Strength Change Explanations in Quantitative Argumentation

Timotheus Kampik
Umeå University
Umeå, Sweden
tkampik@cs.umu.se

Xiang Yin
Imperial College London
London, United Kingdom
x.yin20@imperial.ac.uk

Nico Potyka
Cardiff University
Cardiff, United Kingdom
PotykaN@cardiff.ac.uk

Francesca Toni
Imperial College London
London, United Kingdom
f.toni@imperial.ac.uk

ABSTRACT

In order to make argumentation-based inference contestable, it is crucial to explain what changes can achieve a desired (instead of the contested) inference result. To this end, we introduce *strength change explanations* for quantitative (bipolar) argumentation graphs. Strength change explanations describe changes to the initial strengths of a subset of the arguments in a given graph that can achieve a desired ordering based on the final strengths of some (potentially different) subset of arguments. We show that the existing notions of *inverse* and *counterfactual* problems can be reduced to strength change explanations. We also prove basic soundness and completeness properties of our strength change explanations, and demonstrate their existence and non-existence in some special cases. By applying a heuristic search, we demonstrate that we can often successfully find strength change explanations for layered graphs that are common in typical application scenarios; still, limitations remain for settings where we do not provide guarantees for the presence (or absence) of explanations.

KEYWORDS

Formal Argumentation, Explainable AI, Contestability

ACM Reference Format:
Timotheus Kampik, Xiang Yin, Nico Potyka, and Francesca Toni. 2026. Strength Change Explanations in Quantitative Argumentation. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/OKNZ9792

1 INTRODUCTION

In order to facilitate human-centricity, applications of Artificial Intelligence (AI) need to be *contestable*: not only must machines explain the results of their decision-making processes to humans; in addition, humans must be able to challenge these results [18]. Computational argumentation, in which inferences are drawn from potentially dynamic graphs modelling arguments (nodes), as well as attack and support relationships between them (edges), may have the potential to be a key enabler of contestable AI [10, 16].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/OKNZ9792

One way to achieve contestability is to enable machines to provide, given a decision outcome, explanations in the sense of sets of required changes that lead to a more desirable outcome [12, 29]. We define such explanations for Quantitative Bipolar Argumentation Graphs (QBAGs), graphs with weighted nodes and directed edges representing two binary relations modelling *attack* and *support*, respectively. *Gradual semantics* then draw inferences from QBAGs by updating the weights from *initial strengths* to *final strengths*, given the graph topology of the QBAG. QBAGs play an important role in argumentative eXplainable AI (XAI), a line of research that aims to advance the study and application of computational argumentation in the broader XAI context [9]. A range of works showcases the application of QBAGs to XAI use cases, such as explainable image recognition [4] and recommendation systems [27]. To facilitate contestability in QBAGs, we study which changes to the initial strengths of a subset of the arguments in a QBAG can yield a desired outcome in terms of the ordering arising from the final strengths of another subset of the QBAG's arguments.

For example, assume the arguments a and e that model variables of a credit application decision in an intermediate layer of a layered QBAG; d models a variable influencing both a and e. Finally, b and c in the output layer model the acceptance of the application only if the final strength of c is greater than the final strength of b. This ordering can potentially be affected by changes to the initial strengths of a and e. We may want to identify such initial strength changes that specifically affect a change of the final strength ordering $b \succ c$ (b's final strength is greater than c's, *application rejected*) to $c \succ b$ (*application accepted*), cf. Figure 1.

We can identify such changes, which we call *Strength change eXplanations* (SXs), by generalizing the so-called *inverse problem* in quantitative argumentation, which describes the assignment of initial strengths to all arguments in a QBAG such that a desired final strength ordering of these arguments is achieved [22].

