

# Applying Attribution Explanations in Truth-Discovery Quantitative Bipolar Argumentation Frameworks

Xiang Yin<sup>1</sup>, Nico Potyka<sup>2</sup> and Francesca Toni<sup>1</sup>

<sup>1</sup>Imperial College London, UK

<sup>2</sup>Cardiff University, UK

## Abstract

Explaining the strength of arguments under gradual semantics is receiving increasing attention. For example, various studies in the literature offer explanations by computing the attribution scores of arguments or edges in Quantitative Bipolar Argumentation Frameworks (QBAFs). These explanations, known as *Argument Attribution Explanations (AAEs)* and *Relation Attribution Explanations (RAEs)*, commonly employ *removal-based* and *Shapley-based* techniques for computing the attribution scores. While AAEs and RAEs have proven useful in several applications with acyclic QBAFs, they remain largely unexplored for cyclic QBAFs. Furthermore, existing applications tend to focus solely on either AAEs or RAEs, but do not compare them directly. In this paper, we apply both AAEs and RAEs, to Truth Discovery QBAFs (TD-QBAFs), which assess the trustworthiness of sources (e.g., websites) and their claims (e.g., the severity of a virus), and feature complex cycles. We find that both AAEs and RAEs can provide interesting explanations and can give non-trivial and surprising insights.

## Keywords

Explainable AI, Quantitative Argumentation, Truth Discovery Application

## 1. Introduction

Abstract argumentation Frameworks (AFs) [1] are promising tools in the Explainable AI (XAI) field [2] due to their transparency and interpretability, as well as their ability to support reasoning about conflicting information [3, 4, 5]. Quantitative Bipolar AFs (QBAFs) [6] are an extension of traditional AFs, which consider the *(dialectical) strength* of arguments and the *support* relation between arguments. In QBAFs, each argument has a *base score*, and its dialectical strength is computed by *gradual semantics* based on its base score and the strength of its attackers and supporters [7]. QBAFs can be deployed to support several applications like product recommendation [8], review aggregation [9] or stance aggregation [10].

Another interesting application that has been considered recently are truth discovery networks [11, 12, 13]. Figure 1 shows an example of a *Truth-Discovery QBAFs (TD-QBAF)* to evaluate the trustworthiness of sources and the reliability of claims made about an exhibition. We have 11 sources and 6 claims, each represented as an abstract argument. The nodes on the left represent the 11 source arguments ( $s_0$  to  $s_{10}$ ), while the ones on the right represents the 6 claim arguments. The claim arguments are categorized into three types – year, place, and theme

---

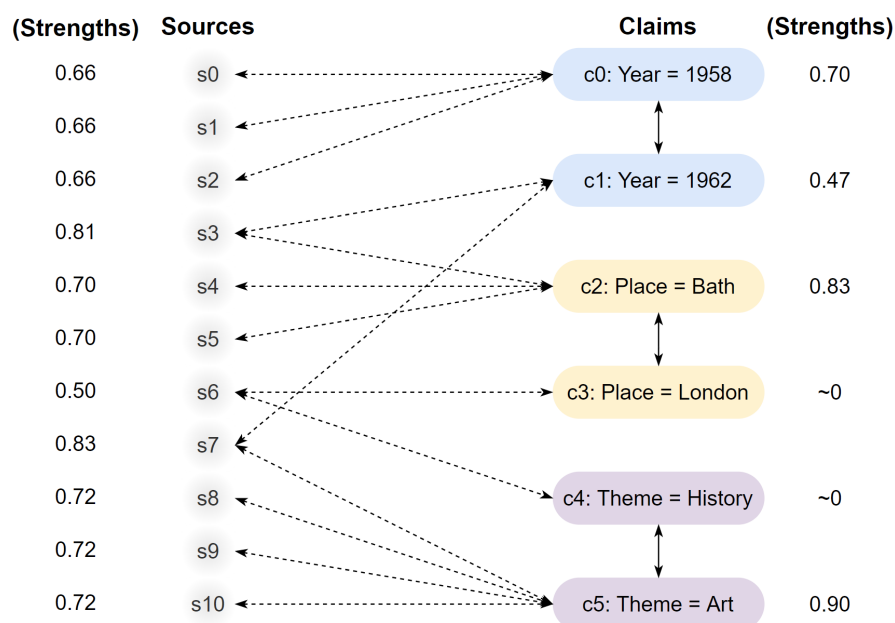
ArgXAI-24: 2nd International Workshop on Argumentation for eXplainable AI

✉ x.yin20@imperial.ac.uk (X. Yin); potykan@cardiff.ac.uk (N. Potyka); ft@imperial.ac.uk (F. Toni)

🆔 000-0002-6096-9943 (X. Yin); 0000-0003-1749-5233 (N. Potyka); 0000-0001-8194-1459 (F. Toni)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** Example of a TD-QBAF. (Nodes are arguments, where the  $s_i$  and  $c_i$  are identifiers for the *source* and *claim* arguments, respectively (for ease of reference). Solid and dashed edges indicate *attack* and *support*, respectively.)

of the exhibition — each distinguished by different colors. For pairs of contradictory claims, where different values are asserted for the same object, a bi-directional attack relationship is introduced between the claims. For each report (one for each pair of source and claim), a bi-directional support relationship is established between the source and the claim. Following [13], we use a base score of 0.5 for source argument (we are initially indifferent about the trustworthiness of a source), and a base score of 0 for claims (we do not believe claims without evidence). We compute the dialectical strength of arguments using the *Quadratic Energy (QE)* gradual semantics [14], and the final strengths of arguments are displayed on their side in Figure 1. While the strength values seem plausible, it can be challenging to understand why certain claims and sources receive higher or lower trust scores.

