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

Xiang Yin[1] , **Nico Potyka**[2] , **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]School of Computer Science and Informatics, Cardiff University, UK
{xy620, ft}@imperial.ac.uk, potykan@cardiff.ac.uk

1 Introduction

Explainable Artificial Intelligence (XAI) (Xu et al. 2019) 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, Russell, and Wachter 2019) due to their transparency and interpretability, as well as their ability to support reasoning about conflicting information (Čyras et al. 2021; Albini et al. 2020; Potyka 2021; Potyka, Yin, and Toni 2023; Ayoobi, Potyka, and Toni 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, Rago, and Toni 2019). QBAFs can be deployed to support several applications. For example, (Cocarascu, Rago, and Toni 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, 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; Čyras, Kampik, and Weng 2022; Yin, Potyka, and Toni 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

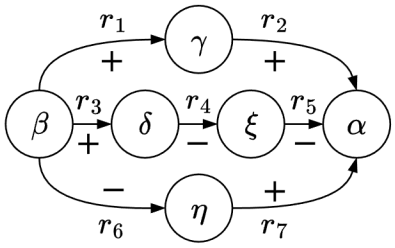

Figure 1: Graphical view of (elements of) a QBAF resulting from aggregating movie reviews (here, nodes are arguments, edges labelled + are supports, edges labelled - are attacks, and the r_i are identifiers for the edges (for ease of reference)).

that the QBAF (partially) depicted therein results from aggregating movie reviews as in (Cocarascu, Rago, and Toni 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 r_1 and r_3 to α, it also has a negative contribution via r_6.

2 Contribution

Motivated by the aforementioned considerations, we make the following contributions:

- We propose a comprehensive theory of RAEs that adapts Shapley values to quantify the contributions of both attack and support relations within QBAFs. This approach addresses the limitations of traditional argument-based attribution methods by providing more detailed insights into the influence of individual relations.
- We propose several desirable properties of RAEs, including some adapted from properties of Shapley values and some defined ex-novo. These properties ensure that RAEs provide reasonable and faithful explanations.
- Recognizing the computational challenges of calculating exact Shapley values in large QBAFs, we introduce a probabilistic algorithm that efficiently approximates

RAEs. This algorithm is validated with theoretical convergence guarantees and empirical evidence of its quick convergence.

- To demonstrate the utility of RAEs, we conduct two detailed case studies: one on fraud detection and another on Large Language Models (LLMs). These case studies illustrate how RAEs can be applied to real-world scenarios, providing actionable insights and enhancing the interpretability of complex decision-making processes.

3 Relevance to KR

Our research is centered on the explainability of formal QBAFs, an area that holds significant relevance for researchers specializing in argumentation theory as well as those involved in XAI. Specifically, it caters to researchers who are dedicated to enhancing the explainability of formal models through the application of XAI techniques. By delving into the intricate mechanisms of QBAFs, our work aims to bridge the gap between formal models and their practical explainability, making it a compelling and valuable study for researchers in these intersecting fields.

4 Significance of Results

This paper introduces the concept of RAEs as a reasonable and faithful method for enhancing the explainability of QBAFs. By adapting Shapley values to the context of argumentation, RAEs provide a detailed, theoretically sound, and computationally feasible approach to attributing argument strength to both attacks and supports within QBAFs. This methodological advancement allows for a nuanced understanding of the role of attacks and supports, offering fine-grained insights into their influences.

The significance of this work is underscored through comprehensive case studies that demonstrate the practical utility of RAEs. These case studies highlight how RAEs can effectively explain the reasoning processes within QBAFs, thereby improving the transparency of QBAFs. This increased transparency is crucial for fostering trust in AI systems, particularly in applications where understanding the underlying decision-making processes is essential. By bridging the gap between theoretical foundations and practical application, this paper contributes a significant advancement to the fields of argumentation theory and XAI, offering researchers and practitioners a reasonable and faithful method for enhancing the interpretability and reliability of 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

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. Deep argumentative explanations. *arXiv:2012.05766*.

Ayoobi, H.; Potyka, N.; and Toni, F. 2023. Sparx: Sparse argumentative explanations for neural networks. In *European Conference on Artificial Intelligence (ECAI)*, volume 372, 149–156.

Baroni, P.; Romano, M.; Toni, F.; Aurisicchio, M.; and Bertanza, G. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6:24–49.

Baroni, P.; Rago, A.; and Toni, F. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* 105:252–286.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 1261–1269.

Čyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: a survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 4392–4399.

Čyras, K.; Kampik, T.; and Weng, Q. 2022. Dispute trees as explanations in quantitative (bipolar) argumentation. In *International Workshop on Argumentation for eXplainable AI*, volume 3209, 1–12.

Delobelle, J., and Villata, S. 2019. Interpretability of gradual semantics in abstract argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: European Conference (ECSQARU)*, volume 11726, 27–38.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77:321–358.

Mittelstadt, B.; Russell, C.; and Wachter, S. 2019. Explaining explanations in ai. In *Conference on fairness, accountability, and transparency*, 279–288.

Potyka, N.; Yin, X.; and Toni, F. 2023. Explaining random forests using bipolar argumentation and markov networks. In *AAAI Conference on Artificial Intelligence*, volume 37, 9453–9460.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *AAAI Conference on Artificial Intelligence*, volume 35, 6463–6470.

Xu, F.; Uszkoreit, H.; Du, Y.; Fan, W.; Zhao, D.; and Zhu, J. 2019. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8*, 563–574. Springer.

Yin, X.; Potyka, N.; and Toni, F. 2023. Argument attribution explanations in quantitative bipolar argumentation frameworks. In *European Conference on Artificial Intelligence (ECAI)*, volume 372, 2898–2905.