in the literature address this question by defining argument-based attribution explanations [Delobelle and Villata, 2019; ˇCyras et al., 2022; Yin et al., 2023], which explain the strength of the topic argument by assigning attribution scores to arguments: the greater the attribution score, the greater the argument’s contribution to the topic argument. However, in many cases, more fine-grained relation-based attribution explanations (RAEs) may be beneficial, or even necessary. For illustration, consider Figure 1, and assume that the QBAF (partially) depicted therein results from aggregating movie reviews as in [Cocarascu et al., 2019], where α is a movie to be rated (topic argument). Here, the review β has a positive argument attribution score by supporting the famous actor γ and the influential director δ, which attacks bad directing ξ, but this argument view conceals the fact that β also weakens α by attacking its genre η, which supports the topic argument. In contrast, (our) RAEs give more fine-grained insights: although β has a positive contribution via r₁

¹We give concrete values for the RAEs in Figure 1 in arxiv.org/abs/2404.14304.
and $r_3$ to $\alpha$, it also has a negative contribution via $r_6$.
Motivated by the aforementioned considerations, we make the following contributions:

- We propose a novel theory of RAEs (Section 4).
- We study desirable properties of RAEs under several gradual semantics (Section 5).
- We propose a probabilistic algorithm to efficiently approximate RAEs (Section 6).
- We carry out two case studies to demonstrate the practical usefulness of RAEs (Section 7).

The proofs of all results are in arxiv.org/abs/2404.14304.

### 2 Related Work

[Cýras et al., 2022] propose the general idea of contribution functions that compute quantitative contributions from one argument to another under a given gradual semantics for QBAFs and study three such functions, described below.

The removal-based contribution function proposed by [Delobelle and Villata, 2019] measures how the strength of the topic argument changes if an argument is removed. In general, removal-based explanations are simple and intuitive for users to understand without a high cognitive burden. However, a problem with them is that removing an argument will also remove paths from its predecessor to the argument. The measure can therefore overestimate the contribution of an argument. To solve this problem, [Delobelle and Villata, 2019] propose to cut off the direct relations to an argument before removing it, to obtain the mere contribution of this argument.

The gradient-based contribution function captures the sensitivity of the topic argument w.r.t. another argument. It is based on the partial derivative of the topic argument’s strength w.r.t. the base score of another argument. Arguments with high sensitivity are seen as important. Following this idea, [Yin et al., 2023] further explored the gradient-based contribution function under the DF-QuAD gradual semantics [Rago et al., 2016] and studied its properties in this setting.

The Shapley-based contribution function uses the Shapley value from coalitional game theory [Shapley, 1951] to assign high sensitivity are seen as important. Following this idea, on the spirit of the contribution function of [Amgoud et al., 2016] to assign attribution scores to its arguments. To solve this problem, [Delobelle and Villata, 2019] measures how the strength of the topic argument changes if an argument is removed. In general, removal-based explanations are simple and intuitive for users.

#### 3 Preliminaries

To begin with, we recall the definition of QBAFs. We focus on QBAFs with strength values in the domain $\mathbb{I} = [0, 1]$.

**Definition 1 (QBAF).** A Quantitative Bipolar AF (QBAF) is a quadruple $Q = (A, R^-, R^+, \tau)$ where:

- $A$ is a set of arguments;
- $R^- \subseteq A \times A$ is a binary attack relation;
- $R^+ \subseteq A \times A$ is a binary support relation;
- $\tau: A \rightarrow \mathbb{R}$ is a function assigning base scores to arguments.

QBAFs are often denoted graphically (see Figure 1 as an example), where arguments are nodes and edges show the attack or support relations, labelled by $-$ and $+$, respectively. The base scores can be seen as a priori strengths of arguments when ignoring all other arguments (and is omitted from graphical representations, as in Figure 1). Seeing QBAFs as graphs allows us to use standard notions such as that of path.

In the remainder, unless specified otherwise, we assume as given a generic QBAF $Q = (A, R^-, R^+, \tau)$ for $\mathbb{I} = [0, 1]$. Also, we let $R = R^- \cup R^+$ and, for any $\alpha \in A$, $R(\alpha) = \{ (\alpha, \beta) \in R \mid \beta \in A \}$ denotes the set of all outgoing edges from $\alpha$.

Gradual semantics evaluate QBAFs by a function $\sigma: A \rightarrow \mathbb{R}$ that assigns a (final) strength to every argument (e.g. see [Leite and Martins, 2011; Baroni et al., 2015; Amgoud and Ben-Naim, 2018]). In most approaches, $\sigma$ is defined via an iterative process that initializes all strength values with the base scores and then updates the strength values based on the strength of attackers and supporters. The final strength is the limit of this process. The process is guaranteed to converge for acyclic graphs after at most $n = |A|$ iterations\(^3\) [Potyka, 2019] and, in practice, also quickly converges for cyclic graphs [Potyka, 2018]. Since we aim to explain the strength, which is only possible when it is defined, we will assume convergence for all arguments in the remainder, amounting to the following.

