Contestability in Edge-Weighted Quantitative Bipolar Argumentation Frameworks

Xiang Yin[1], **Nico Potyka**[2], **Antonio Rago**[3], **Timotheus Kampik**[4] and **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]School of Computer Science and Informatics, Cardiff University, UK
[3]Department of Informatics, King's College London, UK
[4]Department of Computing Science, Umeå University, Sweden
{xy620, ft}@ic.ac.uk, potykan@cardiff.ac.uk, antonio.rago@kcl.ac.uk, timotheus.kampik@umu.se

Abstract

Contestable AI requires that AI-driven decisions align with given subjective user preferences. Various types of argumentation frameworks have been shown to support forms of contestability. In this paper we focus on the little-studied Edge-Weighted Quantitative Bipolar Argumentation Frameworks (EW-QBAFs), where arguments have a base score as in QBAFs but attacks and supports (edges) are weighted. After generalising gradual semantics and properties thereof from QBAFs to EW-QBAFs, we introduce the *contestability problem* for EW-QBAFs, which asks how to modify *edge weights* to achieve a desired *strength* for a specific *topic argument*. To address this problem, we propose *gradient-based relation attribution explanations (G-RAEs)*, which quantify the sensitivity of the topic argument's strength to changes in individual edge weights, thus providing interpretable guidance for weight adjustments towards contestability. Building on G-RAEs, we develop a heuristic algorithm that progressively adjusts the edge weights to attain the desired strength. We evaluate our approach experimentally on synthetic EW-QBAFs that simulate the structural characteristics of personalised recommender systems and multi-layer perceptrons, demonstrating that it can support contestability effectively.

1 Introduction

Contestable AI (Lyons, Velloso, and Miller 2021; Alfrink et al. 2023; Leofante et al. 2024) studies AI systems that allow users to challenge their outputs. This is particularly relevant when these outputs deviate from expected or desirable results. Such deviations may arise from misalignments with subjective user preferences. To enable contestable AI, recent work (Leofante et al. 2024) advocates computational argumentation (see (Atkinson et al. 2017) for an overview) as a promising paradigm, due to its inherent abilities to support conflict resolution, explainability and interactivity (e.g., as in (Cocarascu, Rago, and Toni 2019; Rago, Li, and Toni 2023; Potyka, Yin, and Toni 2023; Chen, Yin, and Toni 2025)), all properties that have been acknowledged as important for contestable AI (Lyons, Velloso, and Miller 2021; Alfrink et al. 2023; Almada 2019).

Among various forms of computational argumentation, Edge-Weighted Quantitative Bipolar Argumentation Frameworks (EW-QBAFs), in the spirit of (Mossakowski and Neuhaus 2018; Potyka 2021; Chi et al. 2021), allow reasoning to be naturally modelled over conflicting and supporting information in a quantitative way. We see EW-QBAFs as consisting of four components: *arguments*, relations (of *attack* and *support*), *base scores* of arguments, and *edge weights* of relations. The *strengths* of arguments can be evaluated by EW-QBAF semantics, similarly as for QBAFs (Mossakowski and Neuhaus 2018), but depending on edge weights as well as base scores and the strengths of attackers and supporters.

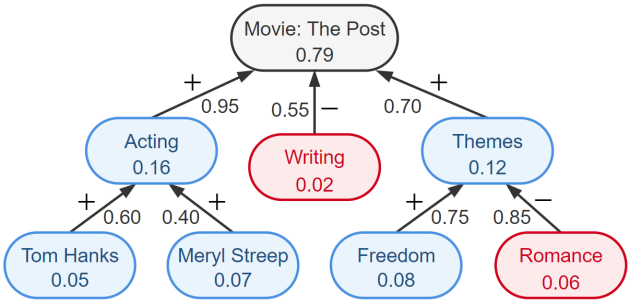

Figure 1: Example of EW-QBAF-based movie PRS (adapted from (Cocarascu, Rago, and Toni 2019)). Blue and red nodes represent supporters and attackers, respectively. Edges labelled + and - indicate support and attack relations, respectively. Values on nodes and edges indicate their base scores and weights, respectively.

EW-QBAFs are well-suited in a number of contexts, e.g. multi-layer perceptrons (MLPs) are EW-QBAFs (Potyka 2021). For illustration, consider Personalised Recommender Systems (PRSs). While black-box PRSs rely on implicit subjective user preference parameters that are hard for users to understand, EW-QBAFs offer an interpretable structure where subjective user preferences are explicitly represented by edge weights [1]. For example, Figure 1 shows an EW-QBAF for movie recommendation. Here, the top-level argument reflects the overall evaluation of a movie, influenced by criteria arguments such as acting and writing, which are further influenced by sub-criteria arguments like specific actors. The base scores reflect the credibility of criteria (e.g.,

[1]"Subjective user preferences" in this paper denotes user-specific degrees of influence encoded by edge weights, rather than preference relations, such as preorders or partial orders.

"the acting quality is high"), possibly derived from aggregated reviews (Cocarascu, Rago, and Toni 2019). In contrast, edge weights represent subjective user preferences – how strongly a user values the influence of each criterion. Although edge weights can be learned from personal behavioural data (e.g., ratings and clicks), such learned subjective user preferences may not always reflect users' actual strength, potentially leading to unsatisfactory recommendations. In such cases, modifying edge weights is a natural way for users to contest the system's reasoning and align recommendations with their subjective preferences.

This motivates the central *contestability problem* in this paper: **given a topic argument α and a desired strength s, how can the edge weights be modified such that the resulting strength of α equals s?** To facilitate its solution, we introduce a notion of *gradient-based relation attribution explanations (G-RAEs)* for EW-QBAFs. Intuitively, G-RAEs quantify the sensitivity of the topic argument's strength to changes in individual edge weights. They thus can serve as interpretable indicators, suggesting how much an edge weight should be increased or decreased. Based on G-RAEs, we design a heuristic algorithm that progressively adjusts the edge weights to reach the desired strength. Experimental results on synthetic EW-QBAFs, designed to simulate MLP and PRS structures, demonstrate that our heuristic algorithm is *effective* in attaining the desired strength, runs in polynomial time, and scales well to large and dense EW-QBAFs.

In summary, our main contributions are as follows:

- We formally define EW-QBAFs and their gradual semantics, generalising those of QBAFs (Section 4).
- We introduce new properties for EW-QBAFs that explain the effects of edge weight changes (Section 5).
- We formally define the contestability problem for EW-QBAFs (Section 6).
- We formally define G-RAEs and study their satisfaction of existing properties (Section 7).
- We develop an algorithm based on G-RAEs to solve the contestability problem (Section 8).
- We empirically show the effectiveness and scalability of our algorithm (Section 9).

All technical proofs and the implementation code are provided in the supplementary material (SM).

2 Related Work

Various proposals for contestable AI have been made (Hirsch et al. 2017; Almada 2019; Alfrink et al. 2023; Russo and Toni 2023). (Leofante et al. 2024) classified contestable AI solutions along three dimensions: the contested entities, the contesting entities, and the contestation method. In our work, the contested entities are the topic arguments' strengths; the contesting entities are humans (e.g., users of PRSs or domain experts using MLPs); and the contesting method is based on the novel notion of G-RAEs.

Argumentation naturally supports contestability. For example, (Shao et al. 2020) proposed structured argumentation-based learning methods that leverage the expressive power of universal function approximators, such as neural networks, while preserving a broad range of tractable inference routines. These methods can provide argumentative explanations for the decision-making process, which can in turn be contested to facilitate model revision, e.g., for a correct output based on incorrect reasoning. Also, (Freedman et al. 2025) employed QBAFs to enhance explainability and contestability in the context of claim verification for large language models (LLMs). In this work, the strength of a topic argument can be contested either through base score modification, or by introducing additional arguments and relations. Our contestability problem is also related to the counterfactual problem (Yin, Potyka, and Toni 2024b; Kampik et al. 2026). This studies how base scores can be changed to change a decision (e.g., revising a credit score in loan approval systems), while our contestability problem focuses on modifying edge weights. The former can be seen as adjusting the apriori belief in an argument, whereas the latter corresponds to adjusting the relevance of an argument for a discussion.

