CE-QArg: Counterfactual Explanations for Quantitative Bipolar Argumentation Frameworks

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

Abstract

There is a growing interest in understanding arguments' strength in Quantitative Bipolar Argumentation Frameworks (QBAFs). Most existing studies focus on attribution-based methods that explain an argument's strength by assigning importance scores to other arguments but fail to explain how to change the current strength to a desired one. To solve this issue, we introduce *counterfactual explanations* for QBAFs. We discuss problem variants and propose an iterative algorithm named *Counterfactual Explanations for Quantitative bipolar Argumentation frameworks (CE-QArg)*. CE-QArg can identify *valid* and *cost-effective* counterfactual explanations based on two core modules, *polarity* and *priority*, which help determine the updating direction and magnitude for each argument, respectively. We discuss some formal properties of our counterfactual explanations and empirically evaluate CE-QArg on randomly generated QBAFs.

1 Introduction

Explainable AI (XAI) aims to enhance the transparency and trustworthiness of AI models by providing explanations for their decision-making process (Adadi and Berrada 2018), which is crucial in high-stakes decision-making domains such as healthcare, finance, and judiciary. Recently, explaining the reasoning process of Quantitative Bipolar Argumentation Frameworks (QBAFs) (Baroni et al. 2015) has received increasing attention (Kampik and Čyras 2022; Yin, Potyka, and Toni 2023; Kampik et al. 2024). QBAFs consist of *arguments*, binary relations (of *support* and *attack*), and a *base score function* that ascribes initial strengths to each argument. QBAFs semantics typically determine each argument's *(final) strength* based on the strength of its attackers and supporters (e.g. see (Leite and Martins 2011; Baroni et al. 2015; Amgoud and Ben-Naim 2018)), which allows quantitative reasoning among contradictory information (Čyras et al. 2021; Potyka 2021; Ayoobi, Potyka, and Toni 2023; Potyka, Yin, and Toni 2023). To explain arguments' final strength, often attribution-based methods are applied (e.g. see (Kampik and Čyras 2022; Kampik et al. 2024; Yin, Potyka, and Toni 2023)). These methods assign "importance scores" to arguments, showing how much they contribute to the final strength of arguments of interest.

To illustrate the idea, Figure 1 shows a QBAF to decide whether a person's loan application will be approved. This QBAF has a hierarchical structure, where at the top level is the *topic argument* α, determining whether the loan will be approved. Two other arguments influence α: (i) β (high annual salary) supports α, but it is also attacked by ρ (high risk of layoffs); (ii) γ (poor credit score) attacks α, and it is supported by ζ (high number of late payments). The base score of α is initially set to 0.5, while the base scores for β, γ, ρ, and ζ are 0.3, 0.6, 0.7, and 0.4, respectively. We apply the DF-QuAD semantics (Rago et al. 2016) (denoted as σ_{DF}) to evaluate this QBAF: the loan will be approved if $\sigma_{DF}(\alpha) \geq 0.5$, and rejected otherwise. In this scenario, the bank rejects the applicant because the final strength is lower than the desired strength ($\sigma_{DF}(\alpha) = 0.165 < 0.5$). To explain such an outcome, for instance, the Shapley-based attribution method (Kampik et al. 2024), finds that β has a positive importance score of 0.0975 wrt. α, while γ, ζ, and ρ all have negative importance scores of -0.34, -0.0525 and -0.04, respectively[1]. Since the negative scores outweigh the positive one, the final strength is lower than the desired strength, resulting in an unsuccessful application.

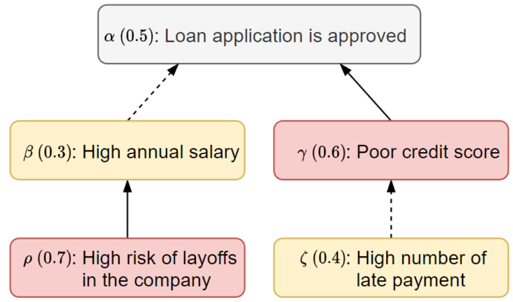
α (0.5): Loan application is approved
β (0.3): High annual salary
γ (0.6): Poor credit score
ρ (0.7): High risk of layoffs in the company
ζ (0.4): High number of late payment

Figure 1: An example QBAF for loan application. (Solid and dashed edges indicate *attack* and *support*, respectively; the numbers in brackets are the arguments' base scores).

While attribution explanations are intuitive, in this example and more generally, they fail to offer guidance on how to modify the topic argument's strength, e.g., in the example, to improve one's chance of getting approved. In contrast, *counterfactual explanations* (Wachter, Mittelstadt, and Rus-

[1] See Section 3 and https://arxiv.org/abs/2407.08497 for more details of the computation of strengths and importance scores, respectively.

sell 2017) explain how an AI model's output would change if one alters the inputs. These explanations are comprehensible because they elicit causal reasoning and thinking in humans (Byrne 2019; Verma, Dickerson, and Hines 2020).

In this work, we introduce counterfactual explanations to the QBAF setting to compensate for the limitations of attribution explanations. Here, counterfactually explaining a QBAF means to identify a base score function that could lead to a specified final strength of a topic argument under a given gradual semantics. For example, in Figure 1, if the applicant had a different base score function that could result in a desired final strength of α, then the loan would have been approved.

The main contributions of our paper are:

- We formally define three counterfactual problems for QBAFs (Section 4).
- We propose formal properties guiding the design of explanation methods to solve counterfactual problems for QBAFs and propose an iterative algorithm (CE-QArg) to generate valid, cost-effective counterfactuals[2] (Section 5).
- We propose formal properties for counterfactuals (Section 6).
- We empirically show the high effectiveness, scalability, and robustness of CE-QArg (Section 7).

The proofs of all technical results are in https://arxiv.org/abs/2407.08497. The code is available at https://github.com/XiangYin2021/CE-QArg.

2 Related Work

The literature on QBAF explanations primarily focuses on attribution-based explanations. These methods aim to explain the reasoning outcome for (final strength of) a particular argument (referred to as the topic argument) in a QBAF by assigning importance scores to arguments. There are various ways to define these importance scores, including removal-based methods (Kampik et al. 2024) that measure the impact of removing an argument; gradient-based methods (Yin, Potyka, and Toni 2023) that measure sensitivity with respect to an argument's base score; and Shapley-based methods (Kampik et al. 2024) that distribute the overall impact among all arguments using tools from game theory. Shapley-based methods can also be used to attribute importance scores to edges (attacks and supports) rather than arguments in a QBAF to explain their impact (Amgoud, Ben-Naim, and Vesic 2017; Yin, Potyka, and Toni 2024). While attribution-based explanations are helpful and intuitive in explaining outcomes (final strength), they cannot guide improving them by altering the inputs (QBAFs). We focus instead on counterfactual explanations.

Counterfactual explanations are typically used in XAI (Wachter, Mittelstadt, and Russell 2017) and *contestable* AI (Alfrink et al. 2023; Leofante et al. 2024) to indicate paths towards "algorithmic recourse". Counterfactual explanations for QBAFs have not been well-studied. A recent study by (Kampik, Čyras, and Alarcón 2024) focuses on explaining why the strengths' partial order of two topic arguments swap after updating QBAFs (e.g. by adding/removing arguments/edges or changing the base scores of arguments), seeing counterfactual explanations as argument sets whose elements, when updated, cause this swap. For example, suppose there are two topic arguments α and β in a QBAF and the final strength of α is smaller than that of β. If a supporter γ for α were added, resulting in α's strength becoming larger than β's, then the set $\{\gamma\}$ is a possible counterfactual explanation for the strength swap between α and β. Instead, we focus on explaining arguments' strength rather than partial orders between arguments' strengths. Additionally, we focus on counterfactuals as base score functions in structure-fixed QBAFs rather than as argument sets in structure-changeable QBAFs. Another work (Oren et al. 2022), implicitly relates to counterfactual explanations for QBAFs under gradual semantics, by studying the inverse problem of identifying base score functions that can lead to a desired ranking of strengths for all arguments. Differently, we focus on reaching a desired strength for topic arguments instead of a desired ranking of all arguments.