**Definition 2 (Well-definedness).** A gradual semantics $\sigma$ is well-defined for $Q$ iff $\sigma(\alpha)$ exists for every $\alpha \in A$.

**Example 1.** Consider the QBAF in Figure 2 where the base scores of all arguments are set to 0.5. Then, the DF-QuAD gradual semantics [Rago et al., 2016], denoted by $\sigma^D$, determines the following strengths:\(^4\) $\sigma^D(\alpha) = 0.8048675$, $\sigma^D(\beta) = 0.375$, $\sigma^D(\gamma) = 0.375$, $\sigma^D(\delta) = 0.25$, $\sigma^D(\zeta) = 0.5$.

We will often need to restrict QBAFs to a subset of the edges or change the base score function, as follows.

---

\(^3\)Since the strength of an argument can only be affected by its parents, it actually suffices to update each argument only once by following a topological ordering of the arguments [Potyka, 2019].

\(^4\)We omit details on how gradual semantics determine strengths, as we focus on explaining these strengths. For details see arxiv.org/abs/2404.14304.
Then, for any gradual semantics, which is a variant of various notions pro-
stricted Euler-based (REB) [Amgoud and Ben-Naim, 2018] (Definition 5)
relation attribution explanations define our In order to find a fair and reasonable attribution method for cyclic QBAFs.
Conjecture 1. DF-QuAD, QE, REB satisfy monotonicity in DF-QuAD, Quadratic Energy (QE) [Potyka, 2018], Re-
that these semantics are also monotonic for cyclic QBAFs. We conjecture
tonic iff for any \( Q \) \( \tau \) \( Q \) \( \tau \) satisfies monotonicity in acyclic QBAFs. We conjecture
Figure 2: An example of QBAF (where all base scores are set to 0.5).

**Definition 3.** For \( S \subseteq \mathcal{R} \), let \( Q_{|S}(\alpha) = (A, R \cap S, R^+ \cap S, \tau) \). For \( \tau' : A \to I \) a base score function, let \( Q_{|\tau'}(\alpha) = (A, R^-, R^+ \vee S, \tau') \). Then, for any \( \alpha \in A \), we let \( \sigma_{Q_{|S}}(\alpha) \) denote the strength of \( \alpha \) in \( Q_{|S}(\alpha) \) and \( \sigma_{Q_{|\tau'}}(\alpha) \) denote the strength of \( \alpha \) in \( Q_{|\tau'}(\alpha) \).

For illustration, in Figure 2, suppose \( S = \{r_1, r_3, r_5\} \). Then \( r_2 \) and \( r_4 \) are not considered when computing \( \sigma_{Q_{|S}}(\alpha) \).

We will consider the following monotonic property of gradual semantics, which is a variant of various notions proposed in the literature (see [Baroni et al., 2019]) suitable for our setting. Roughly speaking, it states that base scores and relations monotonically influence arguments as one would intuitively expect.

**Definition 4 (Monotonicity).** A gradual semantics \( \sigma \) is monotonic if for any \( \alpha, \beta \in A \) such that \( \alpha \neq \beta \) and \( \mathcal{R}(\beta) = \{ (\beta, \alpha) \} \), for any \( \tau' : A \to I \):

1. If \( (\beta, \alpha) \in \mathcal{R}^- \), then \( \sigma(\alpha) \leq \sigma_{\mathcal{R} \setminus \{\beta, \alpha\}}(\alpha) \);
2. If \( (\beta, \alpha) \in \mathcal{R}^+ \), then \( \sigma(\alpha) \geq \sigma_{\mathcal{R} \setminus \{\beta, \alpha\}}(\alpha) \);
3. If \( (\beta, \alpha) \in \mathcal{R}^- \), \( \tau(\beta) \leq \tau'(\beta) \) and \( \tau(\gamma) = \tau'(\gamma) \) for all \( \gamma \in A \setminus \{\beta\} \), then \( \sigma(\alpha) \leq \sigma_{\tau'}(\alpha) \);
4. If \( (\beta, \alpha) \in \mathcal{R}^+ \), \( \tau(\beta) \leq \tau'(\beta) \) and \( \tau(\gamma) = \tau'(\gamma) \) for all \( \gamma \in A \setminus \{\beta\} \), then \( \sigma(\alpha) \leq \sigma_{\tau'}(\alpha) \).

**Proposition 1.** DF-QuAD, QE, REB satisfy monotonicity in acyclic QBAFs.

**Conjecture 1.** DF-QuAD, QE, REB satisfy monotonicity in cyclic QBAFs.

### 4 Relation Attribution Explanations

In order to explain the strength of a topic argument in a QBAF, we quantify the contribution of all edges to the topic argument. In order to find a fair and reasonable attribution method for quantifying these contributions, we build up on the Shapley-value as in [Amgoud et al., 2017; Čyras et al., 2022]. We define our relation attribution explanations as follows.