In a wider sense, our work is related to attribution explanations in machine learning (Baehrens et al. 2010; Ribeiro, Singh, and Guestrin 2016; Potyka, Yin, and Toni 2022; Lundberg and Lee 2017; Peacock et al. 2025), as well as Argument Attribution Explanations (AAEs) (Cyras, Kampik, and Weng 2022; Yin, Potyka, and Toni 2023a; Kampik et al. 2024; Anaissy et al. 2024) and Relation Attribution Explanations (RAEs) (Amgoud, Ben-Naim, and Vesic 2017; Yin, Potyka, and Toni 2024c) in argumentation. AAEs and RAEs aim to quantify the influence of individual arguments and relations, respectively, on a topic argument (Yin, Potyka, and Toni 2024a). For example, gradient-based AAEs assess this influence by computing the gradient of the topic argument's strength with respect to the influencing argument's base score, thereby capturing the former's sensitivity to the latter. In contrast, Shapley-based RAEs quantify the influence of a relation (i.e., an attack or support) on the topic argument by leveraging Shapley values from cooperative game theory (Shapley 1953). Our G-RAEs adopt the idea of gradient-based AAEs, but focus on assessing the importance of relations, as RAEs do, but for EW-QBAFs.

3 Background

A *QBAF* is a quadruple $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ where:

- $\mathcal{A}$ is a finite set of *arguments*;
- $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$ is a binary *attack* relation;
- $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ is a binary *support* relation;
- $\mathcal{R}^- \cap \mathcal{R}^+ = \emptyset$;
- $\tau : \mathcal{A} \to [0, 1]$ is a *base score function*.

Let $\mathcal{R} = \mathcal{R}^- \cup \mathcal{R}^+$ throughout this paper. We will use the following notions, originally defined for QBAFs but also applicable to EW-QBAFs.

Definition 1 (Paths). *For any $\alpha, \beta \in \mathcal{A}$, we let $p_{\alpha \mapsto \beta} = \langle (\gamma_0, \gamma_1), \cdots, (\gamma_{n-1}, \gamma_n) \rangle (n \geq 1)$ denote a* path *from α to β, where $\alpha = \gamma_0$, $\beta = \gamma_n$, $\gamma_i \in \mathcal{A} (1 \leq i \leq n)$ and $(\gamma_{i-1}, \gamma_i) \in \mathcal{R}$.*

Let $|p_{\alpha\mapsto\beta}|$ denote the length of path $p_{\alpha\mapsto\beta}$. Let $P_{\alpha\mapsto\beta}$ and $|P_{\alpha\mapsto\beta}|$ denote the set of all paths from α to β, and the number of paths in $P_{\alpha\mapsto\beta}$, respectively.

Definition 2 (Edge Types)**.** *Let $\alpha, \beta, \gamma \in \mathcal{A}$ be pairwise distinct and $(\beta, \alpha), (\beta, \gamma) \in \mathcal{R}$.*

1. *$(\beta, \alpha) \in \mathcal{R}$ is a* direct *edge w.r.t. α;*
2. *$(\beta, \gamma) \in \mathcal{R}$ is an* indirect *edge w.r.t. α if $|P_{\gamma\mapsto\alpha}| = 1$;*
3. *$(\beta, \gamma) \in \mathcal{R}$ is a* multifold *edge w.r.t. α if $|P_{\gamma\mapsto\alpha}| > 1$;*
4. *$(\beta, \gamma) \in \mathcal{R}$ is an* independent *edge w.r.t. α if $|P_{\gamma\mapsto\alpha}| = 0$.*

Gradual semantics for QBAFs usually belong to the family of modular semantics (Mossakowski and Neuhaus 2018). For these semantics, strength values are computed iteratively by an update procedure that initializes the strength values with the base score and updates them repeatedly. They are called modular because the update function can be decomposed into an *aggregation function* that aggregates the strength values of attackers and supporters, and an *influence function* that adapts the base score based on the aggregate. The former has the form $\mathrm{agg}(A, S)$, where A and S are multisets of attack and support values. Examples include

Sum: $\mathrm{agg}_\Sigma(A, S) = \sum_{x\in S} x - \sum_{x\in A} x$,

Product: $\mathrm{agg}_\Pi(A, S) = \prod_{x\in A}(1 - x) - \prod_{x\in S}(1 - x)$.

Product-aggregation is used for the DF-QuAD semantics (Rago et al. 2016), while sum-aggregation is used for the restricted Euler-based (REB) semantics (Amgoud and Ben-Naim 2016; Amgoud and Ben-Naim 2018), and the quadratic energy (QE) semantics (Potyka 2018).

A and S are initialized with the strength values of attackers and supporters.

$$A^\sigma_\alpha = \{\sigma(\beta) \mid \beta \in \mathcal{R}^-(\alpha)\}, \tag{1}$$

$$S^\sigma_\alpha = \{\sigma(\beta) \mid \beta \in \mathcal{R}^+(\alpha)\}. \tag{2}$$

Influence functions have the form $\mathrm{infl}(B, A)$, where B is a base score and A is a real number (the aggregate computed by the aggregation function). Intuitively, the influence function will increase (decrease) the base score if the aggregate is larger (smaller) than 0 while respecting the bounds 0 and 1 of the strength domain. For example, the influence functions of the DF-QuAD (Rago et al. 2016) and QE (Potyka 2018) semantics have the form

$$\mathrm{infl}(B, A) = B - B \cdot h(-A) + (1 - B) \cdot h(A),$$

where $h(x) = \max\{0, x\}$ for DF-QuAD and $h(x) = \frac{\max\{0,x\}^2}{1+\max\{0,x\}^2}$ for QE. Finally, the general algorithm for computing strength values under modular semantics can be described as follows:

Initialization: for all arguments α, let $\sigma^{(0)}(\alpha) = \tau(\alpha)$.

Update: for $i \in \mathbb{N}_0$: for all arguments α, let $\sigma^{(i+1)}(\alpha) = \mathrm{infl}(\tau(\alpha), \mathrm{agg}(A^{\sigma^{(i)}}_\alpha, S^{\sigma^{(i)}}_\alpha))$. Stop if difference between $\sigma^{(i)}$ and $\sigma^{(i+1)}$ is sufficiently small.

For *acyclic* QBAFs (such that there is no path from an argument to itself), the algorithm is equivalent to a simple forward pass with respect to a topological ordering of the arguments, and the strength values can be computed in linear time (Potyka 2019). For *cyclic* QBAFs, sufficient conditions for convergence exist, but they need to make strong assumptions about the indegree of arguments or the base scores (Mossakowski and Neuhaus 2018; Potyka 2019). While the conditions are not necessary, the literature contains various examples of cyclic QBAFs where the algorithm fails to converge (Mossakowski and Neuhaus 2018; Potyka and Booth 2024). In this case, strength values remain undefined. However, in all known cases, convergence problems can be solved by continuizing the semantics (Potyka 2018; Potyka 2019; Potyka and Booth 2024).

4 EW-QBAFs and Gradual Semantics

We consider EW-QBAFs, in the spirit of (Mossakowski and Neuhaus 2018; Potyka 2021; Chi et al. 2021).

Definition 3 (Edge-Weighted QBAF)**.** *An* Edge-Weighted QBAF *is a quintuple $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w\rangle$ where:*

- *$\langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau\rangle$ is a QBAF;*
- *$w : \mathcal{R}^- \cup \mathcal{R}^+ \to [0, 1]$ is an* edge weight function*.*

In the remainder, unless specified otherwise, we assume an EW-QBAF $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w\rangle$. For any $\alpha, \beta \in \mathcal{A}$, we define the sets of incoming attacks, supports, and edges of α as follows: $\mathcal{R}^-(\alpha) = \{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{R}^-\}$, $\mathcal{R}^+(\alpha) = \{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{R}^+\}$, $\mathcal{R}(\alpha) = \mathcal{R}^-(\alpha) \cup \mathcal{R}^+(\alpha)$.

EW-QBAF gradual semantics assign a strength to every argument, in the spirit of gradual semantics for QBAFs, but factoring in also the edge weights.

Definition 4 (Edge-Weighted Gradual Semantics)**.** *An* edge-weighted gradual semantics *is a function $\sigma : \mathcal{A} \to [0, 1] \cup \{\bot\}$. We call $\sigma(\alpha)$ the* strength *of α and say that it is* undefined *only if $\sigma(\alpha) = \bot$.*

Gradual semantics for EW-QBAFs can also be defined in terms of aggregation function and influence functions as for QBAFs, but now the multisets A and S in the aggregation function $\mathrm{agg}(A, S)$ are initialized with edge-weighted strength values instead of strength values (of attackers and supporters). That is, for a strength function σ and an argument α, we consider the multisets

$$A^\sigma_\alpha = \{\sigma(\beta) \cdot w((\beta, \alpha)) \mid \beta \in \mathcal{R}^-(\alpha)\}, \tag{3}$$

$$S^\sigma_\alpha = \{\sigma(\beta) \cdot w((\beta, \alpha)) \mid \beta \in \mathcal{R}^+(\alpha)\}. \tag{4}$$

The aggregation can then be fed into influence functions as per QBAFs and the general algorithm for computing strength values under modular semantics applies as per QBAFs.