It is also worth mentioning (Sakama 2014), which initially introduced and investigated counterfactual problems in the argumentation area. This work studies what would happen if an initially accepted (rejected) argument were rejected (accepted) (e.g. by adding a new attacker or removing all the attackers towards an argument) and how to explain the corresponding acceptance change of arguments. Differently from our method, this work focuses on abstract argumentation under complete labellings (Dung 1995) and on the consequences of changes rather than their causes as we do.

3 Preliminaries

Formally, a QBAF can be defined as follows.

Definition 1. *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 $\tau : \mathcal{A} \to [0, 1]$.

The base score function in QBAFs ascribes initial strengths (base scores) to arguments therein. QBAFs may be represented graphically (as in Figure 1) using nodes to represent arguments and their base scores and edges to show the relations among arguments. Then QBAFs are said to be *(a)cyclic* if the graphs representing them are (a)cyclic.

Definition 2. *A* gradual semantics σ *is a function that evaluates a QBAF* $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ *by ascribing values* $\sigma(\alpha) \in [0, 1]$ *to every* $\alpha \in \mathcal{A}$ *as their* strength.

Different (gradual) semantics typically ascribe different strengths to arguments. Most semantics define the strength of an argument through an iterative procedure involving two functions: first, an *aggregation function* aggregates the strength of the argument's attackers and supporters; then, an *influence function* combines the aggregation values with the argument's base score to determine its strength. Gradual semantics guarantee convergence for acyclic QBAFs (Potyka

[2] We sometimes refer to counterfactual explanations simply as counterfactuals.

2019). For cyclic QBAFs, the strength values may not converge (Mossakowski and Neuhaus 2018), but when they do, they converge quickly in practice (Potyka 2018). Our focus in this paper is on cases where convergence occurs, as we aim to explain the strength, which is only possible when the strength is defined. Thus, we will assume in the remainder that gradual semantics are *well-defined*, as follows.

Definition 3. *A gradual semantics σ is* well-defined *for a QBAF $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ iff $\sigma(\alpha)$ exists for every $\alpha \in \mathcal{A}$.*

Notation 1. *In the remainder, unless specified otherwise, we use $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau \rangle$ and $\mathcal{Q}_{\tau'} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau' \rangle$ to indicate two QBAFs with different base score functions only, and use σ for any (well-defined) gradual semantics. For $\alpha \in \mathcal{A}$, we let $\sigma_{\tau'}(\alpha)$ denote $\sigma(\alpha)$ in $\mathcal{Q}_{\tau'}$.*

Concretely, we will use the Quadratic Energy (QE) (Potyka 2018), Restricted Euler-based (REB) (Amgoud and Ben-Naim 2018), and Discontinuity-Free Quantitative Argumentation Debates (DF-QuAD) (Rago et al. 2016). To aid understanding, we will show the definition and an example of the DF-QuAD semantics and recap the other two in https://arxiv.org/abs/2407.08497.

In DF-QuAD, for any argument $\alpha \in \mathcal{A}$, $\sigma^{DF}(\alpha)$ is defined as follows:

Aggregation function:

$$E_\alpha = \prod_{\{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^-\}} (1 - \sigma^{DF}(\beta)) - \prod_{\{\beta \in \mathcal{A} | (\beta, \alpha) \in \mathcal{R}^+\}} (1 - \sigma^{DF}(\beta)).$$

Influence function:

$$\sigma^{DF}(\alpha) = \begin{cases} \tau(\alpha) - \tau(\alpha) \cdot |E_\alpha| & \textit{if } E_\alpha \leq 0; \\ \tau(\alpha) + (1 - \tau(\alpha)) \cdot E_\alpha & \textit{if } E_\alpha > 0. \end{cases}$$

An example of applying DF-QuAD is as follows.

Example 1. *Considering the QBAF in Figure 1, where $\tau(\alpha) = 0.5$, $\tau(\beta) = 0.3$, $\tau(\gamma) = 0.6$, $\tau(\rho) = 0.7$, and $\tau(\zeta) = 0.4$. According to the aggregation and influence function of the DF-QuAD semantics*[3]*, we have $E_\alpha = 0.24 - 0.91 = -0.67$, $E_\beta = -0.7$, $E_\gamma = 0.4$, $E_\rho = 0$, $E_\zeta = 0$, $\sigma^{DF}(\alpha) = \tau(\alpha) - \tau(\alpha) \cdot |E_\alpha| = 0.165$, $\sigma^{DF}(\beta) = \tau(\beta) - \tau(\beta) \cdot |E_\beta| = 0.09$, $\sigma^{DF}(\gamma) = \tau(\gamma) + (1 - \tau(\gamma)) \cdot E_\gamma = 0.76$, $\sigma^{DF}(\rho) = \tau(\rho) = 0.7$, $\sigma^{DF}(\zeta) = \tau(\zeta) = 0.4$.*

4 Counterfactuals for QBAFs

We define three counterfactual problems (see Figure 2) and explore the existence of solutions thereto (i.e. *counterfactuals*) and their relationships. Intuitively, given a QBAF, a topic argument and a desired strength therefor, we see counterfactuals as changes to the base score function to obtain the desired strength for the topic argument.

4.1 Strong Counterfactual Problem

Definition 4. *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* strong counterfactual problem *amounts to identifying a base score function $\tau' \neq \tau$ such that $\sigma_{\tau'}(\alpha^*) = s^*$ (in $\mathcal{Q}_{\tau'}$).*

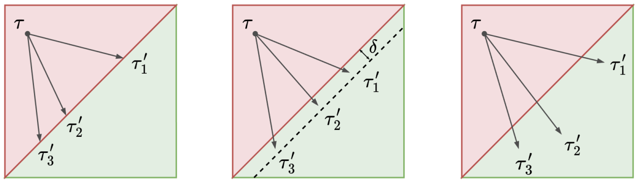

Figure 2: Illustration of strong (left), $\delta-$approximate (middle), and weak (right) counterfactual problems. The squares represent all possible base score functions, with τ the current base score function, and the red (above diagonal) and green (below diagonal) parts as undesirable and desirable alternatives, respectively.

In Figure 2 (left), τ is in the (undesired) red part and $\tau'_1, \tau'_2, \tau'_3$ in the green part are possible counterfactuals (base score functions) as they exactly hit the desired strength.

A trivial solution to the strong counterfactual problem might be just setting the base score of the topic argument to the desired strength and the base score of all others to 0.

Definition 5. *Given a topic argument $\alpha^* \in \mathcal{A}$ and a desired strength s^* for α^*, the* trivial counterfactual *is the base score function $\tau' \neq \tau$ such that $\tau'(\alpha^*) = s^*$ and $\tau'(\alpha) = 0$ for all $\alpha \in \mathcal{A} \setminus \{\alpha^*\}$.*

The trivial counterfactual is a solution to the strong counterfactual problem if the semantics satisfies the following stability property (Amgoud and Ben-Naim 2018).[4]

Definition 6. *A gradual semantics σ satisfies* s-stability *iff for any $\alpha \in \mathcal{A}$, whenever $\sigma(\beta) = 0$ for all $(\beta, \alpha) \in \mathcal{R}^- \cup \mathcal{R}^+$, then $\sigma(\alpha) = \tau(\alpha)$.*

Proposition 1 (Solution Existence)**.** *If σ satisfies s-stability and $\mathcal{Q}$ is acyclic, then the trivial counterfactual is a solution to the strong counterfactual problem.*

Proposition 2. *QE, REB, and DF-QuAD satisfy s-stability.*

Since commonly considered semantics satisfy s-stability, a (trivial) solution to the strong counterfactual problem always exists in acyclic QBAFs. For cyclic QBAFs, the trivial counterfactual is not a solution even in a very simple QBAF with two mutually supporting arguments.