To address this problem, attribution explanations (AEs) have been proposed. Specifically, given an argument of interest (*topic argument*) in a QBAF, AEs can explain the impact of arguments on the topic argument. AEs can be broadly categorized into *Argument Attribution Explanations (AAEs)* (e.g., [15, 16, 17]) and *Relation Attribution Explanations (RAEs)* (e.g., [18, 19]). AAEs 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. Similarly, RAEs assign the attribution scores to edges to measure their contribution. *Removal-based* and *Shapley-based* techniques are commonly used for computing the attribution scores.

However, most existing studies focus on explaining acyclic QBAFs rather than cyclic ones, leaving a gap in understanding the complexities of the latter. In addition, current research typically examines only one type of attribution — either AAEs or RAEs — without providing a

comprehensive comparison of both methods. In this paper, we aim to address these gaps by investigating the applicability of removal and Shapley-based AAEs and RAEs in the context of cyclic TD-QBAFs. Furthermore, we offer a comprehensive comparison between them to better understand the applicability of these AEs.

## 2. Preliminaries

### 2.1. QBAFs and the QE Gradual Semantics

We briefly recall the definition of QBAFs and the QE gradual semantics [14].

**Definition 1 (QBAF).** A Quantitative Bipolar Argumentation Framework (QBAF) is a quadruple  $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$  consisting of a finite set of arguments  $\mathcal{A}$ , binary relations of attack  $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$  and support  $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$  ( $\mathcal{R}^- \cap \mathcal{R}^+ = \emptyset$ ) and a base score function  $\tau : \mathcal{A} \rightarrow [0, 1]$ .

The base score function in QBAFs assigns an apriori belief to arguments. QBAFs can be represented graphically (as in Figure 1) using nodes to represent arguments and edges to show the relations between them. Then QBAFs are said to be (a)cyclic if the graphs representing them are (a)cyclic.

In this paper, we use the QE gradual semantics [14] to evaluate the *strength* of arguments in QBAFs. Like most QBAF semantics, it computes strength values iteratively by initializing the strength value of each argument with its base score and repeatedly applying an update function. Let us represent the strength of arguments in the  $i$ -th iteration by a function

$$\sigma^i : \mathcal{A} \rightarrow [0, 1],$$

where  $\sigma^0(\alpha) = \tau(\alpha)$  for all  $\alpha \in \mathcal{A}$ . In order to compute  $\sigma^{i+1}$  from  $\sigma^i$ , the update function first computes the *energy*  $E_\alpha^i$  of attackers and supporters of each argument  $\alpha$  defined by

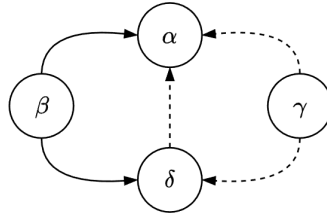
$$E_\alpha^i = \sum_{\{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^+\}} \sigma^i(\beta) - \sum_{\{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^-\}} \sigma^i(\beta).$$

It then computes the strength in the next iteration via

$$\sigma^{i+1}(\alpha) = \begin{cases} \tau(\alpha) - \tau(\alpha) \cdot \frac{(E_\alpha^i)^2}{1+(E_\alpha^i)^2} & \text{if } E_\alpha^i \leq 0; \\ \tau(\alpha) + (1 - \tau(\alpha)) \cdot \frac{(E_\alpha^i)^2}{1+(E_\alpha^i)^2} & \text{if } E_\alpha^i > 0. \end{cases}$$

The final dialectical strength of each argument  $\alpha$  is then defined as the limit  $\lim_{t \rightarrow \infty} \sigma^t(\alpha)$ . In cyclic graphs, the strength values may start oscillating and the limit may not exist [20]. In all known cases, the problem can be solved by *continuizing* the semantics [14, 13]. However, we do not have space to discuss these issues in more detail here and will just restrict to examples where the strength values converge.

To better understand the QE gradual semantics, let us look at an example.



**Figure 2:** Example of a QBAF structure for computing the QE gradual semantics.

**Example 1.** Consider the QBAF in Figure 2, where the base scores are given as  $\tau(\alpha) = 0.8$ ,  $\tau(\beta) = 0.6$ ,  $\tau(\gamma) = 0.9$ , and  $\tau(\delta) = 0.7$ . Since  $\beta$  and  $\gamma$  have no parents, we have  $E_\beta^i = E_\gamma^i = 0$  for all  $i$  and thus  $\sigma(\beta) = \tau(\beta) = 0.6$  and  $\sigma(\gamma) = \tau(\gamma) = 0.9$ . For  $\delta$ , we have  $E_\delta^i = \sigma^i(\gamma) - \sigma^i(\beta) = 0.3$  for all  $i$ , hence  $\sigma(\delta) = \tau(\delta) + (1 - \tau(\delta)) \cdot 0.3^2 / (1 + 0.3^2) = 0.72$ . For  $\alpha$ , we have  $E_\alpha^i = \sigma^i(\gamma) + \sigma^i(\delta) - \sigma^i(\beta) = 1.02$  for all  $i \geq 1$ . Hence,  $\sigma(\alpha) = \tau(\alpha) + (1 - \tau(\alpha)) \cdot 1.02^2 / (1 + 1.02^2) = 0.90$ .