Proposition 1 (Equality of Strength)**.** *Let $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau\rangle$ be a QBAF, and $\mathcal{Q}_w = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w\rangle$ be an EW-QBAF such that $w(r) = 1$ for any $r \in \mathcal{R}$. Let σ be a modular gradual semantics for $\mathcal{Q}$ and $\mathcal{Q}_w$ using the same aggregation and influence functions, denoted by $\sigma_\mathcal{Q}$ and $\sigma_{\mathcal{Q}_w}$, respectively. Then, for any $\alpha \in \mathcal{A}$, $\sigma_\mathcal{Q}(\alpha) = \sigma_{\mathcal{Q}_w}(\alpha)$.*

Example 1. *We evaluate the EW-QBAF in Figure 1 using MLP-based semantics (Potyka 2021). Since there are*

five arguments that have no attackers or supporters, their strengths equal their base scores: $\sigma(Tom\ Hanks) = 0.050$, $\sigma(Meryl\ Streep) = 0.070$, $\sigma(Freedom) = 0.080$, $\sigma(Romance) = 0.060$, *and* $\sigma(Writing) = 0.020$. *Then, by applying aggregation and influence function, we have* $\sigma(Acting) = 0.168$, $\sigma(Themes) = 0.121$, *and* $\sigma(Movie) = 0.826$.

5 Properties of EW-QBAFs

Before introducing the contestability problem, we discuss some properties of EW-QBAFs that will be useful in the following. The properties are motivated by properties in the literature, but pertain to the effect of edge weights, whereas existing properties only cover the effect of base scores. We recall some definitions from (Potyka and Booth 2024).

- For a multiset S of real numbers, we let $Core(S) = \{x \in S \mid x \neq 0\}$ denote the sub-multiset that contains only the non-zero elements.
- Let S, T be multisets of real numbers. S *dominates* T if $Core(S) = Core(T) = \emptyset$ or there is a sub-multiset $S' \subseteq Core(S)$ and a bijective function $f : Core(T) \to S'$ such that $x \leq f(x)$ for all $x \in Core(T)$. If, in addition, $|Core(S)| > |Core(T)|$ or there is an $x \in Core(T)$ such that $x < f(x)$, S *strictly dominates* T. We write $S \succeq T$ ($S \succ T$) if S (strictly) dominates T.
- S and T are *balanced* if $Core(S) = Core(T)$. We write $S \cong T$ in this case.
- An aggregation function agg satisfies balance iff
  1. $A \cong S$ implies $\mathrm{agg}(A, S) = 0$, and
  2. $A \cong A'$ and $S \cong S'$ implies $\mathrm{agg}(A, S) = \mathrm{agg}(A', S')$.
- An aggregation function agg satisfies *Neutrality* iff $\mathrm{agg}(A, S) = \mathrm{agg}(Core(A), Core(S))$.
- An influence function infl satisfies *Balance* if $\mathrm{infl}(b, 0){=}b$.
- An aggregation function agg satisfies *Monotonicity* iff
  1. $\mathrm{agg}(A, S) \leq 0$ if $A \succeq S$, and
  2. $\mathrm{agg}(A, S) \geq 0$ if $S \succeq A$, and
  3. $\mathrm{agg}(A, S) \leq \mathrm{agg}(A', S)$ if $A \succeq A'$,
  4. $\mathrm{agg}(A, S) \geq \mathrm{agg}(A, S')$ if $S \succeq S'$.

  If the inequalities can be replaced by strict inequalities for strict domination, agg satisfies *Strict Monotonicity*.
- An influence function infl satisfies *Monotonicity* iff
  1. $\mathrm{infl}(b, a) \leq b$ if $a < 0$, and
  2. $\mathrm{infl}(b, a) \geq b$ if $a > 0$, and
  3. $\mathrm{infl}(b_1, a) \leq \mathrm{infl}(b_2, a)$ if $b_1 < b_2$, and
  4. $\mathrm{infl}(b, a_1) \leq \mathrm{infl}(b, a_2)$ if $a_1 < a_2$.

  If the inequalities can be replaced by strict inequalities, when excluding $b = 0$ for the first and $b = 1$ for the second item, infl satisfies *strict monotonicity*.

Our first property, *edge-neutrality*, is a variant of *neutrality* (Amgoud and Ben-Naim 2018): while the latter states that arguments with strength 0 have no effect, the former demands that an edge with weight 0 has no impact.

Definition 5 (Edge-Neutrality)**.** *A semantics* σ *satisfies* edge-neutrality *iff, for any* $\mathcal{Q}$ *and* $\mathcal{Q}' = \langle \mathcal{A}, \mathcal{R}^{-\prime}, \mathcal{R}^{+\prime}, \tau, w' \rangle$ *such that* $\mathcal{R}^- \subseteq \mathcal{R}^{-\prime}$, $\mathcal{R}^+ \subseteq \mathcal{R}^{+\prime}$ *and* $w'(r) = w(r)$ *for all* $r \in \mathcal{R}$, *the following holds: for any* $\alpha, \beta \in \mathcal{A}$, *let* $\sigma_{\mathcal{Q}'}(\alpha)$ *denote the strength of* α *in* $\mathcal{Q}'$*; if* $\mathcal{R}^{-\prime} \cup \mathcal{R}^{+\prime} = \mathcal{R}^- \cup \mathcal{R}^+ \cup \{(\beta, \alpha)\}$ *and* $w'((\beta, \alpha)) = 0$, *then* $\sigma(\alpha) = \sigma_{\mathcal{Q}'}(\alpha)$.

Lemma 1. *If a modular semantics is based on an aggregation function* agg *that satisfies neutrality, then the semantics satisfies edge-neutrality.*

The following *edge-stability* property is similar to (base score) *stability* (Amgoud and Ben-Naim 2018).

Definition 6 (Edge-Stability)**.** *A gradual semantics* σ *satisfies* edge-stability *iff for any* $\alpha \in \mathcal{A}$, *whenever* $w(r) = 0$ *for all* $r \in \mathcal{R}(\alpha)$, *then* $\sigma(\alpha) = \tau(\alpha)$.

Lemma 2. *If a modular semantics is based on an aggregation function* agg *and an influence function* infl *which satisfy their associated balance properties, then the semantics satisfies edge-stability.*

The next property is a variant of the *directionality* property (Amgoud and Ben-Naim 2016).

Definition 7 (Edge-Directionality)**.** *A semantics* σ *satisfies* edge-directionality *iff, for any* $\mathcal{Q}$ *and* $\mathcal{Q}' = \langle \mathcal{A}', \mathcal{R}^{-\prime}, \mathcal{R}^{+\prime}, \tau', w' \rangle$ *such that* $\mathcal{A} = \mathcal{A}'$, $\mathcal{R}^- \subseteq \mathcal{R}^{-\prime}$, *and* $\mathcal{R}^+ \subseteq \mathcal{R}^{+\prime}$, *the following holds: for any* $\alpha, \beta, \gamma \in \mathcal{A}$, *let* $\sigma_{\mathcal{Q}'}(\gamma)$ *denote the strength of* γ *in* $\mathcal{Q}'$*; if* $\mathcal{R}^{-\prime} \cup \mathcal{R}^{+\prime} = \mathcal{R}^- \cup \mathcal{R}^+ \cup \{(\alpha, \beta)\}$ *and* $P_{\beta \mapsto \gamma} = \emptyset$, *then* $\sigma(\gamma) \equiv \sigma_{\mathcal{Q}'}(\gamma)$ *for any* $w'((\alpha, \beta)) \in [0, 1]$.

Lemma 3. *Every modular semantics satisfies edge-directionality.*

We next consider two variants of *monotonicity* properties proposed in the literature (Baroni, Rago, and Toni 2018; Yin, Potyka, and Toni 2024b).