**Definition 5 (RAEs).** Let \( \alpha \in A \) be a topic argument and \( r \in \mathcal{R} \). We define the Relation Attribution Explanation (RAE) from \( r \) to \( \alpha \) under \( \sigma \) as:

\[
\phi_{\sigma}^r(\alpha) = \sum_{\mathcal{S} \subseteq \mathcal{R} \setminus \{r\}} \frac{|\mathcal{R}| - |\mathcal{S}| - 1)! |\mathcal{S}|!}{|\mathcal{R}|!} [\sigma_{\mathcal{S} \cup \{r\}}(\alpha) - \sigma_{\mathcal{S}}(\alpha)].
\]

Intuitively, \( \phi_{\sigma}^r(\alpha) \) looks at every subset of edges \( (\mathcal{S}) \) and computes the marginal contribution of \( r \) ([\( \sigma_{\mathcal{S} \cup \{r\}}(\alpha) - \sigma_{\mathcal{S}}(\alpha) \)]). This marginal contribution is weighted by the probability that a random permutation of the edges starts with the subset \( (\mathcal{S}) \) and is followed by \( r \). The main difference between our definition and that in [Amgoud et al., 2017] lies in the potential “causes” of topic arguments. We attribute the strength of a topic argument to all edges (direct and indirect causes) in the QBAF while [Amgoud et al., 2017] attribute it only to the directly incoming edges (direct causes). Furthermore, our definition is suitable not only for attacks but also for supports [Cayrol and Lagasque-Schiex, 2013], which are important in applications [Delobelle and Villata, 2019].

Qualitatively, we distinguish three different (relation) contributions based on the sign of \( \phi_{\sigma}^r(\alpha) \).

**Definition 6 ((Relation) Contribution).** Let \( \alpha \in A \) and \( r \in \mathcal{R} \).

1. If \( \phi_{\sigma}^r(\alpha) > 0 \), we say \( r \) has a positive contribution to \( \alpha \);
2. If \( \phi_{\sigma}^r(\alpha) < 0 \), we say \( r \) has a negative contribution to \( \alpha \);
3. If \( \phi_{\sigma}^r(\alpha) = 0 \), we say \( r \) has a neutral contribution to \( \alpha \).

**Example 2 (Cont.).** Consider again the QBAF in Figure 2 under the DF-QuAD gradual semantics as in Example 1. Let \( \alpha \) be the topic argument. We compute RAEs by Definition 5:

\[
\phi_{\sigma}^{DF}(r_1) = 0.16875 > 0, \phi_{\sigma}^{DF}(r_2) = 0.16875 > 0, \phi_{\sigma}^{DF}(r_3) \approx -0.0318 < 0, \phi_{\sigma}^{DF}(r_4) \approx -0.0318 < 0, \phi_{\sigma}^{DF}(r_5) \approx 0.0307 > 0.
\]

Hence, \( r_1, r_2 \) and \( r_5 \) have a positive contribution to \( \alpha \) while \( r_3 \) and \( r_4 \) have a negative one. We can also see that \( r_1 \) and \( r_2 \) have a more positive contribution to \( \alpha \) than \( r_5 \). We visualize the RAEs in Figure 3.

**Figure 3:** Contributions, drawn from RAEs, for topic argument \( \alpha \) for the QBAF in Figure 2. (Blue/red edges denote positive/negative contributions, respectively. The thickness of edges represents the magnitude of their contributions, i.e. their RAE value.)

### 5 Properties

We now study some properties of RAEs. We start with Shapley-based properties that basically adapt to our setting properties of Shapley values and then move to argumentative properties that we deem interesting in our setting.

#### 5.1 Shapley-based Properties

Similar to [Amgoud et al., 2017], we first transfer the four basic properties of Shapley-values [Shapley, 1951] to our setting. Efficiency is recognized as a desirable property for attribution methods [Ancona et al., 2017]. In our context, it states that the sum of all RAEs corresponds to the deviation of the topic argument’s strength \( \sigma(\alpha) \) from its base score \( \tau(\alpha) \) (namely the explanation distributes the responsibility for the difference among the edges). To prove this property, we assume that the semantics satisfies the Stability property [Amgoud and Ben-Naim, 2018], which states that the final strength of an argument is its base score whenever it has no incoming edges.

**Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)**

3624
Proposition 2 (Efficiency). If $\sigma$ satisfies Stability, then for all $\alpha \in A$ and $r \in R$: $\sigma(\alpha) = \tau(\alpha) + \sum_{r \in R} \phi^{\alpha}_{\sigma}(r)$.

As an illustration, in Example 2, $\sigma^{DF}(\alpha) = 0.8046875$ equals to the the sum of $\tau(\alpha) = 0.5$ and the RAEs for all five edges in $Q$: $\sum_{r \in R} \phi^{\alpha}_{DF}(r) = 0.3046875$.

Efficiency has an interesting implication that is called Justification in [Cyraș et al., 2022]. This demands that whenever the strength of the topic argument differs from its base score, then there is a non-zero RAE explaining the difference.