In the remainder, unless specified otherwise, we assume as given a generic QBAF  $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$  and we let  $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+$ . We will often need to restrict QBAFs to a subset of the arguments or edges, or change the base score function, as follows.

**Notation 1.** For  $\mathcal{U} \subseteq \mathcal{A}$ , let  $\mathcal{Q}^{\upharpoonright \mathcal{U}} = \langle \mathcal{A} \cap \mathcal{U}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ . Then, for any  $\alpha \in \mathcal{A}$ , we let  $\sigma_{\mathcal{U}}(\alpha)$  denote the strength of  $\alpha$  in  $\mathcal{Q}^{\upharpoonright \mathcal{U}}$ .

**Notation 2.** For  $\mathcal{S} \subseteq \mathcal{R}$ , let  $\mathcal{Q}^{\upharpoonright \mathcal{S}} = \langle \mathcal{A}, \mathcal{R}^- \cap \mathcal{S}, \mathcal{R}^+ \cap \mathcal{S}, \tau \rangle$ . Then, for any  $\alpha \in \mathcal{A}$ , we let  $\sigma_{\mathcal{S}}(\alpha)$  denote the strength of  $\alpha$  in  $\mathcal{Q}^{\upharpoonright \mathcal{S}}$ .

**Notation 3.** For  $\tau' : \mathcal{A} \rightarrow [0, 1]$  a base score function, let  $\mathcal{Q}^{\upharpoonright \tau'} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau' \rangle$ . Then, for any  $\alpha \in \mathcal{A}$ , we let  $\sigma_{\tau'}(\alpha)$  denote the strength of  $\alpha$  in  $\mathcal{Q}^{\upharpoonright \tau'}$ .

## 2.2. Truth Discovery QBAFs (TD-QBAFs)

TD-QBAFs allow reasoning about truth discovery problems using quantitative argumentation. Truth discovery problems can be described concisely as *truth discovery networks (TDNs)* [11]. Formally, a TDN is a tuple  $N = (\mathcal{S}, \mathcal{O}, \mathcal{D}, \mathcal{P})$  consisting of a finite set of *sources*  $\mathcal{S}$ , a finite set of *objects*  $\mathcal{O}$ , a set  $\mathcal{D} = \{D_o\}_{o \in \mathcal{O}}$  of *domains* of the objects, and a set of *reports*  $\mathcal{P} \subseteq \mathcal{S} \times \mathcal{O} \times V$ , where  $V = \bigcup_{o \in \mathcal{O}} D_o$ , and for all  $(s, o, v) \in \mathcal{P}$ , we have  $v \in D_o$ , and there is no  $(s, o, v') \in \mathcal{P}$  with  $v \neq v'$ . Given a TDN  $N$ , we are interested in a truth discovery operator that assigns a trust score to each source and each claim [11].

Singleton suggested to reason about TDNs using bipolar argumentation frameworks, where we have bi-directional support edges between sources and their claims (trustworthy sources make claims more believable, and, conversely, believable claims make sources more trustworthy) and contradictory claims attack each other [12]. TD-QBAFs implement this idea with QBAFs, where sources have a base score of 0.5 (we are initially indifferent about the trustworthiness of sources) and claims have a base score of 0 (we do not believe anything without evidence).

**Definition 2 (TD-QBAF induced from a TDN).** *The TD-QBAF induced from the TDN  $N = (\mathcal{S}, \mathcal{O}, \mathcal{D}, \mathcal{P})$  is defined as  $Q = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ , where  $\mathcal{A} = \mathcal{S} \cup \{(o, v) \mid \exists s \in \mathcal{S} : (s, o, v) \in \mathcal{P}\}$ ,  $\mathcal{R}^- = \{(c, c') \in \mathcal{A}^2 \cap \mathcal{C}^2 \mid \text{obj}(c) = \text{obj}(c'), \text{val}(c) \neq \text{val}(c')\}$ ,  $\mathcal{R}^+ = \{(s, (o, v)), ((o, v), s) \mid (s, o, v) \in \mathcal{P}\}$ .  $\tau(s) = 0.5$  for all  $s \in \mathcal{S}$  and  $\tau(c) = 0$  for all  $c \in \mathcal{C}$ .*

Every QBAF semantics gives rise to a truth discovery operator that is defined by associating each source and claim with its final strength under the semantics. The semantical properties of QBAF semantics like balance and monotonicity directly translate to meaningful guarantees for the derived trust scores.

### 2.3. Argument Attribution Explanations

In order to explain trust scores in TD-QBAFs, we recall the removal-based and Shapley-based AAEs. AAEs aim at evaluating the impact of an argument on a given topic argument. The removal-based AAEs proposed by [15] measure how the strength of the topic argument changes if an argument is removed.