Before we commence the formal part of this paper, let us expand on the colloquial intuition of an SX. An SX depends on a QBAG, a gradual semantics, and a subset of the QBAG's arguments, which we call *mutable arguments*. It defines a set of (mutable argument, initial strength)-tuples that, if applied to the QBAG, yield a specific desired ordering given by the final strengths of (some of) the arguments in the QBAG. Roughly, we say that an SX is ϵ-*approximate* if the desired ordering cannot be achieved by a substantially better SX in terms of the sum of all changes to arguments' initial strengths (a smaller sum is better). Intuitively, a 0-*approximate* SX is *optimal*, while a larger ϵ indicates weaker optimality guarantees. Below, we

give an example of SXs, applying a simplistic gradual semantics that (given an acyclic QBAG) traverses the graph in topological order and, given an argument, subtracts the final strengths of all attackers from the argument's initial strength, while adding the final strengths of all supporters[1].

Example 1

Consider the QBAG in Figure 1.1. Nodes in the graphs are arguments, x (i):**f** *represents argument* x *with initial strength* $\tau(\mathrm{x}) = i$ *and final strength* $\sigma(\mathrm{x}) = \mathbf{f}$*, and edges labelled* + *and* − *represent support and attack, respectively. The final strength of* b *is greater than the final strength of* c*:* $\sigma(\mathrm{b}) > \sigma(\mathrm{c})$*. We want to find changes to the initial strengths of the arguments* a *and* e *that yield* $\sigma(\mathrm{b}) < \sigma(\mathrm{c})$*. Such changes are applied in Figures 1.2, 1.3, and 1.4. The changes applied in Figures 1.2 and 1.3 are* ϵ*-approximate, given* $\epsilon = 1$ *(i.e., technically any* $\epsilon \geq 1$ *would work as well). Clearly, the optimal way of achieving the desired ordering is increasing the initial strength of* a *by marginally more than* 1*. As the changes applied in* $\mathcal{G}'$ *and* $\mathcal{G}''$ *are* $|\tau_{\mathcal{G}'}(\mathrm{a}) - \tau_{\mathcal{G}}(\mathrm{a})| + |\tau_{\mathcal{G}'}(\mathrm{e}) - \tau_{\mathcal{G}}(\mathrm{e})| = 2$ *and* $|\tau_{\mathcal{G}''}(\mathrm{a}) - \tau_{\mathcal{G}}(\mathrm{a})| = 2$*, respectively, the changes are still within the approximation "wiggle room" of* 1*. The changes applied to* $\mathcal{G}^*$ *are not* ϵ*-approximate given* $\epsilon = 1$*: we have increased the initial strength of* a *by* 1 *and of* e *by* 2*, but we could have increased the initial strength of* e *by* $< 2 - 1$ *(e.g., by just* 0.5*) and still achieve the desired ordering.*

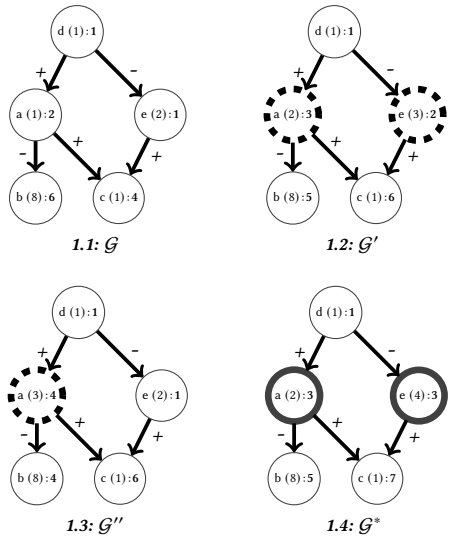


Figure 1: QBAG $\mathcal{G}$ and its updates $\mathcal{G}'$, $\mathcal{G}''$, and $\mathcal{G}^*$. Here and henceforth, a node labelled x (i):**f represents argument** x **with initial strength** $\tau(\mathrm{x}) = i$ **and final strength** $\sigma(\mathrm{x}) = \mathbf{f}$**. Edges labelled + and − represent support and attack, respectively. Arguments with bold borders are strength change explanation arguments, given the desired ordering** $\langle \mathrm{c}, \mathrm{b} \rangle$ **and the *mutable set* {**a, e**}; arguments with bold dashed borders make up** 1***-approximate* strength change explanations.

[1]We use this semantics so that readers can easily verify the examples; our results do not depend on it.