4.2 $\delta-$Approximate Counterfactual Problem

Ensuring that the final strength of a topic argument analytically equals the desired strength may be challenging due to the complexity of the definition of gradual semantics that may involve various linear and nonlinear transformations (e.g., QE, REB and DF-QuAD). Thus, we consider the following relaxation of the strong counterfactual problem.

Definition 7. *Given a constant $\delta > 0$, a topic argument $\alpha^* \in \mathcal{A}$ and a desired strength s^* for α^* such that $\sigma(\alpha^*) \neq s^*$ in $\mathcal{Q}$, the $\delta-$*approximate counterfactual problem *amounts to identifying a base score function $\tau' \neq \tau$ such that*

- *if $\sigma(\alpha^*) < s^*$ in $\mathcal{Q}$, then $s^* \leq \sigma_{\tau'}(\alpha^*) \leq s^* + \delta$ in $\mathcal{Q}_{\tau'}$ and*
- *if $\sigma(\alpha^*) > s^*$ in $\mathcal{Q}$, then $s^* - \delta \leq \sigma_{\tau'}(\alpha^*) \leq s^*$ in $\mathcal{Q}_{\tau'}$.*

[3] We follow the convention that the product of an empty set is 1 (not 0) because 1 is the neutral element with respect to multiplication.

[4] Strictly speaking, the definition in (Amgoud and Ben-Naim 2018) assumes $\mathcal{R}^- = \mathcal{R}^+ = \emptyset$, but if the neutrality property (Amgoud and Ben-Naim 2018) holds, the definitions are equivalent.

As an illustration, in Figure 2 (middle), τ'_1, τ'_2 and τ'_3 are all solutions to the $\delta-$approximate counterfactual problem as they all lie in the δ-interval region of the desired strength.

4.3 Weak Counterfactual Problem

For practical applications, it is also worth further relaxing the $\delta-$approximate counterfactual problem, e.g. when QBAFs are used to solve binary classification problems (Cocarascu, Rago, and Toni 2019; Kotonya and Toni 2019), the topic argument's final strength does not necessarily need to equal or approximate the desired strength. Concretely, in an argumentative movie recommender system of (Cocarascu, Rago, and Toni 2019), as long as a movie's rating (final strength) is higher than some threshold, the movie is considered of good quality. Thus, we next propose the following further relaxation of the strong counterfactual problem.

Definition 8. *Given a topic argument $\alpha^* \in \mathcal{A}$ in $\mathcal{Q}$ and a desired strength s^* for α^* such that $\sigma(\alpha^*) \neq s^*$ in $\mathcal{Q}$, the* weak counterfactual problem *amounts to identifying a base score function $\tau' \neq \tau$ such that*

- *if $\sigma(\alpha^*) < s^*$ in $\mathcal{Q}$, then $\sigma_{\tau'}(\alpha^*) \geq s^*$ in $\mathcal{Q}_{\tau'}$ and*
- *if $\sigma(\alpha^*) > s^*$ in $\mathcal{Q}$, then $\sigma_{\tau'}(\alpha^*) \leq s^*$ in $\mathcal{Q}_{\tau'}$.*

As an illustration, in Figure 2 (right), τ'_1, τ'_2 and τ'_3 are all solution to the weak counterfactual problem as they all cross the threshold (red diagonal line).

4.4 Validity and Problem Relationships

We define notions of validity for counterfactuals and then explore the relationships among these notions.

To be consistent with the literature of counterfactuals in XAI, if a base score function is a solution to a counterfactual problem, we say that this solution is a valid counterfactual.

Definition 9. *A* valid counterfactual for the strong/$\delta-$approximate/weak counterfactual problem *is a base score function τ' which is a solution to the strong/$\delta-$approximate/weak counterfactual problem, respectively.*

To illustrate, in Figure 2, all τ'_1, τ'_2 and τ'_3 are valid counterfactuals as they are solutions to the (respective) problems.

Next, we study relationships among valid counterfactuals.

Proposition 3 (Problem Relationships)**.**

1. *If a counterfactual is valid for the strong counterfactual problem, then it is also valid for the $\delta-$approximate and weak counterfactual problems.*
2. *If a counterfactual is valid for the $\delta-$approximate counterfactual problem, then it is also valid for the weak counterfactual problem.*

Corollary 1. *If σ satisfies s-stability and $\mathcal{Q}$ is acyclic, then the trivial counterfactual is valid for the strong/$\delta-$approximate/weak counterfactual problem.*

However, non-trivial valid counterfactuals do not always exist when not allowing directly setting the base score of the topic argument to the desired strength. For instance, when a topic argument is not connected with any other arguments in the QBAF, then there is no non-trivial counterfactual.

Proposition 4 (Uniqueness)**.** *There is a unique valid counterfactual (for the strong/$\delta-$approximate/weak counterfactual problem) iff the trivial counterfactual is the only valid counterfactual (for the respective problem).*

We leave to future study the identification of special classes of QBAFs for which non-trivial counterfactual explanations can be always guaranteed to exist.

5 Cost-Effective Counterfactuals

In this section we turn to computational challenges: we aim to design an algorithm that can return not only ***valid*** but also ***cost-effective*** counterfactuals for the $\delta-$approximate counterfactual problem. We focus on this problem because, as discussed in Section 4.2, it may be unrealistic to aim at exactly matching the desired strength and solve the strong counterfactual problem and, by Proposition 3, once a valid $\delta-$approximate counterfactual is returned, it is also valid for the weak counterfactual problem. We interpret cost as distance: the shorter the distance, the lower the cost of a counterfactual. This is in line with literature on counterfactual explanations in XAI (Wachter, Mittelstadt, and Russell 2017). To illustrate, in Figure 2 (middle), while τ'_1, τ'_2 and τ'_3 are all valid $\delta-$approximate, their distance to the original base score varies, using the following notion of distance.

Definition 10. *The L_p-*Norm Distance *between τ and τ' is:*

$$d_p(\tau, \tau') = \left(\sum_{\alpha \in \mathcal{A}} ||\tau(\alpha) - \tau'(\alpha)||^p \right)^{\frac{1}{p}}.$$

Minimizing this distance is difficult because there is no closed expression for a topic argument's final strength in general (cyclic) graphs. Here, we will therefore propose an approximate algorithm and evaluate it empirically in Section 7. Before giving the algorithm (in Section 5.2), we will present properties driving its design (Section 5.1 below).

5.1 Algorithm Design Properties

We will design an iterative algorithm that incrementally adapts the weight to find a close counterfactual. To do so, we need to determine an updating ***direction*** and ***magnitude***, which means deciding how much each argument's base score should be increased or decreased. To determine the former, we partition arguments based on their *polarity*. We first define *paths* and *connectivity* between arguments.

Definition 11. *For any $\alpha, \beta \in \mathcal{A}$, we let $p_{\alpha \mapsto \beta} = \langle (\gamma_0, \gamma_1), (\gamma_1, \gamma_2), \cdots, (\gamma_{n-1}, \gamma_n) \rangle (n \geq 1)$ 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 $(\gamma_{i-1}, \gamma_i) \in \mathcal{R}^- \cup \mathcal{R}^+$.*

Notation 2. *We let $|p_{\alpha \mapsto \beta}|$ denote the length of path $p_{\alpha \mapsto \beta}$. We let $P_{\alpha \mapsto \beta}$ and $|P_{\beta \mapsto \alpha}|$ denote the set of all paths from α to β, and the number of paths in $P_{\alpha \mapsto \beta}$, respectively.*

We next distinguish three types of connectivity based on the number of paths from one argument to another.

Definition 12. *For any $\alpha, \beta \in \mathcal{A}$:*

- *β is* disconnected *from α iff $P_{\beta \mapsto \alpha} = \emptyset$;*
- *β is* single-path connected *to α iff $|P_{\beta \mapsto \alpha}| = 1$;*

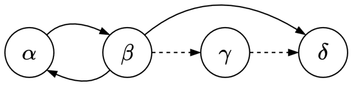

Figure 3: An example QBAF (base scores omitted).