**Definition 3 (Removal-based AAEs).** *Let  $\alpha, \beta \in \mathcal{A}$ . The removal-based AAE from  $\beta$  to  $\alpha$  under  $\sigma$  is:*

$$\varphi_\sigma^\alpha(\beta) = \sigma(\alpha) - \sigma_{\mathcal{A} \setminus \{\beta\}}(\alpha).$$

The Shapley-based AAEs [16, 21] use the Shapley value from coalitional game theory [22] to assign attributions. Each argument in a QBAF is seen as a *player* that can contribute to the strength of the topic argument. Intuitively, Shapley-based AAEs look at all possible ways how the argument could be added to the QBAF and average its impact on the topic argument.

**Definition 4 (Shapley-based AAEs).** *Let  $\alpha, \beta \in \mathcal{A}$ . The Shapley-based AAE from  $\beta$  to  $\alpha$  under  $\sigma$  is:*

$$\psi_\sigma^\alpha(\beta) = \sum_{\mathcal{U} \subseteq \mathcal{A} \setminus \{\alpha, \beta\}} \frac{(|\mathcal{A} \setminus \{\alpha\}| - |\mathcal{U}| - 1)! |\mathcal{U}|!}{|\mathcal{A} \setminus \{\alpha\}|!} [\sigma_{\mathcal{U} \cup \{\beta\}}(\alpha) - \sigma_{\mathcal{U}}(\alpha)].$$

### 2.4. Relation Attribution Explanations

RAEs are similar to AAEs, but measure the impact of edges rather than the impact of arguments. Analogous to the idea of removal-based AAEs [15], we consider the removal-based RAEs.

**Definition 5 (Removal-based RAEs).** *Let  $\alpha \in \mathcal{A}$  and  $r \in \mathcal{R}$ . The removal-based RAE from  $r$  to  $\alpha$  under  $\sigma$  is:*

$$\lambda_\sigma^\alpha(r) = \sigma(\alpha) - \sigma_{\mathcal{R} \setminus \{r\}}(\alpha).$$

Shapley-based RAEs [18, 19] share the same idea with Shapley-based AAEs, but the attribution objects are changed from arguments to edges.

**Definition 6 (Shapley-based RAEs).** *Let  $\alpha \in \mathcal{A}$  and  $r \in \mathcal{R}$ . The Shapley-based RAE from  $r$  to  $\alpha$  under  $\sigma$  is:*

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

### 3. Explaining TD-QBAFs with AAEs and RAEs

#### 3.1. Settings

To compare the different AEs, we explain the strength of argument  $c5$  in Figure 1. Since there are 17 arguments and 32 edges in Figure 1, computing Shapley-based AAEs and RAEs exactly is prohibitively expensive. We therefore apply the approximation algorithm from [19] that approximates the Shapley values using sampling (we set the sample size to 1000).

We report the removal and Shapley-based AAEs and RAEs in Figure 3 and 4<sup>1</sup>. In addition, to provide intuitive explanations for argument  $c5$ , we visualize the removal and Shapley-based AAEs and RAEs as shown in Figure 3 and 4, where blue/red arguments or edges denote positive/negative AAEs or RAEs. The darkness of the color of arguments and the thickness of the edges denote the magnitude of the their AAEs and RAEs, respectively<sup>2</sup>.

#### 3.2. Results and Analysis for AAEs

Figure 3 shows the results of removal and Shapley-based AAEs.