Taking the idea sketched above as a starting point, the contributions of this paper are the following: (i) We provide a formal framework for SXs (Section 4); (ii) We analyse basic properties of *optimal* SXs (Section 5); (iii) We demonstrate existence and non-existence guarantees of SXs for some cases (Section 6); (iv) We implement a heuristics-based search for identifying strength change explanations, drawing from what we have learned from the theoretical analysis and empirically demonstrate the feasibility of finding SXs in small layered QBAGs, as well as some limitations (Section 7); (v) Finally, we formally establish the connection between SXs related approaches in the literature (Section 8).

2 RELATED WORK

The work we present in this paper extends the line of research on argumentative XAI [9, 30] and, more specifically, on the study of explaining inferences drawn from argumentation frameworks—QBAGs, in our case. QBAG explainability has recently been studied in a range of works [3, 13, 14, 32]. Most relevant in our context are studies that define and analyse (i) (any, not only initial strength) changes to QBAGs that can explain changes in the final strength--based ordering of two arguments [14]; (ii) initial strength changes to the arguments in a QBAG that can affect the change of a specific argument's final strength in a desirable way, i.e., so that a desired final strength is achieved or a final strength threshold is exceeded [32]. Our research differs from *(i)* in that we explain how to achieve a *counterfactual*, desired ordering that requires searching for a corresponding graph (changing only initial strengths); in contrast, [14] defines explanations as subsets of changed arguments in a graph, given *actual* (factual) changes. In comparison to *(ii)*, the (strong) counterfactual problem defined in [32] can be reduced to a special case of the SXs we introduce in our work. Finally, our work is conceptually, but not formally, related to the idea of a *semifactual* explanation that, in contrast to a counterfactual explanation, describes a somewhat "maximal" change to a graph that does not affect the outcome of an inference (either in general or to the extent required), as first introduced to argumentative XAI in [1].

Beyond explainability, our work adds to the study of *enforcement*, investigating how desired outcomes can be achieved ("enforced") in different variants of computational argumentation [11]. Here, the most relevant work introduces the *inverse problem* that describes the assignment of initial strengths to a gradual argumentation framework such that a desirable outcome in terms of a final strength-based ordering of the arguments is achieved [22]. Our work extends the inverse problem to generate explanations, in order to support bipolar (instead of attack-only) argumentation graphs, and to allow for constraining the set of arguments whose initial strengths can be changed and whose final strengths are of interest. We provide a more formal integration with [22, 32] by means of an analysis presented in Section 8.

3 PRELIMINARIES

A *Quantitative Bipolar Argumentation Graph (QBAG)* [6, 24] is a tuple $\mathcal{G} = (Args, \tau, Att, Supp)$ consisting of a set of arguments $Args$, an *attack* relation $Att \subseteq Args \times Args$, a *support* relation $Supp \subseteq Args \times Args$ such that $Att \cap Supp = \emptyset$, and an *initial strength* function $\tau : Args \to \mathbb{I}$. We denote the class of QBAGs by $\boldsymbol{Q}$. Given a QBAG

$\mathcal{G} = (Args, \tau, Att, Supp)$ and $x \in Args$, $\mathcal{R}^-_{\mathcal{G}}(x)$ and $\mathcal{R}^+_{\mathcal{G}}(x)$ denote the sets $\{y \mid y \in Args, (y,x) \in Att\}$ and $\{y \mid y \in Args, (y,x) \in Supp\}$, respectively, which we call the attackers/supporters of x. We drop the subscript $\mathcal{G}$ where the context is clear.

For $x, y \in Args$, we say that "x can reach y" iff there is a directed path from x to y in $(Args, Att \cup Supp)$; for $S, S' \subseteq Args$, we say that "S can reach S'" iff $\exists x \in S, y \in S'$ such that x can reach y; analogously, we may say that "x can reach S'", given that $\{x\}$ can reach S'. Given $S \subseteq Args$, we define $\mathcal{G} \downarrow_S := (S, \tau \cap (S \times \mathbb{I}), Att \cap (S \times S), Supp \cap (S \times S))$.