- β *is* multi-path connected *to* α *iff* $|P_{\beta\mapsto\alpha}| > 1$.

Example 2. *Consider the QBAF in Figure 3. γ is disconnected from β because there is no path from γ to β; γ is single-path connected to δ via path $p_{\gamma\mapsto\delta} = \langle(\gamma,\delta)\rangle$; α is multi-path connected to β, because there are infinitely many paths from α to β, namely $p'_{\alpha\mapsto\beta} = \langle(\alpha,\beta)\rangle$, $p''_{\alpha\mapsto\beta} = \langle(\alpha,\beta),(\beta,\alpha),(\alpha,\beta)\rangle$ and so on.*

Inspired by (Rago, Li, and Toni 2023; Yin, Potyka, and Toni 2023), we define polarity to characterize the influence between arguments according to their connectivity and the number of attacks on paths between them.

Definition 13. *The* polarity *from β to α ($\alpha,\beta\in\mathcal{A}$, $\alpha\neq\beta$) is:*

- neutral *(β is* neutral *to α) iff β is disconnected from α;*
- positive *(β is* positive *to α) iff $P_{\alpha\mapsto\beta} \neq \emptyset$ and for every path $p_{\alpha\mapsto\beta} \in P_{\beta\mapsto\alpha}$, $|p_{\beta\mapsto\alpha} \cap \mathcal{R}^-|$ is even;*
- negative *(β is* negative *to α) iff $P_{\beta\mapsto\alpha} \neq \emptyset$ and for every path $p_{\beta\mapsto\alpha} \in P_{\beta\mapsto\alpha}$, $|p_{\beta\mapsto\alpha} \cap \mathcal{R}^-|$ is odd;*
- unknown *(β is* unknown *to α) iff $\exists p'_{\beta\mapsto\alpha} \in P_{\beta\mapsto\alpha}$ such that $\left|p'_{\beta\mapsto\alpha}\cap\mathcal{R}^-\right|$ is even and $\exists p''_{\beta\mapsto\alpha} \in P_{\beta\mapsto\alpha}$ such that $\left|p''_{\beta\mapsto\alpha}\cap\mathcal{R}^-\right|$ is odd.*

So, if β is single-path connected to α, then it is positive or negative to α; if β is multi-path connected to α, unless it is positive or negative to α on every path, it is unknown to α.

Example 3. *In Figure 3, γ is neutral to β; γ is positive to δ as there is only one path $p_{\gamma\mapsto\delta}$ with 0 (even number) attacks; α is negative to β since, although there are infinitely many paths from α to β, each path has an odd number of attacks; β is unknown to δ as for one path $p'_{\beta\mapsto\delta} = \langle(\beta,\delta)\rangle$, it has 1 attack, while for another path $p''_{\beta\mapsto\delta} = \langle(\beta,\gamma),(\gamma,\delta)\rangle$, it has 0 attacks. Hence, β is unknown to δ.*

We can restrict the update direction of arguments based on their polarity assuming that σ respects *directionality* and *monotonicity*. Directionality (Amgoud and Ben-Naim 2016) states that the strength of an argument depends only on its incoming edges.

Definition 14. *A semantics σ satisfies* directionality *iff, for any $\mathcal{Q}$ and $\mathcal{Q}'=\langle\mathcal{A}',\mathcal{R}^{-'},\mathcal{R}^{+'},\tau'\rangle$ such that $\mathcal{A}=\mathcal{A}'$, $\mathcal{R}^-\subseteq\mathcal{R}^{-'}$, and $\mathcal{R}^+\subseteq\mathcal{R}^{+'}$, 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 $\mathcal{R}^{-'}\cup\mathcal{R}^{+'} = \mathcal{R}^-\cup\mathcal{R}^+\cup\{(\alpha,\beta)\}$ and $P_{\beta\mapsto\gamma}=\emptyset$, then $\sigma(\gamma)=\sigma_{\mathcal{Q}'}(\gamma)$.*

Proposition 5. *If a semantics σ satisfies directionality, then for any $\alpha,\beta\in\mathcal{A}$ such that β is neutral to α and for any τ' such that $\tau'(\beta)$ is an arbitrary value from $[0,1]$ and $\tau'(\gamma)=\tau(\gamma)$ for all $\gamma\in\mathcal{A}\setminus\{\beta\}$, $\sigma_{\tau'}(\alpha)=\sigma(\alpha)$ always holds.*

Proposition 6. *QE, REB, DF-QuAD satisfy directionality.*

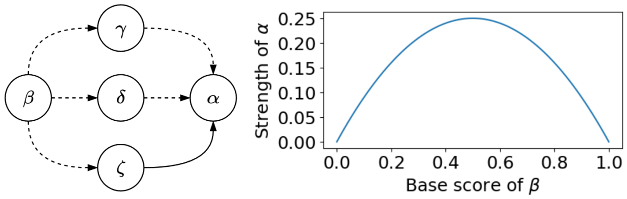

Figure 4: A QBAF evaluated by the DF-QuAD semantics (inspired by (Kampik et al. 2024)).

For semantics satisfying directionality, by Proposition 5, we could just maintain the base scores of **neutral** arguments (wrt. the topic argument) when identifying counterfactuals.

Monotonicity states that increasing (decreasing) the base score of a single-path connected supporter (attacker) will not decrease (increase) the final strength of the topic argument.

Definition 15. *A semantics σ satisfies* monotonicity *iff for any $\alpha,\beta\in\mathcal{A}$ such that β is single-path connected to α, for any τ_1,τ_2 with $\tau_1(\beta)\leq\tau_2(\beta)$ and $\tau_1(\gamma)=\tau_2(\gamma)$ for all $\gamma\in\mathcal{A}\setminus\{\beta\}$:*

- *if $(\beta,\alpha)\in\mathcal{R}^-$, then $\sigma_{\tau_1}(\alpha)\geq\sigma_{\tau_2}(\alpha)$;*
- *if $(\beta,\alpha)\in\mathcal{R}^+$, then $\sigma_{\tau_1}(\alpha)\leq\sigma_{\tau_2}(\alpha)$.*

Proposition 7. *If a semantics σ satisfies monotonicity, then for any τ_1,τ_2 such that $\tau_1(\beta)\leq\tau_2(\beta)$ and $\tau_1(\gamma)=\tau_2(\gamma)$ for all $\gamma\in\mathcal{A}\setminus\{\beta\}$:*

- *if β is negative to α, then $\sigma_{\tau_1}(\alpha)\geq\sigma_{\tau_2}(\alpha)$;*
- *if β is positive to α, then $\sigma_{\tau_1}(\alpha)\leq\sigma_{\tau_2}(\alpha)$.*

Proposition 8. *QE, REB, DF-QuAD satisfy monotonicity.*

For semantics satisfying monotonicity, by Prop. 7, we can increase a topic argument's the strength by increasing (decreasing) the base scores of **positive** (**negative**) arguments.

We now consider unknown arguments. Since their base scores may not be globally monotonic to the strength of a topic argument, we cannot simply decide an invariant updating direction for their base scores. For example, in Figure 4 where the base scores of all arguments are 0, with the increase of β's base score from 0 to 1 while all others remain the same, the final strength of α displays a non-monotonic effect, increasing initially and then decreasing. To overcome this challenge, we introduce the *difference quotient* from one argument to another, enabling us to capture the average changing rate of the strength wrt. the base score within an interval, approximately reflecting local monotonicity at the current base score. Then, the difference quotient can act as indicator for the updating direction of **unknown** arguments.