Definition 8 (Monotonicity)**.** *A gradual semantics* σ *satisfies* monotonicity *iff, for any* $\mathcal{Q}$ *and* $\mathcal{Q}' = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau' \rangle$, *the following holds: for any* $\alpha, \beta \in \mathcal{A}$ $(\alpha \neq \beta)$ *such that* $(\beta, \alpha) \in \mathcal{R}$ *and* $|P_{\beta \mapsto \alpha}| = 1$, *let* $\sigma_{\mathcal{Q}'}(\alpha)$ *denote the strength of* α *in* $\mathcal{Q}'$, *for any* $\tau' : \mathcal{A} \to [0, 1]$*:*

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

Definition 9 (Edge-Monotonicity)**.** *A gradual semantics* σ *satisfies* edge-monotonicity *iff, for any* $\mathcal{Q}$ *and* $\mathcal{Q}' = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w' \rangle$, *the following holds: for any* $\alpha \in \mathcal{A}$, *let* $\sigma_{\mathcal{Q}'}(\alpha)$ *denote the strength of* α *in* $\mathcal{Q}'$, *for any* $r \in \mathcal{R}$ *and* $w' : \mathcal{R} \to [0, 1]$*:*

1. *If* $r \in \mathcal{R}^-(\alpha)$, $w(r) \leq w'(r)$ *and* $w(t) = w'(t)$ *for all* $t \in \mathcal{R} \setminus \{r\}$, *then* $\sigma(\alpha) \geq \sigma_{\mathcal{Q}'}(\alpha)$*;*
2. *If* $r \in \mathcal{R}^+(\alpha)$, $w(r) \leq w'(r)$ *and* $w(t) = w'(t)$ *for all* $t \in \mathcal{R} \setminus \{r\}$, *then* $\sigma(\alpha) \leq \sigma_{\mathcal{Q}'}(\alpha)$.

Lemma 4. *If a modular semantics is based on an aggregation function* agg *and an influence function* infl *which satisfy*

their associated monotonicity properties, then the semantics satisfies monotonicity and edge-monotonicity for the class of acyclic EW-QBAFs.

All commonly considered aggregation and influence functions satisfy their associated balance, monotonicity and neutrality properties (Potyka and Booth 2024, Lemma 10, 14). Thus, our previous lemmas imply the following result.

Proposition 2. *Edge-Weighted QE, REB, DF-QuAD and MLP-based semantics satisfy edge-neutrality, edge-stability, edge-directionality, monotonicity and edge-monotonicity.*

Finally, we will use the fact that the strength function under these semantics is differentiable with respect to edge-weights when the EW-QBAF is acyclic.

Lemma 5. *For acyclic EW-QBAFs, the strength function under Edge-Weighted QE, REB, DF-QuAD and MLP-based semantics is differentiable with respect to edge-weights.*

In the remainder, we will use σ for any of these edge-weighted gradual semantics.

6 Contestability Problem

In this section, we define and study the contestability problem for EW-QBAFs. We will assume that the EW-QBAFs are acyclic. While this is a restriction, many applications such as PRSs (Cocarascu, Rago, and Toni 2019; Battaglia et al. 2024; Rago et al. 2025) and MLPs naturally result in acyclic graphs due to their hierarchical structure.

Intuitively, the contestability problem for EW-QBAFs is to find a modification of edge weights that yields a desired strength for a specified topic argument [2].

Definition 10 (Contestability Problem)**.** *Given a* topic argument $\alpha \in \mathcal{A}$ *and a desired strength* s *for* α *such that* $\sigma(\alpha) \neq s$ *in* $\mathcal{Q}$*, the* contestability problem *is to identify an edge weight function* w' *such that for* $\mathcal{Q}' = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w' \rangle$*, it holds that* $\sigma_{\mathcal{Q}'}(\alpha) = s$*.*

As an example, consider the EW-QBAF in Figure 1, where a user disagrees with the current movie rating score $\sigma(\alpha) = 0.827$ and instead prefers a lower strength $s = 0.3$. This scenario can be viewed as a contestability problem, in which the edge weights can be adjusted to better capture subjective user preferences.

Our first question is whether any desired strength for a given topic argument can be attained. Cocarascu et al. (2019) introduced an *attainability* property that examines whether a desired strength can be reached by selecting a certain set of arguments, while we investigate attainability by adjusting edge weights, shifting the problem from discrete argument selection to continuous weight modification. This transition poses new theoretical challenges, as it requires analysing a high-dimensional, and often non-convex, gradual semantics.

Definition 11 (Attainability)**.** *For any* $\alpha \in \mathcal{A}$ *and* $s \in [0, 1]$*, we say that* s *is* attainable for α *iff there exists an edge weight function* w' *such that* $\sigma_{w'}(\alpha) = s$*. The* attainable set *of* α *is defined as* $S(\alpha) = \{s \mid s \text{ is attainable for } \alpha\}$.

The following proposition presents a special case where the base score is always attainable.

Proposition 3 (Base Score Attainability)**.** *For any* $\alpha \in \mathcal{A}$*, if* σ *satisfies edge-stability, then* $\tau(\alpha) \in S(\alpha)$.

To find the boundaries of a topic argument's attainable strength, we first define the $\max$ and $\min$ edge weight functions (illustrated in Figure 2). These functions represent two extreme cases: one where the topic argument receives maximal supports and no attacks, and another where it receives maximal attack and no support.

Definition 12 (Max and Min Edge Weight Functions)**.** *For any* $\alpha \in \mathcal{A}$*, the* max *and* min edge weight functions $w^\alpha_{\max}$ *and* $w^\alpha_{\min}$ *(respectively) are defined as follows:*

- $w^\alpha_{\max}(r) = 1$ *for all* $r \in \mathcal{R}^+$*;*
- $w^\alpha_{\max}(t) = 0$ *for all* $t \in \mathcal{R}^-$*;*
- $w^\alpha_{\min}(r) = 1$ *for all* $r \in \mathcal{R}^-(\alpha) \cup (\mathcal{R}^+ \setminus \mathcal{R}^+(\alpha))$*;*
- $w^\alpha_{\min}(t) = 0$ *for all* $t \in \mathcal{R}^+(\alpha) \cup (\mathcal{R}^- \setminus \mathcal{R}^-(\alpha))$.

Intuitively, to maximise $\sigma(\alpha)$ (Figure 2, left), we set all supports of α to 1 and attacks to 0 (e.g., $w(r_{10}) = w(r_{12}) = 1$, $w(r_{11}) = 0$). We then maximise its supporters and attackers (e.g., $w(r_4) = 1$, $w(r_8) = 0$; $w(r_6) = w(r_7) = w(r_9) = 1$, $w(r_5) = 0$). We maximise the attackers because attackers may also have indirect positive influence on α, for example, if there was an edge (β_5, β_6). We recursively apply this strategy until all support edges are set to 1 and all attack edges are set to 0, yielding the maximal $\sigma(\alpha)$. A detailed proof is provided in the SM.

The following theorem states that the max and min edge weight functions determine the maximum and minimum of the attainable set.

Theorem 1 (Maximum and Minimum of Attainability)**.** *For any* $\alpha \in \mathcal{A}$ *and any* σ *satisfying edge-monotonicity, monotonicity, and edge-neutrality,* $\sigma_{w^\alpha_{\max}}(\alpha)$ *and* $\sigma_{w^\alpha_{\min}}(\alpha)$ *are the maximum and minimum of* $S(\alpha)$*, respectively.*

While Theorem 1 characterises the boundaries of the attainable set, the next natural question is whether any value between the maximum and the minimum can be attained, which is confirmed by the following theorem.

Theorem 2 (Completeness of Attainability)**.** *For any* $\alpha \in \mathcal{A}$*, let* m *and* M *denote the minimum and maximum of* $S(\alpha)$*, respectively. If* σ *satisfies continuity, then* $S(\alpha) = [m, M]$.

In this section, we introduced the contestability problem and investigated its solvability. Motivated by the PRS application, our focus is on the edge weight modification as it naturally represents subjective user preferences. Although previous work (Freedman et al. 2025) suggests that modifying the base scores is also theoretically possible, this is not appropriate in our PRS setting as they represent review-based ratings of a movie. Moreover, even in scenarios where base scores can be adjusted (Civit et al. 2025b; Civit et al. 2025a), edge weight modification enables more fine-grained control. For example, in the movie The Town (2010), the argument "Ben Affleck" is associated with both

[2] In general, the contestability problem admits multiple solutions, i.e., different edge weight configurations may result in the same strength for a given topic argument.

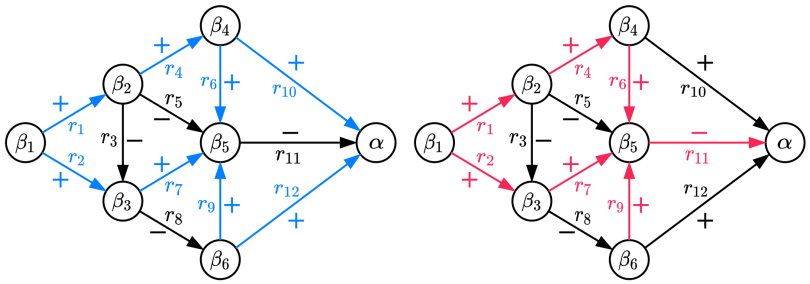

Figure 2: Illustration of the max (left) and min (right) edge weight functions. Blue or red edges are assigned a weight of 1, while black edges are assigned a weight of 0.

acting and writing, supporting one while attacking the other. A user may wish to revise the assessment of acting alone, rather than his overall contribution to the movie. While other approaches, such as adding or removing arguments (i.e., evaluation criteria) (Freedman et al. 2025), are also conceptually feasible, they often lack clarity from an application perspective: it is unclear where such arguments should originate from or which should be removed. Therefore, we assume a fixed EW-QBAF structure and focus exclusively on modifying edge weights, which provides a user-driven and interpretable mechanism for adapting to personalised user strengths. Next, we will introduce an explanation method that can guide the search for a solution.