Corollary 1 (Justification). Let $\alpha \in A$ and $r \in R$.
1. If $\sigma(\alpha) > \tau(\alpha)$, then $\exists r \in R$ such that $\phi^{\alpha}_{\sigma}(r) > 0$;
2. If $\sigma(\alpha) < \tau(\alpha)$, then $\exists r \in R$ such that $\phi^{\alpha}_{\sigma}(r) < 0$.

As an illustration, in Example 2, we have $\sigma^{DF}(\alpha) > \tau(\alpha)$ and $r_1$ with $\phi^{\alpha}_{DF}(r_1) = 0.16875 > 0$ justifies the difference.

Dummi, also known as Missingness in [Lundberg and Lee, 2017], guarantees that if an edge does not make any contribution to the topic argument, then its RAE is 0.

Proposition 3 (Dummi). Let $\alpha \in A$ and $r \in R$. If $\sigma_{\mathcal{S}\cup\{i\}}(\alpha) = \sigma_{\mathcal{S}}(\alpha)$ holds for all $\mathcal{S} \subseteq \mathcal{R}$, then $\phi^{\alpha}_{\sigma}(r) = 0$.

As an illustration, in Example 2, if we explain $\sigma^{DF}(\beta)$ (i.e. the topic argument is $\beta$), then $\phi^{\alpha}_{DF}(r_1) = 0$.

Symmetry states that if two edges share the same contribution to the topic argument, then their RAEs are equal.

Proposition 4 (Symmetry). Let $\alpha \in A$ and $r_1, r_2 \in R$ with $r_1 \neq r_2$. If $\sigma_{\mathcal{S}\cup\{i\}}(\alpha) = \sigma_{\mathcal{S}\cup\{j\}}(\alpha)$ holds for any $\mathcal{S} \subseteq \mathcal{R} \setminus \{r_1, r_2\}$, then $\phi^{\alpha}_{\sigma}(r_1) = \phi^{\alpha}_{\sigma}(r_2)$.

As an illustration, in Example 2, $r_1$ and $r_2$ have symmetrical effects, thus $\phi^{\alpha}_{DF}(r_1) = \phi^{\alpha}_{DF}(r_2) = 0.1608$.

Dominance states that if one edge always makes a larger contribution than another, then this should be reflected in the magnitude of the RAE.

Proposition 5 (Dominance). Let $\alpha \in A$ and $r_1, r_2 \in R$ with $r_1 \neq r_2$. If $\exists \mathcal{S} \subseteq \mathcal{R} \setminus \{r_1, r_2\}$ such that $\sigma_{\mathcal{S}\cup\{i\}}(\alpha) > \sigma_{\mathcal{S}\cup\{j\}}(\alpha)$ and $\forall \mathcal{S}' \subseteq \mathcal{R} \setminus \{r_1, r_2\}$ ($\mathcal{S}' \neq \mathcal{S}$) such that $\sigma_{\mathcal{S}'\cup\{i\}}(\alpha) \geq \sigma_{\mathcal{S}'\cup\{j\}}(\alpha)$, then $\phi^{\alpha}_{\sigma}(r_1) > \phi^{\alpha}_{\sigma}(r_2)$.

As an illustration, in Example 2, let $\tau(\beta) = 1.0$ while all other base scores remain 0.5. For $\mathcal{S}' = \{r_3, r_4, r_5\}$ and $\forall \mathcal{S}' \subseteq \{r_3, r_4, r_5\}$, we have $\phi^{\alpha}_{DF}(r_1) > \phi^{\alpha}_{DF}(r_2)$ and $\phi^{\alpha}_{DF}(r_1) \geq \phi^{\alpha}_{DF}(r_2)$, thus $\phi^{\alpha}_{DF}(r_1) = 0.3375 > \phi^{\alpha}_{DF}(r_2) = 0.1292$.

5.2 Argumentative Properties

We now study some argumentative properties that we deem interesting in our setting. When assessing these properties, we distinguish three edge types in QBAFs based on the form and number of paths to the topic argument.

Definition 7 (Edge Types). Let $\alpha, \beta, \gamma \in A$, $\alpha \neq \beta$. Then
1. $(\beta, \gamma)$ is a direct edge w.r.t. $\alpha$ if $(\beta, \gamma) \in R$ and there is only one path from $\gamma$ to $\alpha$ in $Q$ (and $\gamma = \alpha$);
2. $(\beta, \gamma)$ is an indirect edge w.r.t. $\alpha$ if there is only one path from $\gamma$ to $\alpha$ in $Q$ (and $\gamma \neq \alpha$);
3. $(\beta, \gamma)$ is a multifold edge w.r.t. $\alpha$ if there is more than one path from $\gamma$ to $\alpha$ in $Q$ (and $\gamma \neq \alpha$).

Example 3 (Cont). Given the QBAF in Figure 2, $r_1$ and $r_2$ are direct edges w.r.t. $\alpha$ as they bring direct support to $\alpha$; $r_3$ and $r_4$ are indirect edges w.r.t. $\alpha$ as they are on single paths to $\alpha$ (while not bringing support or attack to it); $r_5$ is a multifold edge w.r.t. $\alpha$ because it starts two different paths $(r_5, r_3, r_1$ and $r_5, r_4, r_2)$ to $\alpha$.