Definition 16. *For any $\alpha,\beta\in\mathcal{A}$, given a constant $h\in[-1,0)\cup(0,1]$ and a base score function τ' such that $\tau'(\beta)=\tau(\beta)+h\in[0,1]$ and $\tau'(\gamma)=\tau(\gamma)$ for all $\gamma\in\mathcal{A}\setminus\{\beta\}$, the* difference quotient *from β to α is:*

$$\Delta^h_{\beta\mapsto\alpha} = \frac{\sigma_{\tau'}(\alpha)-\sigma(\alpha)}{h}.$$

As an illustration, if the difference quotient from an unknown argument to the topic argument is greater (less) than 0, we increase (decrease) its base score when the desired strength is higher than the current. However, if the difference quotient is exactly equal to 0, we do not update. To highlight the compatibility of the difference quotient with our previous ideas, let us note that neutral arguments always have 0 difference quotient, while positive (negative) arguments always have positive (negative) quotients.

Proposition 9 (Sign Invariance)**.** *For any $\alpha, \beta \in \mathcal{A}$, for any $h \in [-1, 0) \cup (0, 1]$, if a semantics σ satisfies directionality and monotonicity, then*

- *if β is neutral to α, $\Delta^h_{\beta \mapsto \alpha} = 0$;*
- *if β is positive to α, $\Delta^h_{\beta \mapsto \alpha} \geq 0$;*
- *if β is negative to α, $\Delta^h_{\beta \mapsto \alpha} \leq 0$.*

After determining the updating directions for all four types of arguments, we consider updating the magnitude. To do so, we use a *priority* strategy that assigns higher updating magnitude to arguments closer to the topic argument in terms of path length. We expect that an appropriate priority can help identify cost-effective counterfactuals in terms of L_p-norm distance, under the hypothesis that updating closer arguments is more efficient in achieving the desired strength. We empirically verify this hypothesis in Section 7.

We define *priority* as the reciprocal of the shortest path length from one argument to another, which is at most 1 for the attackers or supporters.

Definition 17. *For any $\alpha, \beta \in \mathcal{A}$, $\alpha \neq \beta$:*

- *if β is disconnected from α, then the* priority *from β to α is* 0*;*
- *if β is single-path or multi-path connected to α via $n(n \geq 1)$ paths $p^1_{\beta \mapsto \alpha}, \cdots, p^n_{\beta \mapsto \alpha} \in P_{\beta \mapsto \alpha}$, then the* priority *from β to α is $1/min\{|p_{\beta \mapsto \alpha}| \mid p_{\beta \mapsto \alpha} \in P_{\beta \mapsto \alpha}\}$.*

Example 4. *Consider the QBAF in Figure 3. Let us first consider the priority from α to δ. There are infinitely many paths from α to δ. Among them, path $p_{\alpha \mapsto \delta} = \langle (\alpha, \beta), (\beta, \delta) \rangle$ has the minimal length 2, thus, the priority from α to δ is* 0.5*. Since the minimal length of paths from β to δ and from γ to δ are both* 1*, the priorities are both* 1.

5.2 Algorithms

Algorithm 1 computes the polarity from α to β ($\alpha \neq \beta$) in three steps. Firstly, we compute all the non-cyclic paths from α to β with *Depth-First Search (DFS)* and store them in a set of paths called $P_{\alpha \mapsto \beta}$. If there is no path, then the polarity is neutral. Secondly, all the nodes in all the paths are checked for cycles with the function *find_cycles* [5], which outputs a set of elementary cycles. If any of the node is part of a cycle, then we check whether each cycle contains an odd number of attacks. If this is the case, then the polarity from α to β is unknown because the cycle will contain both odd and even numbers of attacks by going through the cycle an odd or even number of times. For instance, suppose a QBAF consists of two arguments α and β, where α attacks β, and β supports α. In this case, path $\langle (\alpha, \beta) \rangle$ is negative while $\langle (\alpha, \beta), (\beta, \alpha), (\alpha, \beta) \rangle$ is positive, thus α is unknown to β. Finally, we check the number of attacks on every path: if all the paths contain an odd (even) number of attacks, α is negative (positive) to β, and it is unknown otherwise.

Algorithm 1 Polarity Computation Algorithm

```
Input: A QBAF $\mathcal{Q}$, two arguments $\alpha, \beta \in \mathcal{A}$
Output: The polarity from $\alpha$ to $\beta$
$P_{\alpha \mapsto \beta} \leftarrow DFS(\mathcal{Q}, \alpha, \beta)$
if $P_{\alpha \mapsto \beta} == \emptyset$ then
    return $-2$ //neutral
for $p_{\alpha \mapsto \beta}$ in $P_{\alpha \mapsto \beta}$ do
    for $node$ in $p_{\alpha \mapsto \beta}$ do
        $cycles \leftarrow find\_cycles(\mathcal{Q}, node)$
        for $cycle$ in $cycles$ do
            if $cycle$ contains odd number attacks then
                return $0$ // unknown
$polarity \leftarrow [\,]$ // empty array
for $p_{\alpha \mapsto \beta}$ in $P_{\alpha \mapsto \beta}$ do
    if $p_{\alpha \mapsto \beta}$ contains odd number attacks then
        $polarity.append(-1)$ // negative
    else if $p_{\alpha \mapsto \beta}$ contains even number attacks then
        $polarity.append(1)$ // positive
if all items in $polarity == -1$ then
    return $-1$ // negative
else if all items in $polarity == 1$ then
    return $1$ // positive
else
    return $0$ // unknown
```

Algorithm 2 Priority Computation Algorithm

```
Input: A QBAF $\mathcal{Q}$, two arguments $\alpha, \beta \in \mathcal{A}$, a constant $c$
Output: The priority from $\alpha$ to $\beta$
if $\alpha == \beta$ then
    return $c$
$P_{\alpha \mapsto \beta} \leftarrow DFS(\mathcal{Q}, \alpha, \beta)$
return $1/min\{length(p_{\alpha \mapsto \beta}) \mid p_{\alpha \mapsto \beta} \in P_{\alpha \mapsto \beta}\}$
```

Algorithm 2 computes the priority from α to β. We define the self-priority of an argument as a constant greater than 1. Thus an argument has the highest priority to itself than any others. Next, we perform a DFS to compute all the non-cyclic paths from α to β and return the reciprocal of the shortest path length.

Algorithm 3 (which we call ***CE-QArg*** for Counterfactual Explanations for Quantitative bipolar Argumentation frameworks) is an iterative updating algorithm for identifying valid and cost-effective δ-approximate counterfactuals. CE-QArg essentially involves determining the direction and magnitude of arguments. For brevity, we assume the current strength of the topic argument is less than the desired one. Firstly, we compute the polarity and priority for each argument with the function *func_polarity* and *func_priority* from Algorithm 1 and 2, respectively. Secondly, the key part of

[5] See https://github.com/XiangYin2021/CE-QArg for details. A quicker implementation could involve applying (Johnson 1975), which is able to find all the elementary cycles of a directed graph in time bounded by O((n+e)(c+1)), where n, e, and c are the number of nodes, edges, and elementary cycles (nodes occur once except for the starting node) in the graph, respectively.