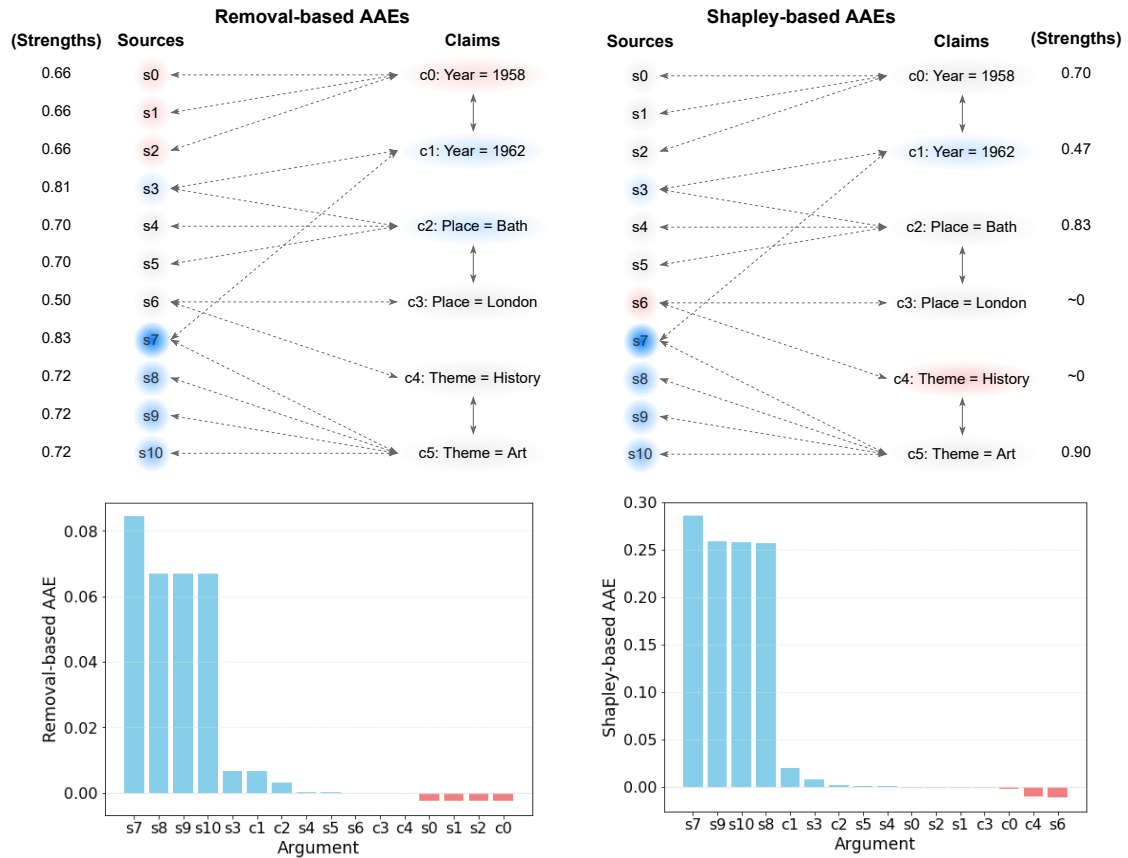
For the **removal-based AAEs**, we observe that  $s7$ ,  $s8$ ,  $s9$ , and  $s10$  have noticeably positive influences on  $c5$ , followed by minor positive influences from  $s3$ ,  $c1$ , and  $c2$ . This is because  $s7$  to  $s10$  are direct supporters for  $c5$ , whereas  $s3$ ,  $c1$ , and  $c2$  indirectly support  $c5$ . Specifically,  $c2$  supports  $s3$ ,  $s3$  supports  $c1$ ,  $c1$  supports  $s7$ , and then  $s7$  supports  $c5$ , meaning  $s3$ ,  $c1$ , and  $c2$  all indirectly support  $c5$ . These indirect influences also explain why the AAEs of  $s3$ ,  $c1$ , and  $c2$  are much smaller than those of  $s7$  to  $s10$ . Besides, since  $s7$  is supported by  $c1$ , its AAE is slightly larger than those of  $s8$  to  $s10$ , which have consistent AAEs due to their symmetrical structure to  $c5$ . In contrast,  $s0$ ,  $s1$ ,  $s2$ , and  $c0$  have minor negative influences on  $c5$  because  $c0$  attacks  $c1$ , an indirect supporter for  $c5$ . Furthermore,  $s0$  to  $s2$  support  $c0$ , and thus they have negative influences on  $c5$  as well. However, their negative influences are not obvious due to the indirect influences. Finally, the remaining arguments have AAEs close to 0, indicating their negligible influences on  $c5$ .

When considering the **Shapley-based AAEs**, the results are similar to those of removal-based AAEs, where  $s7$  to  $s10$  still have significant influences on  $c5$ . Unlike removal-based AAEs, however, we notice that both  $c4$  and  $s6$  have minor negative influences on  $c5$ . This is because  $c4$  directly attacks  $c5$ , while  $s6$  indirectly attacks  $c5$  by supporting  $c4$ , although the QE strength of  $c4$  is very small (close to 0). Also, the negative influences of  $s0$  to  $s2$  and  $c0$  and positive influence of  $c2$  are relatively negligible compared with those of in removal-based AAEs due to their indirect connection to  $c5$ .

In this case study, both removal and Shapley-based AAEs can effectively capture the main influential arguments despite having some tiny differences in those low contributing arguments. This is mainly because of their different mechanisms of computing the AAEs. Another important reason is probably due to the approximation algorithm used for Shapley-based AAEs, leading to different AAEs even with the same sample size for the coalitions. We also noticed that the qualitative influence (the sign) of those Shapley-based AAEs close to 0 is sensitive when

<sup>1</sup>The numerical AAEs and RAEs can be found in the Appendix

<sup>2</sup>The code of all experiments is available at <https://github.com/XiangYin2021/TD-QBAF-AAE-RAE>.



**Figure 3:** Removal and Shapley-based AAEs for the topic argument  $c_5$  of TD-QBAF in Figure 1. (Blue/red/grey nodes denote positive/negative/negligible AAEs, respectively. The darkness of nodes represents the magnitude of their AAE values.)