The first argumentative property is Sign Correctness, demanding that the sign of an edge reflects its polarity.

Property 1 (Sign Correctness). Let $\alpha \in A$ and $r \in R$.
1. If $r \in R^-$, then $\phi^{\alpha}_{\sigma}(r) \leq 0$;
2. If $r \in R^+$, then $\phi^{\alpha}_{\sigma}(r) \geq 0$.

Naturally, sign correctness cannot be satisfied if the gradual semantics does not behave in the intended way, but it does so if it satisfies monotonicity, for direct edges. Note that we are defining this and later properties w.r.t. specific arguments and edges, considering their satisfaction w.r.t. classes of edges, e.g. direct edges as in the next result.

Proposition 6. Let $r$ be a direct edge w.r.t. $\alpha$. $\phi^{\alpha}_{\sigma}(r)$ satisfies sign correctness if $\sigma$ satisfies monotonicity.

For indirect edges, we need to make several case differentiations, as the meaning of edges can be inverted along paths (e.g. an attacker of an attacker, actually serves as a supporter).

Proposition 7. Let $r$ be an indirect edge w.r.t. $\alpha$. Suppose the path sequence from $r$ to $\alpha$ is $r, r_1, \cdots, r_n (n \geq 1)$. Let $\lambda = \{r_1, \cdots, r_n\} \cup R^-$. Then the following statements hold if $\sigma$ satisfies monotonicity.
1. If $r \in R^-$ and $\lambda$ is odd, then $\phi^{\alpha}_{\sigma}(r) \geq 0$;
2. If $r \in R^-$ and $\lambda$ is even, then $\phi^{\alpha}_{\sigma}(r) \leq 0$;
3. If $r \in R^+$ and $\lambda$ is odd, then $\phi^{\alpha}_{\sigma}(r) \leq 0$;
4. If $r \in R^+$ and $\lambda$ is even, then $\phi^{\alpha}_{\sigma}(r) \geq 0$.

Example 4 (Cont). Consider the QBAF in Figure 2 and the RAЕs in Example 2. $r_1 \in R^+$ is a direct edge w.r.t. $\alpha$, hence $\phi^{\alpha}_{DF}(r_1) \geq 0$; while $r_3 \in R^-$ is an indirect edge w.r.t. $\alpha$, and $\lambda$ is 0 (even), hence $\phi^{\alpha}_{DF}(r_3) \leq 0$.

Essentially, these results show that RAEs correctly explain the behavior of direct and indirect edges under monotonicity. For multifold edges, however, monotonicity may not help.

Proposition 8. Let $r$ be a multifold edge w.r.t. $\alpha$. $\phi^{\alpha}_{\sigma}(r)$ may violate sign correctness even if $\sigma$ satisfies monotonicity.

Counterfactuality is a natural property which states that the strength of a topic argument will not be increased (decreased) if an edge with positive (negative) contribution is removed.

Property 2 (Counterfactuality). Let $\alpha \in A$ and $r \in R$.
1. If $\phi^{\alpha}_{\sigma}(r) < 0$, then $\sigma(\alpha) \leq \sigma_{\mathcal{R}\setminus\{r\}}(\alpha)$;
2. If $\phi^{\alpha}_{\sigma}(r) > 0$, then $\sigma(\alpha) \geq \sigma_{\mathcal{R}\setminus\{r\}}(\alpha)$.

Proposition 9. Let $r$ be a direct or indirect edge w.r.t. $\alpha$. $\phi^{\alpha}_{\sigma}(r)$ satisfies counterfactuality if $\sigma$ satisfies monotonicity.

Example 5 (Cont). Consider the QBAF in Figure 2 and the RAЕs in Example 2. $r_1$ is a direct edge w.r.t. $\alpha$ and $\phi^{\alpha}_{DF}(r_1) > 0$. If $r_1$ is removed, then $\sigma^{DF}(\alpha) = 0.8046875 > \sigma_{\mathcal{R}\setminus\{r_1\}}(\alpha) = 0.6875$. $r_3$ is an indirect edge w.r.t. $\alpha$ and $\phi^{\alpha}_{DF}(r_3) < 0$. If $r_3$ is removed, then $\sigma^{DF}(\alpha) = 0.8046875 < \sigma_{\mathcal{R}\setminus\{r_3\}}(\alpha) = 0.84375$. 3625
Proposition 10. Let $r$ be a multifold edge w.r.t. $\alpha$. $\phi^\sigma_\alpha(r)$ may violate counterfactuality even if $\sigma$ satisfies monotonicity.

From a debugging angle, it is worth exploring how the RAE can be adjusted by a user. We find that the RAE of an edge $(\beta, \gamma)$ is closely related to the base score of its source argument $\beta$, in the sense of the properties of Qualitative Invariability and Quantitative Variability defined below.

Qualitative Invariability states that an edge with positive RAE will never make a negative contribution to the topic argument even if the base score of its source argument changes.