Gradual semantics determine the final strengths of arguments in a QBAG.

Definition 1 (Gradual Semantics and Strength Function [6, 24])
A gradual semantics σ *defines for* $\mathcal{G} = (Args, \tau, Att, Supp)$ *a (possibly partial)* final strength function $\sigma_{\mathcal{G}} : Args \to \mathbb{I} \cup \{\bot\}$ *that assigns the* final strength $\sigma_{\mathcal{G}}(x)$ *to each* $x \in Args$*, where* $\bot$ *is a reserved symbol meaning 'undefined'.*

There is a variety of gradual semantics [2, 7, 23, 28], most of which belong to the class of *modular semantics* [20]. Modular semantics define the strengths of arguments by an iterative process. The strength values of all arguments are initialized with their initial strengths. Then the arguments' strength values are updated based on the strengths of their parents and the base score until they converge. Since the procedure can fail to converge in cyclic graphs, Definition 1 defines the final strength function as a partial function.

Modular semantics are called *modular* because their update function can be decomposed into an aggregation function that aggregates the strength values of attackers and supporters, and an influence function that uses the aggregate to adapt the initial strength. Intuitively, supporters increase the aggregate while attackers decrease it based on their strengths. A positive aggregate will increase the initial strength, while a negative aggregate will decrease it (cf. Example 1). Table 1 provides some aggregation and influence functions from the literature [20, 24]. By combining them, we can obtain the semantics listed in Table 2 including DF-QuAD [28], Euler-based [2] and quadratic energy [23] semantics.

Aggregation Functions	
Sum	$\alpha^{\Sigma}_v(s) = \sum_{i=1}^{n} v_i \times s_i$
Product	$\alpha^{\Pi}_v(s) = \prod_{i:v_i=-1}(1-s_i) - \prod_{i:v_i=1}(1-s_i)$
Influence Functions	
Linear(k)	$\iota^{l}_w(s) = w - \frac{w}{k} \times max\{0,-s\} + \frac{1-w}{k} \times max\{0,s\}$
Euler-based	$\iota^{e}_w(s) = 1 - \frac{1-w^2}{1+w\times e^s}$
p-Max(k)	$\iota^{p}_w(s) = w - w \times h(-\frac{s}{k}) + (1-w) \times h(\frac{s}{k})$,
for $p \in \mathbb{N}$	where $h(x) = \frac{max\{0,x\}^p}{1+max\{0,x\}^p}$

Table 1: Common aggregation α and influence ι functions. Intuitively, $s \in [0,1]^n$ is a strength vector (associating each argument with its current strength), $v \in \{-1,0,1\}^n$ is a relationship vector indicating which arguments attack (-1), support (1) or are in no relationship to (0) the argument of interest, and w is an initial strength.

We can compare gradual semantics based on their satisfaction of argumentation *principles*. Such principles can help us find classes of cases for when, and when not, we can find strength change explanations. Below we provide the definitions of the principles that are relevant for our work.

Semantics	Aggregation	Influence
DF-QuAD (DFQuAD)	Product	Linear(1)
Euler-Based (EB)	Sum	EulerBased
QuadraticEnergy (QE)	Sum	2-Max(1)

Table 2: Examples of gradual semantics.

Directionality describes that adding attacks or supports can only affect their directed successors.

Principle 1 (Directionality)
A gradual semantics σ *satisfies* directionality *iff for all* $\mathcal{G} = (Args, \tau, Att, Supp)$, $\mathcal{G}' = (Args, \tau, Att', Supp')$, *and* $x \in Args$ *s.t.* $Att \cup Supp = Att' \cup Supp' \cup (y,z)$ *for some* $y \in Args$, $z \in Args \setminus \{x\}$ *s.t. there is no directed path from* z *to* x, *we have* $\sigma_{\mathcal{G}}(x) = \sigma_{\mathcal{G}'}(x)$.

Let us note that directionality is a very weak assumption and is satisfied by all modular semantics.