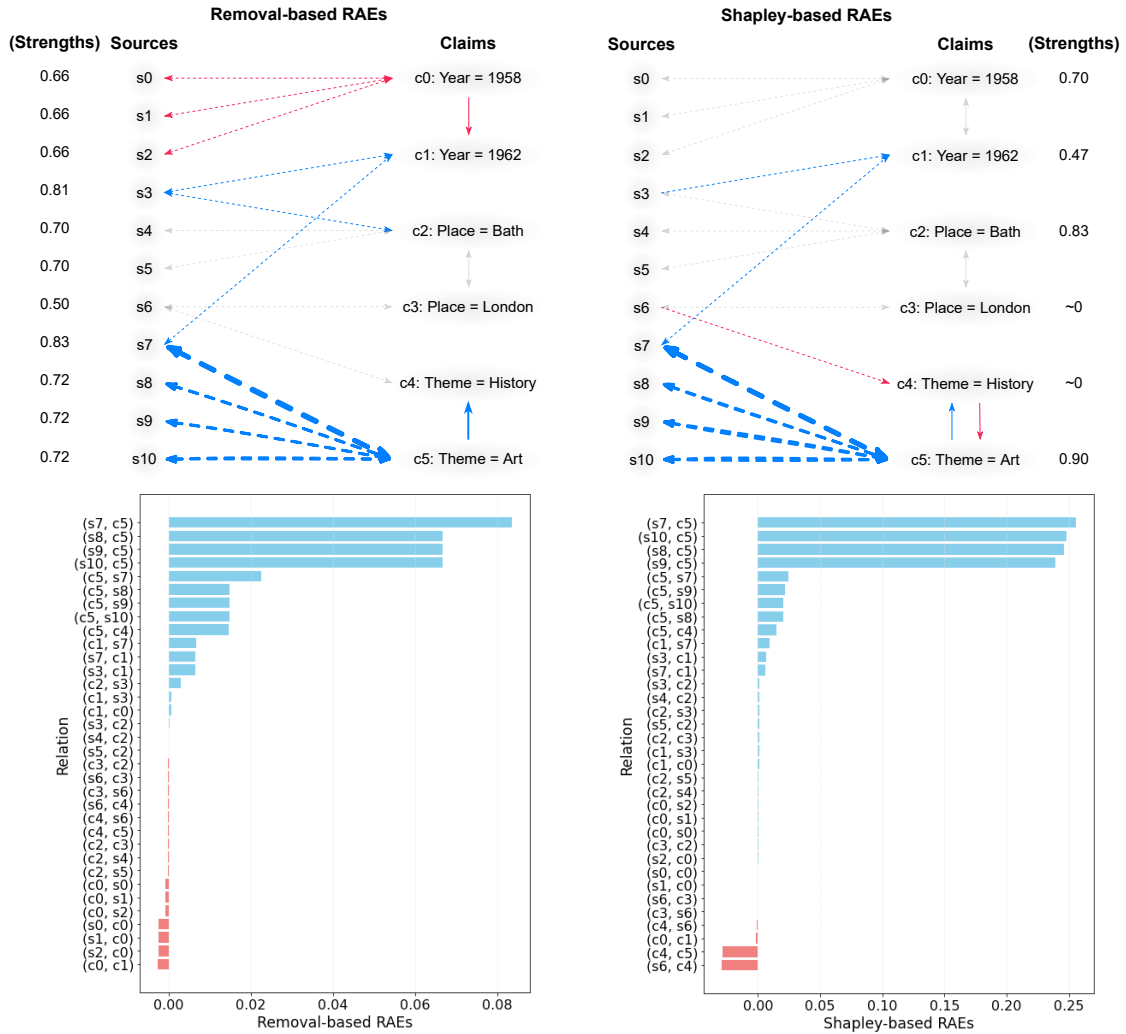
applying the approximation algorithm, thus we do not visualize those close to 0. However, this should not be a concern since their influence is negligible.

### 3.3. Results and Analysis for RAEs

Figure 4 shows the results of removal and Shapley-based RAEs.

Let us first discuss the **removal-based RAEs**. We see that  $(s_7, c_5)$  has the largest positive impact on  $c_5$ . Following closely are  $(s_8, c_5)$ ,  $(s_9, c_5)$ , and  $(s_{10}, c_5)$ , which also have notably positive influences on  $c_5$  because they are direct incoming supports to  $c_5$ . There are also four outgoing supports from  $c_5$ , namely  $(c_5, s_7)$ ,  $(c_5, s_8)$ ,  $(c_5, s_9)$ , and  $(c_5, s_{10})$ , with positive influences but their RAEs are greatly smaller than that of the previous four as they are indirect supports. For instance,  $c_5$  first supports  $s_7$ , and then  $s_7$  supports  $c_5$ , indicating the indirect positive influence of  $(c_5, s_7)$ . Additionally,  $(c_5, c_4)$  also contributes positively to  $c_5$  because  $c_5$  attacks its attacker  $c_4$ , thus the attack from  $c_4$  to  $c_5$  is weakened. We can also observe





**Figure 4:** Removal and Shapley-based RAEs for the topic argument  $c_5$  of TD-QBAF in Figure 1. (Blue/red/grey edges denote positive/negative/negligible RAEs, respectively. The darkness of edges represents the magnitude of their RAE values.)

some marginal influences, such as the positive influences provided by  $(c_1, s_7)$ ,  $(s_3, c_1)$ , and  $(c_2, s_3)$  on  $c_5$ , while the negative influences from  $(c_0, c_1)$ ,  $(s_0, c_0)$ ,  $(s_1, c_0)$ , and  $(s_2, c_0)$ . The remaining edges have RAEs close to 0, showing their negligible influence on  $c_5$ .

When it comes to the **Shapley-based RAEs**, which have similar effects to removal-based RAEs, the four incoming supports to  $c_5$  are still the major contributors, and the four outgoing supports from  $c_5$  have minor RAEs. Different from removal-based RAEs, Shapley-based RAEs capture some different negligible influences, such as the negative influence by  $(c_4, c_5)$  and  $(s_6, c_4)$ . However, Shapley-based RAEs also disregard some tiny influences, like  $(s_0, c_0)$ ,  $(s_1, c_0)$ , and  $(s_2, c_0)$ , which are shown by removal-based RAEs.



In this case study, both removal and Shapley-based RAEs have a consistent ranking for the main influential edges despite having some tiny differences in those low contributing edges. The reasons are the same as we discussed above.