7 Explanations and Properties

In this section, we propose a novel notion of a *gradient-based relation attribution explanation (G-RAE)* and study its properties adapted from the literature.

7.1 G-RAEs

In order to attain a desired strength of a topic argument in acyclic EW-QBAFs, we combine the ideas of gradient-based AAEs and RAEs, and propose a novel notion of *G-RAEs*, which capture the sensitivity of the strength of a topic argument with respect to the changes of individual edge weights.

Definition 13 (Gradient-based Relation Attribution Explanations (G-RAEs))**.** *Let $r \in \mathcal{R}$ and $\alpha \in \mathcal{A}$ be a* topic argument. *For a perturbation $\varepsilon \in [-w(r), 0) \cup (0, 1 - w(r)]$, let w' be an edge weight function such that $w'(r) = w(r) + \varepsilon$ and $w'(t) = w(t)$ for all $t \in \mathcal{R} \setminus \{r\}$. The* G-RAE *from r to α under σ is*

$$\nabla^{\sigma}_{r \mapsto \alpha} = \lim_{\varepsilon \to 0} \frac{\sigma_{w'}(\alpha) - \sigma(\alpha)}{\varepsilon}.$$

Let us note that, by Lemma 5, $\nabla^{\sigma}_{r \mapsto \alpha} \in \mathbb{R}$ is always well-defined in acyclic EW-QBAFs. We next distinguish attribution influence based on the sign of G-RAE.

Definition 14 (Attribution Influence)**.** *We say that the* attribution influence *from r to α is* positive *if $\nabla^{\sigma}_{r \mapsto \alpha} > 0$,* negative *if $\nabla^{\sigma}_{r \mapsto \alpha} < 0$, and* neutral *if $\nabla^{\sigma}_{r \mapsto \alpha} = 0$.*

Like explanations scores, G-RAEs reveal both the direction and magnitude of each edge's influence on the topic argument, which could be practically useful in various domains such as PRS contestation or MLP debugging, where identifying and adjusting the most influential edges may directly enhance user satisfaction or improve model performance.

Example 2. *Consider the EW-QBAF in Figure 1 where σ is given by MLP-based semantics. The G-RAEs of edges are computed by Definition 13: $(Acting, Movie)$=0.02413, $(Themes, Movie)$=0.01738, $(Meryl, Acting) = 0.00134$, $(Tom, Acting) = 0.00095$, $(Freedom, Themes) = 0.00086$, $(Romance, Themes) = -0.00064$, $(Writing, Movie) = -0.00287$.*

7.2 Properties of G-RAEs

We next adapt some commonly used properties of argumentative attribution explanations from (Yin, Potyka, and Toni 2023b; Yin, Potyka, and Toni 2024c), and examine the satisfaction of these properties.

We start by analysing the influence of edges with different connectivities (Definition 2) in the following propositions. Direct Influence (Yin, Potyka, and Toni 2023b; Yin, Potyka, and Toni 2024c) shows that our G-RAEs correctly capture the qualitative effects of direct connectivity: a direct attack (support) always yields non-positive (non-negative, respectively) influence on the topic argument.

Proposition 4 (Direct Influence)**.** *Let $r \in \mathcal{R}$ be a direct edge w.r.t. $\alpha \in \mathcal{A}$. Then the following statements hold if σ satisfies edge-monotonicity:*

1. *If $r \in \mathcal{R}^+$, then $\nabla^{\sigma}_{r \mapsto \alpha} \geq 0$;*
2. *If $r \in \mathcal{R}^-$, then $\nabla^{\sigma}_{r \mapsto \alpha} \leq 0$.*

We next make several differentiations for indirect connectivity (Rago, Li, and Toni 2023; Yin, Potyka, and Toni 2024c) because the polarity of an edge may invert along the path. For example, an attacker of an attacker actually serves as a supporter.

Proposition 5 (Indirect Influence)**.** *Let r be an indirect edge w.r.t. $\alpha \in \mathcal{A}$. Suppose the path sequence from r to α is $\langle r, r_1, \cdots, r_n \rangle (n \geq 1)$. Let $\lambda = |\{r_1, \cdots, r_n\} \cap \mathcal{R}^-|$. Then the following statements hold if σ satisfies both monotonicity and edge-monotonicity.*

1. *If $r \in \mathcal{R}^+$ and λ is odd, then $\nabla^{\sigma}_{r \mapsto \alpha} \leq 0$;*
2. *If $r \in \mathcal{R}^+$ and λ is even, then $\nabla^{\sigma}_{r \mapsto \alpha} \geq 0$;*
3. *If $r \in \mathcal{R}^-$ and λ is odd, then $\nabla^{\sigma}_{r \mapsto \alpha} \geq 0$;*
4. *If $r \in \mathcal{R}^-$ and λ is even, then $\nabla^{\sigma}_{r \mapsto \alpha} \leq 0$.*

Example 3. *In Example 2, since* $(Acting, Movie)$ *and* $(Writing, Movie)$ *are direct support and attack w.r.t. the topic argument* $Movie$*, their G-RAEs are positive and negative, respectively, by Proposition 4. In addition, since the path from* $Romance$ *to* $Movie$ *contains one attack (i.e., an odd number of attacks), the G-RAE of* $(Romance, Movie)$ *is negative by Proposition 5.*

Irrelevance states that any edge independent of the topic argument has no influence.

Proposition 6 (Irrelevance)**.** *Let* $\alpha \in \mathcal{A}$*. If* $r \in \mathcal{R}$ *is an independent edge w.r.t.* α *and* σ *satisfies edge-directionality, then* $\nabla^{\sigma}_{r\mapsto\alpha} = 0$*.*

In general, multifold edges may influence a topic argument through multiple paths with mixed polarities. As illustrated in (Yin, Potyka, and Toni 2024b, Figure 4), such mixtures can lead to non-monotonic effects, even under monotonic semantics, making it difficult to reason about their overall influence. To address this challenge, we focus on a more tractable subclass — single-polarity multifold edges — where all paths share the same polarity. This restriction allows us to derive formal guarantees on the direction of influence. We formalize this notion as follows:

Definition 15 (Single-Polarity Multifold Edge)**.** *Let* r *be a multifold edge w.r.t.* $\alpha \in \mathcal{A}$ *via paths* $p_1, \cdots, p_m$*. For each path* p_i *(*$i \in \{1, \ldots, m\}$*), suppose the sequence from* r *to* α *is* $\langle r, r_{i1}, \cdots, r_{i|p_i|}\rangle (|p_i| \geq 1)$*. Let* $\lambda_i = \left|\{r_{i1}, \cdots, r_{i|p_i|}\} \cap \mathcal{R}^-\right|$*. We say that* r *is a* single-polarity multifold edge *w.r.t.* α *if either* λ_i *is odd for all* i*, or* λ_i *is even for all* i*.*

Example 4. *Figure 3 shows a single-polarity multifold edge* r *w.r.t.* α *as every path contains even numbers of attacks.*

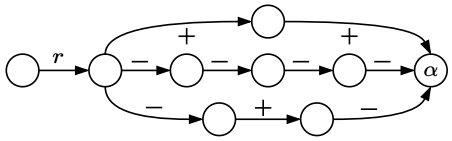

Figure 3: Single-polarity multifold edge.

The following proposition states that the influence of a single-polarity multifold edge on a topic argument is either consistently positive or consistently negative.

Proposition 7 (Single-Polarity Multifold Influence)**.** *Let* r *be a single-polarity multifold edge w.r.t.* $\alpha \in \mathcal{A}$*.*
1. If $r \in \mathcal{R}^+$ *and* λ_i *is odd for all* i*, then* $\nabla^{\sigma}_{r\mapsto\alpha} \leq 0$*;*
2. If $r \in \mathcal{R}^+$ *and* λ_i *is even for all* i*, then* $\nabla^{\sigma}_{r\mapsto\alpha} \geq 0$*;*
3. If $r \in \mathcal{R}^-$ *and* λ_i *is odd for all* i*, then* $\nabla^{\sigma}_{r\mapsto\alpha} \geq 0$*;*
4. If $r \in \mathcal{R}^-$ *and* λ_i *is even for all* i*, then* $\nabla^{\sigma}_{r\mapsto\alpha} \leq 0$*.*

Next, we adapt two existing properties to EW-QBAFs and examine their satisfaction in the direct and indirect cases. Counterfactuality, inspired by (Yin, Potyka, and Toni 2023a; Yin, Potyka, and Toni 2024c), considers how the strength of a topic argument changes when an edge is removed, corresponding, in our setting, to setting the edge weight to 0.