Algorithm 3 CE-QArg

Input: A QBAF $\mathcal{Q}$, a gradual semantics σ, a topic argument α^* and a desired strength s^* for α^*
Parameter: An updating step ε, a change h
Output: A counterfactual τ^*

```
 1: $update, polarity, priority \leftarrow \{\}, \{\}, \{\}$ //dictionaries
 2: for $\alpha$ in $\mathcal{A}$ do
 3:     $polarity[\alpha] \leftarrow func\_polarity(\alpha, \alpha^*)$
 4:     $priority[\alpha] \leftarrow func\_priority(\alpha, \alpha^*)$
 5:     if $polarity[\alpha] == -2$ then
 6:         $update[\alpha] \leftarrow 0$ // neutral
 7:     else if $polarity[\alpha] == 1$ then
 8:         $update[\alpha] \leftarrow 1$ // positive
 9:     else if $polarity[\alpha] == -1$ then
10:         $update[\alpha] \leftarrow -1$ // negative
11: while $\sigma(\alpha^*) < s^*$ do
12:     for $\alpha$ in $\mathcal{A}$ do
13:         if $polarity[\alpha] == 0$ then
14:             $dquo[\alpha] \leftarrow func\_dquo(\alpha, \alpha*, h)$
15:             if $dquo[\alpha] > 0$ then
16:                 $update[\alpha] \leftarrow 1$
17:             else if $dquo[\alpha] < 0$ then
18:                 $update[\alpha] \leftarrow -1$
19:             else
20:                 $update[\alpha] \leftarrow 0$
21:     for $\alpha$ in $\mathcal{A}$ do
22:         $\tau^*(\alpha) \leftarrow max(0, min(1, \tau^*(\alpha) + update[\alpha] \cdot \varepsilon \cdot$
            $priority[\alpha]))$
23:     compute $\sigma(\alpha^*)$
24: return $\tau^*$
```

this algorithm is to determine the updating direction in every updating iteration. For this, we need an *update* list to record the updating direction in each iteration. For positive, negative and neutral arguments (lines 5-10), we only need to identify the updating direction once since they remain invariant, whereas, for unknown arguments (lines 13-20), we need to compute their updating direction in every iteration by difference quotient using the function *func_dquo*, which can be intuitively implemented by Definition 16 (so we omit the details). Once the updating direction of every argument is determined in an iteration, we update the base scores of all arguments all in one go by a small step multiplied by their priority (line 22) and make sure they are within the bounds. We assume the step is small enough that the interactions among arguments can be neglected. We repeat this procedure iteratively until the current strength reaches the desired strength, after which we return a possible counterfactual (lines 11-24).

6 Formal Properties for Explanations

We study the properties for counterfactuals, with a focus on the neutral, negative, and positive arguments. Here, we assume the gradual semantics considered satisfy both directionality and monotonicity. Note that the properties of explanations are distinct from those of semantics, despite the satisfaction of the former being dependent on the latter.

Existence (Čyras, Kampik, and Weng 2022; Kampik et al. 2024) is a commonly considered property for explanations. It says that if the strength of an argument differs from its base score, then there must exist an argument that caused the change. *Alteration Existence* states that if a valid counterfactual increases the strength of the topic argument, then there must exist a positive (negative) argument whose base score is also increased (decreased) in the counterfactual whenever the QBAF does not have any unknown arguments.

Proposition 10 (Alteration Existence)**.** *Given a valid counterfactual τ^* (for the strong/$\delta-$approximate/weak counterfactual problem), a topic argument $\alpha^* \in \mathcal{A}$ such that $\tau^*(\alpha^*) = \tau(\alpha^*)$, and $\nexists \beta \in \mathcal{A}$ ($\beta \neq \alpha^*$) such that β is unknown to α^*:*

1. *If $\sigma(\alpha^*) \neq \sigma_{\tau^*}(\alpha^*)$, then $\exists \gamma \in \mathcal{A}$ such that $\tau(\gamma) \neq \tau^*(\gamma)$;*
2. *If $\sigma(\alpha^*) < \sigma_{\tau^*}(\alpha^*)$, then either $\exists \gamma \in \mathcal{A}$ such that γ is positive to α^* and $\tau(\gamma) \leq \tau^*(\gamma)$ or $\exists \gamma \in \mathcal{A}$ such that γ is negative to α^* and $\tau(\gamma) \geq \tau^*(\gamma)$;*
3. *If $\sigma(\alpha^*) > \sigma_{\tau^*}(\alpha^*)$, then either $\exists \gamma \in \mathcal{A}$ such that γ is positive to α^* and $\tau(\gamma) \geq \tau^*(\gamma)$ or $\exists \gamma \in \mathcal{A}$ such that γ is negative to α^* and $\tau(\gamma) \leq \tau^*(\gamma)$.*

Given that validity is fundamental for counterfactuals, we introduce two properties associated with validity.

For attribution-based explanations for QBAFs, it is interesting to explore the effects of removing or nullifying an argument (by setting its base score to 0 – e.g. see removal-based contribution functions in (Kampik et al. 2024) and *agreement* in (Yin, Potyka, and Toni 2023)). For counterfactual explanations, it is essential to consider the validity of a counterfactual if it is perturbed. Combining both ideas, we propose a novel property called *nullified-validity*, which examines whether a valid counterfactual remains valid even after nullifying an argument. For example, nullifying a positive argument in a valid counterfactual could still result in a valid counterfactual if a smaller strength is expected for the topic argument in the weak counterfactual problem.

Proposition 11 (Nullified-Validity)**.** *Given a valid counterfactual τ^* (for the strong/$\delta-$approximate/weak counterfactual problem), a topic argument $\alpha^* \in \mathcal{A}$ and a desired strength s^* for α^*, and another base score function τ^0 such that for some $\beta \in \mathcal{A}$ ($\beta \neq \alpha^*$), $\tau^0(\beta) = 0$ and $\tau^0(\gamma) = \tau^*(\gamma)$ for all $\gamma \in \mathcal{A} \setminus \{\beta\}$:*

1. *If β is neutral to α^*, then τ^0 is still valid for the strong/$\delta-$approximate/weak counterfactual problem;*
2. *If β is negative to α^* and $\sigma_{\tau^*}(\alpha^*) \geq s^*$, then τ^0 is still valid for the weak counterfactual problem;*
3. *If β is positive to α^* and $\sigma_{\tau^*}(\alpha^*) \leq s^*$, then τ^0 is still valid for the weak counterfactual problem.*

In attribution-based explanations for QBAFs, it is also interesting to compare two related explanations and study their properties (e.g., see *monotonicity* in (Yin, Potyka, and Toni 2023) and *dominance* in (Amgoud, Ben-Naim, and Vesic 2017; Yin, Potyka, and Toni 2024)). We extend this idea to counterfactuals by focusing on the validity property. *Related-validity* identifies a valid counterfactual by comparing it with another already valid counterfactual without recomputing the strengths of arguments. To illustrate, suppose we have a valid counterfactual for the weak counterfactual problem, where the strength of the topic argument is as small as desirable. Then we can compare it with another counterfactual: if the latter counterfactual has a smaller base score in a positive argument while keeping all other base scores the same, it is still considered valid.

Proposition 12 (Related-Validity)**.** *Given a valid counterfactual explanation τ^* (for the weak counterfactual problem), a topic argument $\alpha^* \in \mathcal{A}$ and a desired strength s^* for α^* such that $\sigma_{\tau^*}(\alpha^*) \geq s^*$, for every other base score function τ' in $\mathcal{Q}_{\tau'}$, and all $\beta \in \mathcal{A}$ ($\beta \neq \alpha^*$):*

1. *If $\beta \in \mathcal{A}$ is neutral to α^*, $\tau'(\beta) \geq \tau^*(\beta)$ and $\tau'(\gamma) = \tau^*(\gamma)$ for all $\gamma \in \mathcal{A} \setminus \{\beta\}$, then τ' is also valid (for the weak counterfactual problem);*
2. *If $\beta \in \mathcal{A}$ is negative to α^*, $\tau'(\beta) \leq \tau^*(\beta)$ and $\tau'(\gamma) = \tau^*(\gamma)$ for all $\gamma \in \mathcal{A} \setminus \{\beta\}$, then τ' is also a valid counterfactual (for the weak counterfactual problem);*
3. *If $\beta \in \mathcal{A}$ is positive to α^*, $\tau'(\beta) \geq \tau^*(\beta)$ and $\tau'(\gamma) = \tau^*(\gamma)$ for all $\gamma \in \mathcal{A} \setminus \{\beta\}$, then τ' is also a valid counterfactual (for the weak counterfactual problem).*

7 Evaluations

We show effectiveness (Experiment 1), scalability (Experiment 2) and robustness (Experiment 3) of CE-QArg. In this section, we focus on the $\delta-$approximate counterfactual problem and valid explanations therefor with $\delta = 0.1$.