Let us further compare the results of AAEs and RAEs. In this case study, we observe some connections between AAEs and RAEs. For example, in both AAEs, the top-4 influential arguments are  $s7$  to  $s10$ , while in both RAEs, the outgoing edges from these arguments ( $(s7, c5)$ ,  $(s8, c5)$ ,  $(s9, c5)$ , and  $(s10, c5)$ ) also rank in the top-4. In addition,  $s0$  to  $s4$  and  $c0$  to  $c2$  have minor influences in the removal-based AAEs, while their incoming or outgoing edges also have minor influences in the removal-based RAEs. A similar phenomenon can be found in the Shapley-based AAEs and RAEs. While it is expected that the RAEs for outgoing edges of important arguments are relatively high, the consistency observed across different sets of arguments and edges is noteworthy. Besides, we found that removal or Shapley-based AAE of an argument does not necessarily equate to the sum of RAEs of all its incoming and outgoing edges, which goes against a reasonable expectation. We will leave the investigation of their formal relationships for future work.

## 4. Conclusion

Since most existing applications of AAEs and RAEs focus on acyclic QBAFs, this paper investigated their applicability in cyclic QBAFs. First, we found that AAEs and RAEs can provide intuitive explanations. By displaying the ranking of arguments or edges, it is easy to identify the most influential arguments or edges in the QBAF without delving into the complex (cyclic) structure of the QBAFs, particularly in TD-QBAFs where the number of arguments is typically large and the connections between source arguments and claim arguments are bi-directional. Second, AAEs and RAEs can provide interesting or even surprising explanations. For example, in the case study provided earlier, one might overlook the influence between claim arguments  $c1$  and  $c5$  because they are in different topics (*Year=1962* and *Theme=Art*), but AEs demonstrate that  $c1$  can contribute to  $c5$  through  $s7$ . Third, RAEs provide more fine-grained explanations than AAEs. This is because when computing AAEs, such as removal-based AAEs, removing an argument means removing all the incoming and outgoing edges associated with that argument, whereas RAEs offer a more detailed insight by processing every incoming and outgoing edge individually. One can choose between them depending on the granularity for practical use.

For future work, it would be worthwhile to investigate how different gradual semantics influence AAEs and RAEs, because the property satisfaction of semantics have an influence on the property satisfaction of explanations. Additionally, the formal relationship between AAEs and RAEs requires further exploration. However, we believe AAEs and RAEs can complement each other, providing a deeper and more comprehensive understanding of the internal mechanisms of QBAFs, particularly the interactions between arguments and edges in complex QBAFs.

## 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

- [1] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, *Artificial Intelligence* 77 (1995) 321–358.
- [2] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: *Conference on fairness, accountability, and transparency*, 2019, pp. 279–288.
- [3] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: a survey, in: *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4392–4399.
- [4] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: *AAAI Conference on Artificial Intelligence*, volume 35, 2021, pp. 6463–6470.
- [5] N. Potyka, X. Yin, F. Toni, Explaining random forests using bipolar argumentation and markov networks, in: *AAAI Conference on Artificial Intelligence*, volume 37, 2023, pp. 9453–9460.
- [6] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, *Argument & Computation* 6 (2015) 24–49.
- [7] P. Baroni, A. Rago, F. Toni, From fine-grained properties to broad principles for gradual argumentation: A principled spectrum, *International Journal of Approximate Reasoning* 105 (2019) 252–286.
- [8] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: Fantastic explanations and how to find them, in: *IJCAI 2018*, 2018, pp. 1949–1955.
- [9] O. Cocarascu, A. Rago, F. Toni, Extracting dialogical explanations for review aggregations with argumentative dialogical agents, in: *AAMAS 2019*, 2019, pp. 1261–1269.
- [10] N. Kotonya, F. Toni, Gradual argumentation evaluation for stance aggregation in automated fake news detection, in: *Workshop on Argument Mining*, 2019, pp. 156–166.
- [11] J. Singleton, R. Booth, Towards an axiomatic approach to truth discovery, *Autonomous Agents and Multi-Agent Systems* 36 (2022) 1–49. URL: <https://doi.org/10.1007/s10458-022-09569-3>.
- [12] J. Singleton, On the link between truth discovery and bipolar abstract argumentation, *Online Handbook of Argumentation for AI* (2020) 43.
- [13] N. Potyka, R. Booth, An empirical study of the behaviour of quantitative bipolar argumentation frameworks for truth discovery, in: *Computational Models of Argument - Proceedings of COMMA*, 2024, p. To appear.
- [14] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 2018, pp. 148–157.
- [15] J. Delobelle, S. Villata, Interpretability of gradual semantics in abstract argumentation, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: European Conference (ECSQARU)*, volume 11726, 2019, pp. 27–38.

- [16] K. Čyras, T. Kampik, Q. Weng, Dispute trees as explanations in quantitative (bipolar) argumentation, in: *ArgXAI 2022, 1st International Workshop on Argumentation for eXplainable AI*, Cardiff, Wales, September 12, 2022, volume 3209, CEUR-WS, 2022.
- [17] X. Yin, N. Potyka, F. Toni, Argument attribution explanations in quantitative bipolar argumentation frameworks, in: *European Conference on Artificial Intelligence (ECAI)*, volume 372, 2023, pp. 2898–2905.
- [18] L. Amgoud, J. Ben-Naim, S. Vesic, Measuring the intensity of attacks in argumentation graphs with shapley value, in: *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 63–69.
- [19] 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, pp. 3622–3630.
- [20] T. Mossakowski, F. Neuhaus, Modular semantics and characteristics for bipolar weighted argumentation graphs, *arXiv preprint arXiv:1807.06685* (2018).
- [21] T. Kampik, N. Potyka, X. Yin, K. Čyras, F. Toni, Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis, *arXiv preprint arXiv:2401.08879* (2024).
- [22] L. S. Shapley, *Notes on the N-person Game*, 1951.