Property 3 (Qualitative Invariability). Let $\alpha, \beta \in \mathcal{A}$ and $r \in \mathcal{R}(\beta)$. Let $\delta_\beta$ denote $\phi^\sigma_\alpha(r)$ when setting $\tau(\beta)$ to some $\delta \in \mathcal{I}$.
1. If $\phi^\sigma_\alpha(r) > 0$, then $\forall \delta \in \mathcal{I}, \delta_\beta \geq 0$;
2. If $\phi^\sigma_\alpha(r) > 0$, then $\forall \delta \in \mathcal{I}, \delta_\beta \geq 0$.

Proposition 11. Let $r$ be a direct or indirect edge w.r.t. $\alpha$. $\phi^\sigma_\alpha(r)$ satisfies qualitative invariability if $\sigma$ satisfies monotonicity.

Quantitative variability states that the RAE of an edge will not be increased (decreased) if the base score of its source argument is increased (decreased).

Property 4 (Quantitative Variability). Let $\alpha, \beta \in \mathcal{A}, \alpha \neq \beta,$ and $\mathcal{R}(\beta) = \{r\}$. Let $\delta_\beta$ denote $\phi^\sigma_\alpha(r)$ when setting $\tau(\beta)$ to some $\delta \in \mathcal{I}$.
1. If $\delta < \tau(\beta)$, then $|\delta_\beta| \leq |\phi_\beta(r)|$;
2. If $\delta > \tau(\beta)$, then $|\delta_\beta| \geq |\phi_\beta(r)|$.

We assess the satisfaction of this property when $r$ is a direct, indirect or manifold edge w.r.t. the topic argument.

Proposition 13. Let $r$ be a direct or indirect edge w.r.t. $\alpha$. $\phi^\sigma_\alpha(r)$ satisfies quantitative invariability if $\sigma$ satisfies monotonicity.

Example 7 (Cont). Consider the QBAF in Figure 2 and the RAEs in Example 2. Since $r_1$ is a direct edge w.r.t. $\alpha$ and $\phi^\sigma_{\alpha,DF}(r_1) > 0$, then if $\tau(\beta)$ is increased to some $\delta > \tau(\beta)$, the new RAE $\delta_\beta$ will not decrease by Proposition 13.

Proposition 14. Let $r$ be a multifold edge w.r.t. $\alpha$. $\phi^\sigma_\alpha(r)$ may violate quantitative variability even if $\sigma$ satisfies monotonicity.

Let us note that many properties may be violated for the multifold case even if the underlying gradual semantics is monotonic. This is because monotonicity only guarantees the direct effects of attacks and supports on an argument. Since a single edge can be involved in a large number of paths to the topic argument, one cannot make a reasonable demand about its effect without a long list of case differentiations. We therefore focus on the special cases where there is only a single path from the argument under investigation to the topic argument (i.e. the cases with direct and indirect edges).

Algorithm 1 An Approximation Algorithm for RAEs

Input: A QBAF $Q = (A, R^-, R^+, \tau)$; a gradual semantics $\sigma$; a topic argument $\alpha$; the sample size $N$.
Output: Approximate RAEs attribution_dict.

1: $\text{attribute}_\text{dict} \leftarrow \{\}$ \% empty dictionary
2: for $r$ in $R$ do
3: \hspace{1cm} $\text{sum} \leftarrow 0$
4: while $N > 0$ do
5: \hspace{1cm} $S \leftarrow \text{random_sample}(R \setminus \{r\})$
6: \hspace{1cm} $\text{sum} \leftarrow \text{sum} + [\sigma_{S \cup \{r\}}(\alpha) - \sigma_S(\alpha)]$
7: \hspace{1cm} $N \leftarrow N - 1$
8: end while
9: \hspace{1cm} $\text{attribute}_\text{dict}[r] \leftarrow \text{sum}/N$
10: end for
11: return $\text{attribute}_\text{dict}$

6 Approximating RAEs Probabilistically

Here, we look at how we can compute RAEs efficiently.

Computing RAEs involves computing the final strength values of arguments. The runtime for computing these values depends on the graph structure of the QBAF and on the gradual semantics. Let $n = |R|$ and $m = |A|$. The strength values can be computed in linear time $O(n + m)$ for acyclic QBAFs [Potyka, 2019, Proposition 3.1]. For some pathological examples of cyclic QBAFs, the strength computation may not converge resulting in infinite runtime [Mossakowski and Neuhaus, 2018; Potyka, 2019]. However, for randomly generated cyclic QBAFs, the strength values typically converge in subquadratic time [Potyka, 2018, Figure 7]. In fact, if the outdegree of arguments in the QBAF is not too large, the strength values are guaranteed to converge in linear time [Potyka, 2019, Proposition 3.3]. To avoid a large number of case differentiations, we just denote the runtime for computing strength values by $T(m, n)$ in the following.

Concerning RAEs, let us first note that we can compute them exactly in exponential time by inspecting all subsets of edges (excluding cases where determining strength values fails to converge).