Property 1 (Counterfactuality)**.** *Let* $\alpha \in \mathcal{A}$ *and* $r \in \mathcal{R}$*. Let* w' *be an edge weight function such that* $w'(r) = 0$ *and* $w'(t) = w(t)$ *for all* $t \in \mathcal{R} \setminus \{r\}$*.*
1. If $\nabla^{\sigma}_{r\mapsto\alpha} < 0$*, then* $\sigma(\alpha) \leq \sigma_{w'}(\alpha)$*;*
2. If $\nabla^{\sigma}_{r\mapsto\alpha} > 0$*, then* $\sigma(\alpha) \geq \sigma_{w'}(\alpha)$*.*

Proposition 8. *Let* r *be a direct, indirect, or single-polarity multifold edge w.r.t.* α*.* $\nabla^{\sigma}_{r\mapsto\alpha}$ *satisfies counterfactuality if* σ *satisfies both edge-monotonicity and monotonicity.*

As an illustration, if we set $w(Acting, Movie) = 0$, then $\sigma(Movie)$ decreases from 0.827 to 0.802.

Qualitative invariability, inspired by (Yin, Potyka, and Toni 2023a; Yin, Potyka, and Toni 2024c), states that the influence of an edge on the topic argument remains qualitatively consistent (i.e., always positive or always negative), irrespective of any variations in its weight.

Property 2 (Qualitative Invariability)**.** *Let* $\alpha \in \mathcal{A}$ *and* $r \in \mathcal{R}$*. Let* ∇_δ *denote* $\nabla^{\sigma}_{r\mapsto\alpha}$ *when setting* $w(r)$ *to some* $\delta \in [0, 1]$*.*
1. If $\nabla^{\sigma}_{r\mapsto\alpha} < 0$*, then* $\forall \delta \in [0, 1]$*,* $\nabla_\delta \leq 0$*;*
2. If $\nabla^{\sigma}_{r\mapsto\alpha} > 0$*, then* $\forall \delta \in [0, 1]$*,* $\nabla_\delta \geq 0$*.*

Proposition 9. *Let* r *be a direct, indirect, or single-polarity multifold edge w.r.t.* α*.* $\nabla^{\sigma}_{r\mapsto\alpha}$ *satisfies qualitative invariability if* σ *satisfies edge-monotonicity and monotonicity.*

As an illustration, even when the edge weights in Figure 1 are altered, their G-RAEs satisfy $\nabla^{\sigma}_{r\mapsto\alpha} \geq 0$ or $\nabla^{\sigma}_{r\mapsto\alpha} \leq 0$.

The final property shows that our G-RAEs can be computed efficiently.

Proposition 10 (Tractability)**.** *Let* $|\mathcal{A}| = m$ *and* $|\mathcal{R}| = n$*, then G-RAEs can be generated in linear time* $\mathcal{O}(m + n)$ *for acyclic EW-QBAFs.*

To summarise, in this section, we defined G-RAEs and examined the influence of three types of connectivity. While the general multifold case is analytically challenging, as discussed earlier, we focus on the single-polarity subclass as this is the setting in which monotonicity guarantees can be meaningfully applied. Exploring other subclasses among multifold edges is left for future work.

8 Contestability Algorithms

In this section, we propose an approximation algorithm for G-RAEs that can also be applied to cyclic EW-QBAFs and a heuristic algorithm for solving the contestability problem.

To simplify the implementation and avoid handling the structure variations of different EW-QBAFs, we adopt a perturbation-based method to approximate G-RAEs. This choice also leaves open the possibility of extending the method to cyclic EW-QBAFs, where exact analytical gradients may be less straightforward to derive. Perturbation-based approaches are also widely used in the literature for gradient estimation tasks when the gradient cannot be computed analytically (e.g., (Ozbulak et al. 2020; Minervini, Franceschi, and Niepert 2023)). Algorithm 1 computes the G-RAE for all edges with respect to the topic argument in an EW-QBAF. The algorithm begins by computing the original strength of the topic argument α under the gradual semantics σ (line 2). Then, for an edge r, the algorithm perturbs $w(r)$ by a value ε and recomputes $\sigma(\alpha)$ based on the perturbed edge weight (lines 4-5). The approximate G-RAE of

Algorithm 1 G-RAE Approximation

Input: An EW-QBAF $\mathcal{Q}$, an edge-weighted gradual semantics σ, a topic argument α.
Parameter: A perturbation value ε.
Output: Approximate G-RAEs g_rae.
Function: $gRAE(\mathcal{Q}, \sigma, \alpha, \varepsilon)$

```
1: g_rae ← {}                                  % attribution scores
2: s_0 ← σ(α)                                  % original strength
3: for r in R do
4:     w[r] ← w[r] + ε                         % perturb w(r)
5:     s' ← σ(α)                               % perturbed strength
6:     g_rae[r] ← (s' − s_0)/ε                 % compute G-RAE
7:     w[r] ← w[r] − ε                         % restore w(r)
8: end for
9: return g_rae
```

Algorithm 2 Contestability Algorithm

Input: An EW-QBAF $\mathcal{Q}$, an edge-weighted gradual semantics σ, a topic argument α and a desired strength s for α.
Parameter: A maximum iteration limit M, an updating step h, an error threshold δ.
Output: An edge weight function w′.

```
1: w' ← w
2: s' ← σ(α)
3: while |s' − s| > δ and M > 0 do
4:     g_rae ← gRAE(Q, σ, α, ε)
5:     for r in R do
6:         update ← w'[r] + g_rae[r] · h
7:         w'[r] ← max(0, min(1, update))
8:     end for
9:     s' ← σ(α)
10:    M ← M − 1
11: end while
12: return w'
```

r is then obtained by dividing the strength change of α by ε (line 6). After computing the G-RAE of edge r, the original weight $w(r)$ is restored for proceeding to the next edge (line 7). The above procedure is iteratively applied to compute the G-RAE for all edges.

Proposition 11 (G-RAE Approximation Complexity)**.** *Let $|\mathcal{A}| = m$ and $|\mathcal{R}| = n$, then all approximate G-RAEs can be generated in time $\mathcal{O}(n \cdot (m + n))$ for acyclic EW-QBAFs.*

Since G-RAEs quantify the influence of edge weights on the strength changes of the topic argument, we utilize them to guide a heuristic algorithm that adjusts edge weights to solve the contestability problem. Algorithm 2 begins by computing the strength of the topic argument α under the gradual semantics σ (line 2). For brevity, we assume the current strength of the topic argument is less than the desired one. If the difference between this strength and the desired strength exceeds a predefined error threshold δ and the iteration count has not yet reached a maximum iteration limit M (introduced to prevent the algorithm from running indefinitely or getting stuck in a local minimum) (line 3), the edge weights are updated by adding their respective G-RAEs multiplied by an updating step h (line 6). Since the edge weights are constrained within the range $[0, 1]$, we apply the $\max$ and $\min$ functions to ensure the updated weights remain within bounds (line 7). Once all edge weights are updated, $\sigma(\alpha)$ is recomputed and the iteration count M is decremented (lines 9-10). The algorithm terminates when either $\sigma(\alpha)$ is sufficiently close to the desired strength or the iteration limit is reached. The algorithm's time complexity is as follows.

Proposition 12 (Contestability Algorithm Complexity)**.** *Let $|\mathcal{A}| = m$ and $|\mathcal{R}| = n$, and suppose the maximum number of iterations is M. Then the time complexity of Algorithm 2 is $\mathcal{O}(M \cdot n \cdot (m + n))$ for acyclic EW-QBAFs.*

9 Experimental Evaluation

We evaluated effectiveness (ability to achieve desired strength) and scalability (performance on dense EW-QBAFs) of our Algorithm 2 in two application scenarios: PRSs (Experiment 1) and MLP debugging (Experiment 2).[3]

Algorithm Parameter Settings In Algorithm 1, the perturbation value ε was set to 10^{-5} [4]. In Algorithm 2, the error threshold δ was set to 0.01 and the maximum number of iterations M was set to 1000. As general guidance, $\varepsilon = 10^{-5}$ is a standard choice for finite-difference approximation, δ should be set according to the desired precision, and M can start at a moderate value (e.g., 1000) and be increased if convergence is not reached. To accelerate convergence, we used a dynamic updating step schedule, applying a larger/smaller step when the current strength was farther from/closer to (respectively) the desired strength.