## Additional Results for AAEs and RAEs

**Table 1**

Comparison of removal-based AAEs and Shapley-based AAEs (in descending order) for the argument  $c_5$  of TD-QBAF in Figure 1. Note that they are in different scales.

| Argument | Removal-based AAE | Argument | Shapley-based AAE |
|----------|-------------------|----------|-------------------|
| s7       | 0.084304029       | s7       | 0.285373360       |
| s8       | 0.066738248       | s9       | 0.259206533       |
| s9       | 0.066738248       | s10      | 0.257762474       |
| s10      | 0.066738248       | s8       | 0.256392126       |
| s3       | 0.006673061       | c1       | 0.020544840       |
| c1       | 0.006635552       | s3       | 0.007852405       |
| c2       | 0.002938110       | c2       | 0.001789997       |
| s4       | 0.000076913       | s5       | 0.000810191       |
| s5       | 0.000076913       | s4       | 0.000789093       |
| s6       | -0.000008421      | s0       | -0.000593859      |
| c3       | -0.000008421      | s2       | -0.000917430      |
| c4       | -0.000008423      | s1       | -0.001164825      |
| s0       | -0.002444482      | c3       | -0.001309005      |
| s1       | -0.002444482      | c0       | -0.001711158      |
| s2       | -0.002444482      | c4       | -0.010154039      |
| c0       | -0.002476209      | s6       | -0.010892383      |

**Table 2**

Comparison of removal-based RAEs and Shapley-based RAEs (in descending order) for the argument  $c_5$  of TD-QBAF in Figure 1. Note that they are in different scales.

| Relation  | Removal-based RAE | Relation  | Shapley-based RAE |
|-----------|-------------------|-----------|-------------------|
| (s7, c5)  | 0.083473613       | (s7, c5)  | 0.255513421       |
| (s8, c5)  | 0.066745475       | (s10, c5) | 0.247761961       |
| (s9, c5)  | 0.066745475       | (s8, c5)  | 0.245927825       |
| (s10, c5) | 0.066745475       | (s9, c5)  | 0.238930772       |
| (c5, s7)  | 0.022507211       | (c5, s7)  | 0.024524066       |
| (c5, s8)  | 0.014968725       | (c5, s9)  | 0.022019375       |
| (c5, s9)  | 0.014968725       | (c5, s10) | 0.020724751       |
| (c5, s10) | 0.014968725       | (c5, s8)  | 0.020059231       |
| (c5, c4)  | 0.014703252       | (c5, c4)  | 0.015153831       |
| (c1, s7)  | 0.006793938       | (c1, s7)  | 0.009541657       |
| (s7, c1)  | 0.006577898       | (s3, c1)  | 0.006460594       |
| (s3, c1)  | 0.006576020       | (s7, c1)  | 0.005974419       |
| (c2, s3)  | 0.002946488       | (s3, c2)  | 0.001282426       |
| (c1, s3)  | 0.000805779       | (s4, c2)  | 0.001206074       |
| (c1, c0)  | 0.000695966       | (c2, s3)  | 0.001140980       |
| (s3, c2)  | 0.000212036       | (s5, c2)  | 0.001083422       |
| (s4, c2)  | 0.000085191       | (c2, c3)  | 0.000928671       |
| (s5, c2)  | 0.000085191       | (c1, s3)  | 0.000881815       |
| (c3, c2)  | -0.000007913      | (c1, c0)  | 0.000834337       |
| (s6, c3)  | -0.000007913      | (c2, s5)  | 0.000670102       |
| (c3, s6)  | -0.000007913      | (c2, s4)  | 0.000631869       |
| (s6, c4)  | -0.000007913      | (c0, s2)  | 0.000586827       |
| (c4, s6)  | -0.000007913      | (c0, s1)  | 0.000558687       |
| (c4, c5)  | -0.000007915      | (c0, s0)  | 0.000558250       |
| (c2, c3)  | -0.000115977      | (c3, c2)  | 0.000164534       |
| (c2, s4)  | -0.000120350      | (s2, c0)  | 0.000144884       |
| (c2, s5)  | -0.000120350      | (s0, c0)  | -0.000017152      |
| (c0, s0)  | -0.000641070      | (s1, c0)  | -0.000036753      |
| (c0, s1)  | -0.000641070      | (s6, c3)  | -0.000276491      |
| (c0, s2)  | -0.000641070      | (c3, s6)  | -0.000465298      |
| (s0, c0)  | -0.002436075      | (c4, s6)  | -0.001147533      |
| (s1, c0)  | -0.002436075      | (c0, c1)  | -0.001659705      |
| (s2, c0)  | -0.002436075      | (c4, c5)  | -0.028383722      |
| (c0, c1)  | -0.002598473      | (s6, c4)  | -0.029099303      |