Proposition 15 (Computing RAEs exactly). RAEs can be computed in time $O(2^n \cdot T(m, n))$.

For larger QBAFs, we can apply approximation methods. It is folklore in algorithmic game theory that Shapley values can be seen as expected values. Under this view, the probability of a subset $S \subseteq R \setminus \{r\}$ is defined by $P(S) = \frac{|\{r \leftarrow \{r\} \setminus S\}|}{|\mathcal{R}|}$ and the marginal contribution of $S$ is defined by the function $m(S) = \sigma_{S\cup\{r\}}(\alpha) - \sigma_S(\alpha)$. Our RAEs then correspond to the expected value $E_P[m]$ of the marginal contribution function under $P$. This interpretation allows approximating the Shapley values by the approximation Algorithm 1, returning a dictionary attribution_dict used iteratively to accumulate pairs assigning estimate values to edges. By the Law of Large Numbers (e.g., Theorem 5.1 in [Gut and Gut, 2005]), the estimates converge in probability to the true Shapley values. That is, for every $\epsilon > 0$, the probability that the estimates deviate by more than $\epsilon$ from the true value approaches 0 as the number of samples approaches infinity. By the Central Limit Theorem (e.g., Theorem 5.2 in [Gut and Gut, 2005]), the distribution of the samples approaches a normal distribution with mean
equal to the true Shapley values. This means, in particular, that the estimator is unbiased. However, the variance can be quite large and is better evaluated empirically. We thus do not give a precise formula and simply state the following guarantees.

**Proposition 16.** The estimates generated by Algorithm 1 converge in probability to the true Shapley values.

**Proposition 17** (Approximating RAEs). If the number of samples for each edge is \( N \), then approximate RAEs can be generated in time \( O(n \cdot N \cdot T(m, n)) \).

In particular, when the QBAF is acyclic or meets the conditions on the outdegree of arguments in [Potyka, 2019, Proposition 3.3], we can compute RAEs in time \( O(n \cdot N \cdot (n + m)) \).

Proposition 16 guarantees that Algorithm 1 converges to the true RAEs but does not tell us how many iterations we require to reach a good approximation. In order to evaluate the convergence speed empirically, we conducted experiments with randomly generated QBAFs of increasing size. Figure 4 shows, for cyclic QBAFs (see arxiv.org/abs/2404.14304 for acyclic QBAFs), how the absolute difference (y-axis) between estimates at every 10-th iteration evolves with an increasing number of samples (x-axis), pointing to convergence within a few hundreds iterations. For each iteration, it approximately took 14ms and 0.9ms for cyclic and acyclic QBAFs, respectively, with 15 arguments and 25 edges. We give hardware specifications and additional experiments for runtime, acyclic QBAFs, and differently-sized QBAFs in arxiv.org/abs/2404.14304.

### 7 Case Studies

Finally, we carry out two case studies, including a large QBAF and a non-tree QBAF, to show some practical use of our RAEs.

#### 7.1 Case Study 1: Fraud Detection

**Background** Automatic fraud detection plays an important role in e-commerce. [Chi et al., 2021] propose to use QBAFs for fraud detection because of their intrinsic interpretability. We take the QBAF from [Chi et al., 2021], shown in Figure 5, where argument 1 (‘It is a fraud case’) is the topic argument, and arguments 2 – 48 represent evidence for or against this case. The specific content of these arguments can be found in arxiv.org/abs/2404.14304 or in the original [Chi et al., 2021].

**Settings** We set the base score for each argument to 0.5 in line with [Chi et al., 2021]. Since we do not consider in this paper edge-weighted argumentation [Dunne et al., 2011], we apply DF-QuAD semantics here instead of O-QuAD [Chi et al., 2021] which is a variant of DF-QuAD with weights on edges. The given case is considered fraud if and only if \( \sigma^{DF}(1) > \tau(1) = 0.5 \). Under DF-QuAD, we have \( \sigma^{DF}(1) \approx 0.2544 < \tau(1) \), which means the case is not considered fraud. Since there are 47 edges in Figure 5, computing RAEs exactly is prohibitively expensive. Thus, we apply the approximate Algorithm 1, setting the sample size \( N \) to 1000. We chose \( N \) experimentally to be large enough to guarantee that the estimates converge.\(^5\)

**Explanations** We apply our RAEs and the contributions derived from them to give quantitative explanations for \( \sigma^{DF}(1) \) (see Figure 6; for more details see arxiv.org/abs/2404.14304).

Figure 6 shows that 18 red edges make a negative contribution while 29 blue edges make a positive contribution to argument 1, and negative contributions overwhelm the positive ones. Among the positive contributions, (2, 1) makes the largest, with \( \text{RAE} = 2.55 \times 10^{-1} \), because it directly supports argument 1. (40, 26) makes the smallest contribution with \( \text{RAE} = 2.83 \times 10^{-5} \) because it is indirect and far away from argument 1. Among the negative contributions, (3, 1) makes the largest (\( \text{RAE} = -4.56 \times 10^{-1} \)) since argument 3 directly attacks argument 1, whereas (43, 32) makes the smallest (\( \text{RAE} = -5.84 \times 10^{-5} \)) as there is only one (odd number) attack from argument 43 to argument 1.