9.1 Experiment 1

In this experiment, we aimed to assess performance of our algorithm on acyclic EW-QBAFs that mirror the structure of PRSs to show the potential of addressing misalignments between the EW-QBAF outcomes and subjective user preferences.

Setups In order to randomly generate acyclic EW-QBAFs, we first created n arguments $(\alpha_1, \cdots, \alpha_n)$, each assigned a random base score drawn uniformly from the interval [0,1]. For every pair of arguments (α_i, α_j) such that $i < j$, an edge was generated with probability p. Since QBAFs in many applications typically contain fewer than 100 arguments and have roughly equal numbers of arguments and edges (e.g., for movie recommender systems (Cocarascu, Rago, and Toni 2019), fake news detection (Kotonya and Toni 2019), fraud detection (Chi et al. 2021)), we generated EW-QBAFs of varying size with $n \in \{10, 20, \ldots, 100\}$ and $p = \frac{2}{n}$. Each edge was randomly designated as either attack or support with equal probability, and its weight was uniformly drawn from the interval [0,1]. The topic argument was designated as α_n, and its desired strength was set to the mean of the maximum and minimum attainable strength.

[3] For detailed results in this section, see Table 1 and 2 in the SM.

[4] For acyclic EW-QBAFs, exact gradients can also be computed via backpropagation.

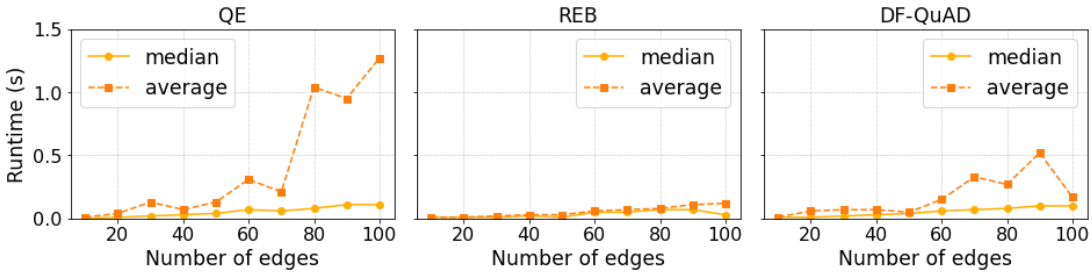

Figure 4: Average and median runtime over 100 randomly generated PRS-like QBAFs under different semantics and edge sizes.

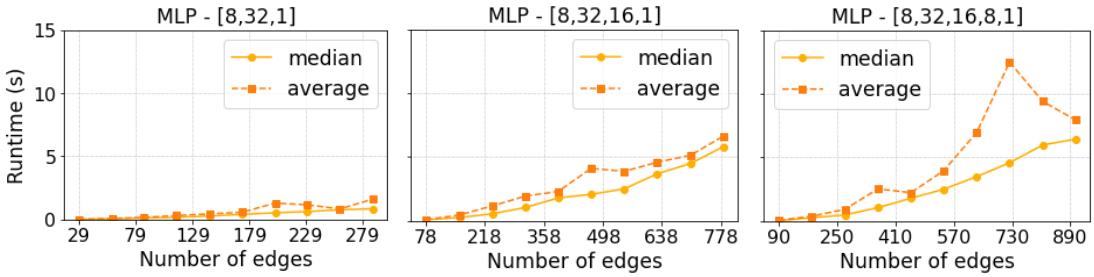

Figure 5: Average and median runtime over 100 randomly generated MLP-like QBAFs, evaluated under MLP-based semantics.

This choice was motivated by Theorem 2, which guarantees that the mean value is always attainable. Note that, depending on p, the topic argument may have no or only a few predecessors in this setting. To mitigate the effects of randomness, we generated 100 QBAFs for each size. To compute argument strengths, we applied three commonly considered gradual semantics - QE (Potyka 2018), REB (Amgoud and Ben-Naim 2018), and DF-QuAD (Rago et al. 2016) - each adapted to incorporate edge weights to suit our setting.

Results and Analysis We evaluated (i) *validity* (i.e., whether the desired strength was attained), (ii) *runtime*, and (iii) number of *attempts* of our algorithm across different semantics and varying edge sizes.

(i) Across all 100 EW-QBAFs of each structure, our algorithm's validity consistently reached 100%, regardless of semantics or number of edges, indicating that it was always able to identify the desired strength, in line with Theorem 2.

(ii) Since our algorithm is based on gradient descent, there is a potential risk of converging to a local minimum. To address this, whenever the solution was not found within the maximum number of iterations M, we adjusted the initial search point and re-executed the algorithm. Under the REB semantics, our algorithm consistently succeeded on the first attempt across all EW-QBAF sizes, followed by the QE semantics with an average of 1.01 attempts. Although DF-QuAD required the highest average number of attempts (1.014), the maximum never exceeded 4 - matching that of the QE semantics - showing that the algorithm can find solutions with a reasonable number of attempts in all cases.

(iii) Figure 4 shows performance, under the same EW-QBAF structure. The REB semantics generally exhibited the shortest average runtime, followed by DF-QuAD, while QE incurred the longest. Despite these differences, the average runtime across all semantics and all EW-QBAF sizes remained under 1.5 seconds. Due to the random generation of EW-QBAFs, certain instances required significantly more time to process. Thus, we also reported the median runtime. Across all three semantics, the median runtimes were comparable and exhibited an overall increasing trend with minor fluctuations[5] as the number of edges increased. In all cases, the median runtime remained below 0.12 seconds.

9.2 Experiment 2

In this experiment, we aimed to evaluate our algorithm on denser QBAFs that mirror the structure of MLPs. As shown in (Potyka 2021), MLPs can be translated into EW-QBAFs, and our algorithm has the potential to assist in debugging MLPs by achieving a desired output for a given instance.

Setups We generated denser QBAFs compared to those in Experiment 1. Specifically, we designed three MLP-like QBAFs structures: $[8, 32, 1]$, $[8, 32, 16, 1]$ and $[8, 32, 16, 8, 1]$, varying in the numbers of neurons and hidden layers. For example, $[8, 32, 1]$ consists of 8 input arguments, three hidden layers with 32, 16, and 8 arguments respectively, and 1 output argument, yielding a total of $8 + 32 + 16 + 8 + 1 = 65$ arguments and $8 \times 32 + 32 \times 16 + 16 \times 8 + 8 \times 1 = 904$ edges in the fully-connected case. We chose these structures because they are suitable for many binary classification tasks such as Pima Indians diabetes classification problem (Potyka, Yin, and Toni 2022).

[5]Minor measurement inaccuracies when execution time falls below 0.1 seconds may contribute to the observed fluctuations.

For each structure, the base scores of arguments, as well as the polarity and weights of edges, were assigned in the same manner as in Experiment 1. The topic argument was set as the output-layer argument. We varied the connection density by setting the probability p of an edge between any two arguments in adjacent layers, with $p \in \{0.1, 0.2, \ldots, 1.0\}$. A value of $p = 1.0$ amounted to a fully-connected MLP-like EW-QBAF. We generated 100 EW-QBAFs for each structure to avoid randomness and applied MLP-based gradual semantics (Potyka 2021) to compute argument strengths.

Results and Analysis The validity consistently reached 100% across all EW-QBAF structures, indicating our algorithm's ability to attain the desired strengths, even in significantly denser configurations compared to those in Experiment 1. We next evaluate the number of attempts and running time. For all structures, our algorithm succeeded on the first attempt, except for the case of $[8, 32, 1]$, where a single fully-connected ($p = 0.7$) MLP-like EW-QBAF required one additional attempt. Overall, the results outperformed those of Experiment 1. A possible explanation lies in the structural constraint imposed in this setting: each argument only connects to arguments in the next layer, resulting in a simpler and more hierarchical structure. As shown in Figure 5, the average runtime increased with the number of edges across all structures, with the maximum average reaching 12.50 seconds. Meanwhile, the median runtime exhibited a strictly increasing trend, peaking at 6.44 seconds. The polynomial runtime results showed our algorithm remains scalable as the density of EW-QBAFs increases.

10 Conclusion

We explored the use of EW-QBAFs for supporting contestability. Specifically, we formalized the contestability problem and proposed G-RAEs as interpretable guidance. Building on G-RAEs, we gave an algorithm to adjust edge weights and showed its effectiveness on synthetic scenarios, highlighting the practical potential of our approach.

Our method could be exploited to covertly steer outcomes without user awareness. Thus, exploring mechanisms to track and expose such modifications would help ensure transparency and trust. Second, user studies could help identify which solution characteristics users find most intuitive or acceptable when contesting model outputs to inform the development of selection criteria that prioritize user-preferred modifications to edge weights.