Settings We conducted experiments separately using acyclic and cyclic QBAFs. For **acyclic** QBAFs, we generated tree-like QBAFs as they occur in many applications of QBAFs (e.g. see (Kotonya and Toni 2019; Cocarascu, Rago, and Toni 2019; Chi et al. 2021)). We created full binary, ternary, and quaternary trees with different widths (2, 3, and 4) and depths (from 1 to 8) where each edge was randomly set to an attack or a support. The topic argument was set as the root of the tree. To improve the credibility of the results and reduce the impact of random errors, we randomly created each tree-like QBAF with a specified width and depth 100 times with different base scores in $[0, 1]$ over different arguments. For **cyclic** QBAFs, we created varying numbers of arguments (from 100, 200, to 1000), each repeated 100 times with random attacks or supports, random base scores in $[0, 1]$, and a randomly designated topic argument. The argument-relation ratio was set as 1:1 to avoid dense QBAFs which can impact the explainability because of their high structural complexity, making them difficult to comprehend and thus less suitable for explainability.

We report on experiments with the QE semantics[6]. We set the updating step ε to 0.01 in each iteration. Besides, we used both L_1 and L_2-norm distance as the metric for cost.

7.1 Experiment 1: Effectiveness

We show the effectiveness of CE-QArg by conducting ablation studies on polarity and priority. We first propose the Baseline method (BL) based on Proposition 9, which directly computes the difference quotient for all arguments as the updating indicators without considering their polarity and priority. We then separately applied priority or polarity on the BL to show their individual efficacy (denoted as BL+pri and BL+pol). Finally, we showed the performance of our CE-QArg which incorporates both polarity and priority. We evaluated CE-QArg on validity, L_1, L_2-norm distance, and runtime.

	Validity	L_1	L_2	Runtime (s)
BL	**1.00**/0.78	16.52/17.75	1.65/1.87	1.52/79.57
BL+pri	**1.00**/0.93	**4.04/0.51**	**0.49/0.30**	1.16/32.30
BL+pol	**1.00/1.00**	22.03/1.42	2.05/0.46	0.02 /1.58
Ours	**1.00/1.00**	4.33/0.54	0.50/**0.30**	**0.01/0.80**

Table 1: Ablation studies for polarity and priority on **acyclic/cyclic** QBAFs: Comparison of average validity, L_1, L_2-norm distance, and runtime over 100 random generated acyclic full binary tree-like QBAFs with a depth of 7 and cyclic QBAFs with 100 arguments and 100 relations.[7](The best results are shown in bold.)

We first discuss the results of acyclic QBAFs (left side of slash in Table 1). All methods had the best validity of 1.00. The use of priority is expected to shorten the L_1 and L_2-norm distance by updating the arguments close to the topic argument with a larger step. The results of BL and BL+pri showed that applying priority resulted in a 75.5% and 70.3% decrease in L_1 and L_2-norm distance, respectively, which also reduced the runtime by 23.7% in the meanwhile. These findings are in line with our hypothesis in that priority enables reaching a desired strength more cost-effectively, making the distance shorter and thus requiring fewer iterations. Utilizing polarity is expected to lower the runtime by computing the polarity only once for neutral, positive, and negative arguments in a QBAF. The results of BL and BL+pol showed that applying polarity significantly decreased the runtime by 98.7% as expected because the BL computes the difference quotient in every iteration for every argument, which causes the runtime wastage. For our CEQArg, we see that the runtime is better than that of any of the previous three algorithms. Compared to BL+pri, although the L_1 and L_2-norm distance is slightly longer in our CE-QArg, the runtime is substantially decreased.

The results of cyclic QBAFs (right side of the slash) are similar. We can observe that using priority alone decreased the L_1 and L_2-norm distance by 97.1% and 84.0%, respectively; and solely applying polarity significantly decreased the runtime by 98.0%. Finally, applying both priority and polarity can yield the desired counterfactual explanations in terms of the L_1 and L_2-norm distance and shorten the algorithm runtime. However, it is interesting to note that the validity was violated without applying polarity (BL and BL+pri). This is probably because of the computation error of arguments' strength, especially for cyclic QBAFs where an analytical value does not generally exist. Then, the strength error will cause the wrong difference quotient

[6] We give results with DF-QuAD and REB in https://arxiv.org/abs/2407.08497.

[7] We show the effectiveness on larger QBAFs in https://arxiv.org/abs/2407.08497.

thus the wrong updating direction. As a result, the current strength just oscillates around the desired strength. The invalidity of counterfactuals also explains why the L_1 and L_2-norm distance performs better than others in BL+pri.

Overall, the ablation studies show the effectiveness of priority and polarity on both acyclic and cyclic QBAFs, which allows finding valid and cost-effective counterfactuals.

7.2 Experiment 2: Scalability

We evaluated the scalability of CE-QArg on QBAFs of varying sizes. We show both validity and the runtime performance. First, all tested acyclic and cyclic QBAFs achieved a validity score of 1.00. We next present the average runtime for acyclic and cyclic QBAFs with different sizes in Figure 5.

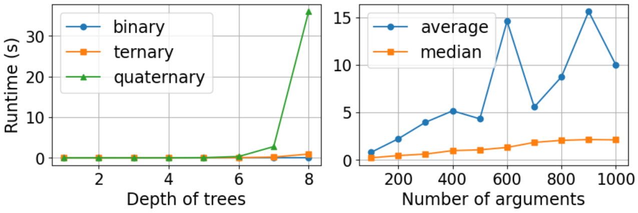

Figure 5: Scalability evaluation for CE-QArg on acyclic (left) and cyclic (right) QBAFs: comparison of average runtime over 100 randomly generated acyclic and cyclic QBAFs.

Figure 5 (left) shows the runtime for binary, ternary, and quaternary tree-like QBAFs with depths ranging from 1 to 8. Notably, binary and ternary trees have a runtime less than 1s across all depths. However, the runtime for quaternary trees increases sharply from depth 7 (2.76s) to depth 8 (36.00s), which is expected since the number of arguments increases substantially from $(4^7 - 1)/3 = 5,461$ to $(4^8 - 1)/3 = 21,845$. In Figure 5 (right), we observe the average runtime (blue line) increases as the number of arguments in QBAFs increases. However, there are two outliers at the number of 600 and 900 which may be due to the randomness of the QBAF generation. Therefore, we added the median runtime, which shows a stable rising trend as expected. Note that the runtime line plot can vary dramatically with different densities of QBAFs because CE-QArg involves the traverse of the QBAF using DFS when computing the polarity and priority. This could be improved by adding pruning strategies while traversing to reduce the runtime cost: we leave this improvement to future work. In summary, both acyclic and cyclic QBAFs exhibit reasonable runtime performance, demonstrating good scalability of CE-QArg.

7.3 Experiment 3: Robustness

Robustness is a crucial and commonly considered metric for counterfactual explanation methods (Artelt et al. 2021; Jiang et al. 2023a; Jiang et al. 2023b; Jiang et al. 2024). An explanation method is robust against input perturbations if similar inputs leading to the same outputs give rise to similar explanations. For instance, if two loan applicants with similar conditions are rejected, they should obtain similar counterfactuals. However, non-robust methods may generate completely different counterfactuals for similar rejected applicants (see an example in Figure 1(b) of (Slack et al. 2021)), which is unfair as they have different updating costs to obtain the desired decision. As an illustration, the counterfactual explanation method in Figure 6 (left) is more robust against input perturbations than that of Figure 6 (right) as the new counterfactual is still close to the previous counterfactual after the input is perturbed. We propose a *robustness against input perturbations* metric which uses the L_p-norm distance to evaluate the robustness of CE-QArg under the perturbation of the input base score function.

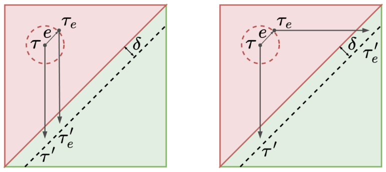

Figure 6: Comparison of robust (left) and non-robust (right) counterfactual explanation methods against input perturbations.