In Figure 6, edges close to argument 1 make a greater contribution than those further away. This is because removing close edges will also remove their predecessors on the path to the topic argument. The RAEs of edges such as (4, 2) and (12, 4) show they also play a role, which is different from the

\(^5\)As the additional experiments for large QBAFs in arxiv.org/abs/2404.14304 show, the estimates converge typically after a few hundreds iterations even with more than a thousand edges, so \( N = 1000 \) here is appropriate.
contribution function in [Amgoud et al., 2017].

Note that direct edges do not always make greater contribution than those further away, especially in multifold scenarios. In such cases, we believe explaining the strength by considering all edges rather than only direct edges is a better choice.

7.2 Case Study 2: Large Language Models (LLMs)

Background and Settings LLMs’ ability to process and generate text can contribute to the development of various AI models [Naveed et al., 2023] and help address the knowledge acquisition bottleneck. The idea is that we can query an LLM with a particular claim and use the answers to build up a QBAF. The QBAF can then be used to visualize (potentially contradictory) arguments that the LLM generated and compute final strength of these arguments, and RAEs can be used to explain the relevance of particular edges for the strength of the claim (seen as the topic argument). To present our RAE approach with QBAFs containing more intricate relationships among arguments than the simple tree-like structure resulting in case study 1, we force a certain structure and generate a non-tree QBAF by ChatGPT(GPT-3.5) [OpenAI, 2022], for the claim ‘It is easy for children to learn a foreign language well’ (topic argument α), prompted to create arguments satisfying the following requirements:

1. Provide one argument β attacking α and two arguments γ and δ supporting α.
2. Let β and γ attack and support δ, respectively.
3. Give confidence for all arguments, ranging from 0 to 1.

We obtained the following arguments and confidence values (which we use as base scores).

β (0.6): Learning a foreign language requires cognitive maturity, which children lack. Hence, it’s difficult for them to excel.
γ (0.9): Studies show that young children possess higher neuroplasticity, making language learning more effective.
δ (0.7): Children immersed in a foreign language environment from an early age have better language acquisition.

We used the QE semantics (σ_QE) to compute the strength of arguments and visualize the QBAF and strengths in Figure 7.

Explanations Figure 8 visualises RAEs and contributions and gives a ranking of the edges based on their contribution. There are two paths from β to α: p_1 = r_5 and p_2 = r_4, r_2. The cumulative contributions of p_1 and p_2 are -0.1078 and 0.0884, respectively, obtained by adding up the RAEs on the path. Thus, p_1 and p_2 make, respectively, a negative and positive contribution to α. Also, p_2 makes a greater contribution when considering absolute values. Although p_2 positively contributes to α, we find r_4 makes a negative contribution on this path, which is not obvious if we only compute argument-based attributions. Indeed, we believe that RAEs are better suited than argument-based attribution explanations in this scenario because they provide a deeper insight into the way arguments affect one another along different reasoning paths.

Property Verification Let us check the satisfaction of some properties introduced previously, under σ_QE used in this case study. First, the sum of all RAEs (0.10) corresponds to the deviation from τ(α) = 0.80 to σ_QE(α) = 0.90, which satisfies efficiency. β directly attacks α so r_5 has a negative RAE while γ directly supports α thus r_4 has a positive RAE by sign correctness. According to counterfactuality, if r_1 is removed, then σ_QE(α) will decrease (to 0.80). If the τ(γ) is increased from 0.9 to 0.95, then φ^{τ}_Q(γ) and φ^{α}_Q(γ) are still positive by qualitative invariance, and φ^{α}_Q(γ) increases from 0.1134 to 0.1182 and φ^{α}_Q(γ) increases from 0.008389 to 0.008438 by quantitatively variability.

8 Conclusion

We introduced RAEs to quantitatively explain the role of attack and support relations under gradual semantics for QBAFs, resulting in more fine-grained insights into the contribution of arguments, along different reasoning paths, than argument-based attribution explanations. We proposed several properties for RAEs, including some adapted from properties of Shapley values and some defined ex-novo. The satisfaction and violation of these properties theoretically shows that our RAEs are reasonable and faithful explanations, which is crucial to explanation methods. We also proposed an efficient probabilistic algorithm to approximate RAEs, proved theoretical convergence guarantees and demonstrated experimentally that it converges quickly. Finally, we carried out two case studies to evaluate and show the practical use of our RAEs.

Our work paves the way to many future directions. First, it would be interesting to explore joint Shapley values [Zhang et al., 2021] for sets of attacks and supports and to investigate interactions among edges. Second, it would be worth exploring formal relationships between RAEs and argument-based attribution explanations. Third, it would be interesting to generalize our RAEs to edge-weighted QBAFs [Amgoud et al., 2017]. Lastly, it would be important to carry out user studies as explanations should be easily understood and accepted by human users [Chen et al., 2022].
Acknowledgments

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.

References