Acknowledgments

Yin and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX). Toni was also funded by EPSRC (grant UKRI3928, NeSyDebates). Kampik was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Any views or opinions expressed herein are solely those of the authors.

References

Alfrink, K.; Keller, I.; Kortuem, G.; and Doorn, N. 2023. Contestable AI by design: Towards a framework. *Minds Mach.* 33(4):613–639.

Almada, M. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL 2019, Montreal, QC, Canada, June 17-21, 2019*, 2–11. ACM.

Amgoud, L., and Ben-Naim, J. 2016. Evaluation of arguments from support relations: Axioms and semantics. In Kambhampati, S., ed., *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 900–906. IJCAI/AAAI Press.

Amgoud, L., and Ben-Naim, J. 2018. Evaluation of arguments in weighted bipolar graphs. *Int. J. Approx. Reason.* 99:39–55.

Amgoud, L.; Ben-Naim, J.; and Vesic, S. 2017. Measuring the intensity of attacks in argumentation graphs with shapley value. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 63–69. ijcai.org.

Anaissy, C. A.; Delobelle, J.; Vesic, S.; and Yun, B. 2024. Impact measures for gradual argumentation semantics. *arXiv preprint arXiv:2407.08302*.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Mag.* 38(3):25–36.

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and Müller, K. 2010. How to explain individual classification decisions. *J. Mach. Learn. Res.* 11:1803–1831.

Baroni, P.; Rago, A.; and Toni, F. 2018. How many properties do we need for gradual argumentation? In McIlraith, S. A., and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 1736–1743. AAAI Press.

Battaglia, E.; Baroni, P.; Rago, A.; and Toni, F. 2024. Integrating user preferences into gradual bipolar argumentation for personalised decision support. In Destercke, S.; Martinez, M. V.; and Sanfilippo, G., eds., *Scalable Uncertainty Management - 16th International Conference, SUM 2024, Palermo, Italy, November 27-29, 2024, Proceedings*, volume 15350 of *Lecture Notes in Computer Science*, 14–28. Springer.

Chen, L.; Yin, X.; and Toni, F. 2025. Latent debate: A surrogate framework for interpreting llm thinking. *arXiv preprint arXiv:2512.01909*.

Chi, H.; Lu, Y.; Liao, B.; Xu, L.; and Liu, Y. 2021. An optimized quantitative argumentation debate model for fraud

detection in e-commerce transactions. *IEEE Intell. Syst.* 36(2):52–63.

Civit, A.; Andriella, A.; Sierra, C.; and Alenyà, G. 2025a. Multi-user personalisation in human-robot interaction: Resolving preference conflicts using gradual argumentation. *arXiv preprint arXiv:2511.03576*.

Civit, A.; Rago, A.; Andriella, A.; Alenyà, G.; and Toni, F. 2025b. From user preferences to base score extraction functions in gradual argumentation. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2026*.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In Elkind, E.; Veloso, M.; Agmon, N.; and Taylor, M. E., eds., *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, 1261–1269. International Foundation for Autonomous Agents and Multiagent Systems.

Cyras, K.; Kampik, T.; and Weng, Q. 2022. Dispute trees as explanations in quantitative (bipolar) argumentation. In Cyras, K.; Kampik, T.; Cocarascu, O.; and Rago, A., eds., *1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), Cardiff, Wales, September 12, 2022*, volume 3209 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2025. Argumentative large language models for explainable and contestable claim verification. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 14930–14939. AAAI Press.

Goodfellow, I. J.; Bengio, Y.; and Courville, A. C. 2016. *Deep Learning*. Adaptive computation and machine learning. MIT Press.

Hirsch, T.; Merced, K.; Narayanan, S. S.; Imel, Z. E.; and Atkins, D. C. 2017. Designing contestability: Interaction design, machine learning, and mental health. In Mival, O. H.; Smyth, M.; and Dalsgaard, P., eds., *Proceedings of the 2017 Conference on Designing Interactive Systems, DIS '17, Edinburgh, United Kingdom, June 10-14, 2017*, 95–99. ACM.

Kampik, T.; Potyka, N.; Yin, X.; Cyras, K.; and Toni, F. 2024. Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. *Int. J. Approx. Reason.* 173:109255.

Kampik, T.; Yin, X.; Potyka, N.; and Toni, F. 2026. Strength change explanations in quantitative argumentation.

Kotonya, N., and Toni, F. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In Stein, B., and Wachsmuth, H., eds., *Proceedings of the 6th Workshop on Argument Mining, ArgMining@ACL 2019, Florence, Italy, August 1, 2019*, 156–166. Association for Computational Linguistics.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; Yin, X.; Zhang, D.; and Toni, F. 2024. Contestable AI needs computational argumentation. In Marquis, P.; Ortiz, M.; and Pagnucco, M., eds., *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. 4765–4774.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum. Comput. Interact.* 5(CSCW1):106:1–106:25.

Minervini, P.; Franceschi, L.; and Niepert, M. 2023. Adaptive perturbation-based gradient estimation for discrete latent variable models. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 9200–9208. AAAI Press.

Mossakowski, T., and Neuhaus, F. 2018. Modular semantics and characteristics for bipolar weighted argumentation graphs. *CoRR* abs/1807.06685.

Ozbulak, U.; Gasparyan, M.; Neve, W. D.; and Messem, A. V. 2020. Perturbation analysis of gradient-based adversarial attacks. *Pattern Recognit. Lett.* 135:313–320.

Peacock, D.; Mansi, N. P.; Toni, F.; and Yin, X. 2025. On the impact of sparsification on quantitative argumentative explanations in neural networks.

Potyka, N., and Booth, R. 2024. An empirical study of quantitative bipolar argumentation frameworks for truth discovery. In Reed, C.; Thimm, M.; and Rienstra, T., eds., *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18-20, 2024*, volume 388 of *Frontiers in Artificial Intelligence and Applications*, 205–216. IOS Press.

Potyka, N.; Yin, X.; and Toni, F. 2022. Towards a theory of faithfulness: Faithful explanations of differentiable classifiers over continuous data. *CoRR* abs/2205.09620.

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

Potyka, N. 2018. Continuous dynamical systems for weighted bipolar argumentation. In Thielscher, M.; Toni, F.; and Wolter, F., eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018*, 148–157. AAAI Press.

Potyka, N. 2019. Extending modular semantics for bipolar weighted argumentation. In Elkind, E.; Veloso, M.; Agmon, N.; and Taylor, M. E., eds., *Proceedings of the 18th International Conference on Autonomous Agents and Mul-*

tiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019, 1722–1730. International Foundation for Autonomous Agents and Multiagent Systems.

Potyka, N. 2021. Interpreting neural networks as quantitative argumentation frameworks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 6463–6470. AAAI Press.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In Baral, C.; Delgrande, J. P.; and Wolter, F., eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, 63–73. AAAI Press.

Rago, A.; Cocarascu, O.; Oksanen, J.; and Toni, F. 2025. Argumentative review aggregation and dialogical explanations. *Artif. Intell.* 340:104291.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive explanations by conflict resolution via argumentative exchanges. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, 582–592.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B.; Shah, M.; Smola, A. J.; Aggarwal, C. C.; Shen, D.; and Rastogi, R., eds., *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144. ACM.

Russo, F., and Toni, F. 2023. Causal discovery and knowledge injection for contestable neural networks. In Gal, K.; Nowé, A.; Nalepa, G. J.; Fairstein, R.; and Radulescu, R., eds., *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*. IOS Press. 2025–2032.

Shao, X.; Rienstra, T.; Thimm, M.; and Kersting, K. 2020. Towards understanding and arguing with classifiers: Recent progress. *Datenbank-Spektrum* 20(2):171–180.

Shapley, L. S. 1953. A value for n-person games. *Contribution to the Theory of Games* 2.

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

Yin, X.; Potyka, N.; and Toni, F. 2023b. Argument attribution explanations in quantitative bipolar argumentation frameworks. In *ECAI 2023*. IOS Press. 2898–2905.

Yin, X.; Potyka, N.; and Toni, F. 2024a. Applying attribution explanations in truth-discovery quantitative bipolar argumentation frameworks. *arXiv preprint arXiv:2409.05831*.

Yin, X.; Potyka, N.; and Toni, F. 2024b. Ce-qarg: Counterfactual explanations for quantitative bipolar argumentation frameworks. In Marquis, P.; Ortiz, M.; and Pagnucco, M., eds., *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*.

Yin, X.; Potyka, N.; and Toni, F. 2024c. Explaining arguments' strength: Unveiling the role of attacks and supports. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, 3622–3630. ijcai.org.