Metric 1. *Given a perturbation score $e > 0$, two base score functions τ and $\tau_e = \tau(\alpha) + e$ for all $\alpha \in \mathcal{A}$, and two counterfactual explanation τ' for τ and τ'_e for τ_e, the* robustness against input perturbations *is measured by $d_p(\tau', \tau'_e)$.*

In addition, inspired by (Leofante and Lomuscio 2023; Pawelczyk et al. 2022), we propose a *robustness against noisy execution* metric, which requires that the generated counterfactual could still lead to similar output (final strength for the topic argument) even if the counterfactual is perturbed. Still taking the loan application as an example, robustness against noisy execution ensures that when a rejected applicant is very close to the provided counterfactual, then the output of this counterfactual should also be close to the desired final strength. We evaluate robustness against noisy execution by the absolute difference between the strength of a topic argument obtained with two similar counterfactuals.

Metric 2. *Given a topic argument $\alpha^* \in \mathcal{A}$, a perturbation score e, and two counterfactual explanations τ' and $\tau'_e = \tau'(\alpha) + e$ for all $\alpha \in \mathcal{A}$, the* robustness against noisy execution *is measured by $|\sigma_{\tau'_e}(\alpha^*) - \sigma_{\tau'}(\alpha^*)|$.*

We applied Metrics 1 and 2 with an increasing perturbation e from 10^{-8} to 10^{-1} over acyclic and cyclic QBAFs. Figure 7 (left) shows the robustness against input perturbations through the average explanation difference measured by Metric 1, while Figure 7 (right) shows the robustness against noisy execution through the average strength difference of the topic argument measured by Metric 2. With the increase of e, the explanation difference and strength difference both showed an approximate linear and stable increasing trend, albeit in Figure 7 (right), we observed a bit of unsteadiness at the beginning when $e = 10^{-8}$ and $e = 10^{-7}$. Overall, CE-QArg exhibited robustness against input perturbations and noisy execution on both cyclic and acyclic QBAFs.

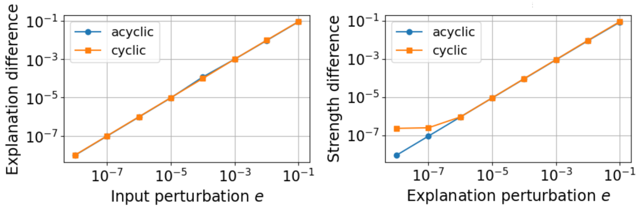

Figure 7: Robustness evaluation of perturbing base score functions (left) and perturbing generated counterfactuals (right) over 100 random generated QBAFs with increasing perturbation e from 10^{-8} to 10^{-1}. Acyclic QBAFs are binary trees with a depth of 7 while cyclic QBAFs contain 100 arguments and relations.[8]

8 Conclusions

We formally defined three counterfactual problem variants and discussed their relationships. We proposed an iterative algorithm CE-QArg to identify valid and cost-effective counterfactuals. We discussed some formal properties of our counterfactual explanations and empirically evaluate CE-QArg on random generated QBAFs. Experimental results show that CE-QArg has a desirable performance on effectiveness, scalability, and robustness. While the identification of valid and cost-effectiveness counterfactuals still lacks sufficient theoretical guarantees because there is no closed expression for a topic argument's final strength in general (cyclic) graphs, we improved the searching process by applying polarity and priority so that the L_1 and L_2-norm distance and the runtime decreased significantly compared to the baseline (BL) method.

There are a few avenues for future work. First, it would be interesting to explore identifying counterfactuals in QBAFs when their structure can be changed by adding or removing arguments or edges. Second, it would be worth exploring identifying counterfactuals for multiple topic arguments simultaneously. However, this would be challenging as it involves the interactive effect among topic arguments. Third, it would also be interesting to explore the relationship between argument attribution explanations and counterfactual explanations (Kommiya Mothilal et al. 2021). Finally, it would be important to carry out case studies and user experiments as explanations should finally help humans understand and make better decisions.

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

Adadi, A., and Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access* 6:52138–52160.

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

Amgoud, L., and Ben-Naim, J. 2016. Evaluation of arguments from support relations: Axioms and semantics. In *25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, 900–906.

Amgoud, L., and Ben-Naim, J. 2018. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning* 99:39–55.

Amgoud, L.; Ben-Naim, J.; and Vesic, S. 2017. Measuring the intensity of attacks in argumentation graphs with shapley value. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 63–69.

Artelt, A.; Vaquet, V.; Velioglu, R.; Hinder, F.; Brinkrolf, J.; Schilling, M.; and Hammer, B. 2021. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–09. IEEE.

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.

Byrne, R. M. 2019. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, 6276–6282.

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 Intelligent Systems* 36:52–63.

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. *arXiv preprint arXiv:2105.11266*.

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.

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(2):321–357.

Jiang, J.; Lan, J.; Leofante, F.; Rago, A.; and Toni, F. 2023a. Provably robust and plausible counterfactual explanations

[8] We show the robustness on larger QBAFs in https://arxiv.org/abs/2407.08497.

for neural networks via robust optimisation. In Yanikoglu, B., and Buntine, W. L., eds., *Asian Conference on Machine Learning, ACML 2023, 11-14 November 2023, Istanbul, Turkey*, volume 222 of *Proceedings of Machine Learning Research*, 582–597. PMLR.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2023b. Formalising the robustness of counterfactual explanations for neural networks. 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*, 14901–14909. AAAI Press.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024. Robust counterfactual explanations in machine learning: A survey. *CoRR* abs/2402.01928.

Johnson, D. B. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing* 4(1):77–84.

Kampik, T., and Čyras, K. 2022. Explaining change in quantitative bipolar argumentation. In *9th International Conference on Computational Models of Argument, COMMA 2022, Cardiff, Wales, September 14-16 2022.*, volume 353, 188–199. IOS Press.

Kampik, T.; Potyka, N.; Yin, X.; Čyras, K.; and Toni, F. 2024. Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. *arXiv preprint arXiv:2401.08879.*

Kampik, T.; Čyras, K.; and Alarcón, J. R. 2024. Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations. *International Journal of Approximate Reasoning* 164:109066.

Kommiya Mothilal, R.; Mahajan, D.; Tan, C.; and Sharma, A. 2021. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 652–663.

Kotonya, N., and Toni, F. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Workshop on Argument Mining*, 156–166.

Leite, J., and Martins, J. 2011. Social abstract argumentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2287–2292.

Leofante, F., and Lomuscio, A. 2023. Towards robust contrastive explanations for human-neural multi-agent systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2343–2345.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; et al. 2024. Contestable ai needs computational argumentation. *arXiv preprint arXiv:2405.10729.*

Mossakowski, T., and Neuhaus, F. 2018. Modular semantics and characteristics for bipolar weighted argumentation graphs. *arXiv preprint arXiv:1807.06685.*

Oren, N.; Yun, B.; Vesic, S.; and Baptista, M. S. 2022. Inverse problems for gradual semantics. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 2719–2725. ijcai.org.

Pawelczyk, M.; Datta, T.; Van den Heuvel, J.; Kasneci, G.; and Lakkaraju, H. 2022. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. In *The Eleventh International Conference on Learning Representations*.

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. 2018. Continuous dynamical systems for weighted bipolar argumentation. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*, 148–157.

Potyka, N. 2019. Extending modular semantics for bipolar weighted argumentation. In *International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*, 1722–1730.

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

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-free decision support with quantitative argumentation debates. In *International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, 63–73.

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*, 582–592.

Sakama, C. 2014. Counterfactual reasoning in argumentation frameworks. In *COMMA*, 385–396.

Slack, D.; Hilgard, A.; Lakkaraju, H.; and Singh, S. 2021. Counterfactual explanations can be manipulated. *Advances in neural information processing systems* 34:62–75.

Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* 2.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* 31:841.

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.

Yin, X.; Potyka, N.; and Toni, F. 2024. Explaining arguments' strength: Unveiling the role of attacks and supports. In *International Joint Conference on Artificial Intelligence (IJCAI).*