Proposition 3.1 ([26], Theorem 20)**.** *Every modular semantics satisfies directionality.*

Below, we assume an arbitrary QBAG $\mathcal{G} = (Args, \tau, Att, Supp)$ and $x \in Args$. *Stability* describes that any differences between initial and final strength of an argument depend on the presence of attackers or supporters.

Principle 2 (Stability)
A gradual semantics σ *satisfies the* stability principle *iff* $\mathcal{R}^-(x) = \mathcal{R}^+(x) = \emptyset$ *implies that* $\sigma(x) = \tau(x)$.

Stability is a special case of the *balance* property [5], which states that if the attackers and supporters of an argument are equally strong, then its final strength is just its base score.

Principle 3 (Balance)
A gradual semantics σ *satisfies the* balance principle *iff it holds that if the multisets* $\{\sigma(y) \mid y \in \mathcal{R}^-(x)\}$ *and* $\{\sigma(y) \mid y \in \mathcal{R}^+(x)\}$ *are equal then* $\sigma(x) = \tau(x)$.

Balance does not hold for arbitrary modular semantics, but for every *elementary* modular semantics.

Proposition 3.2 ([26], Theorem 20)**.** *Every elementary modular semantics satisfies balance.*

Most QBAG semantics are elementary modular semantics including the Euler-based [2], DF-QuAD [28], Quadratic Energy [23], MLP-based [25], TAC and CRC semantics [26]. They therefore satisfy both directionality and balance.

4 STRENGTH CHANGE EXPLANATIONS (SXS)

Strength change explanations explain how to establish a desired *final strength ordering* of arguments of interest, which we may call *topic arguments*. Here and henceforth, we assume a QBAG $\mathcal{G} = (Args, \tau, Att, Supp)$ and a final strength function $\sigma_{\mathcal{G}}$, omitting the subscript $\mathcal{G}$ where the context is clear.

Definition 2 (Final Strength Ordering, $\preceq$-Satisfaction)
Given $S \subseteq Args$, *we define the* final strength ordering $\preceq^S_{\mathcal{G},\sigma}$ *as follows:*

$$\preceq^S_{\mathcal{G},\sigma} := \{(x,y) \mid x, y \in S, \sigma_{\mathcal{G}}(x) \leq \sigma_{\mathcal{G}}(y)\}.$$

For $\preceq \subseteq Args \times Args$ *we say that "*$\mathcal{G}$ *satisfies* $\preceq$ *w.r.t.* σ*" iff* $\preceq^S_{\mathcal{G},\sigma} = \preceq$ *for some* $S \subseteq Args$. *For* $x, y \in Args$ *s.t.* $(x,y) \in \preceq^S_{\mathcal{G},\sigma}$ *we use the short-hand* $x \preceq^S_{\mathcal{G},\sigma} y$*; we drop the subscript* σ *where the context is clear.*

We go back to the initial figure to give an example of a final strength ordering.

Example 2

Consider again $\mathcal{G}$ from Figure 1.1. Because $\preceq_{\mathcal{G}}^{Args}$ is a total preorder, we can represent it as a sequence of sets, i.e., we can denote $\preceq_{\mathcal{G}}^{Args}$ by $\langle\{d, e\}, \{a\}, \{c\}, \{b\}\rangle$.

Our strength change explanations will change the initial strengths of some the arguments of a QBAG. We first define such *strength changes* in general.

Definition 3 (Strength Change)

A strength change *of a QBAG $\mathcal{G} = (Args, \tau, Att, Supp)$ is a (possibly partial) function $\delta_{\mathcal{G}} : Args \to \mathbb{I}$ such that $\delta_{\mathcal{G}}(x) \neq \tau(x)$ whenever $\delta_{\mathcal{G}}(x) \neq \bot$. We call* $\text{ddom}(\delta_{\mathcal{G}}) = \{y \mid y \in Args, \delta_{\mathcal{G}}(y) \neq \bot\}$ *its* defined domain.

We let $\delta_\emptyset := \emptyset$, which we call the *empty strength change*.

When a strength change is applied to a QBAG, it yields a QBAG with an updated initial strength function.

Definition 4 (Strength Change Application)

We define the application of a strength change $\delta_{\mathcal{G}}$ to $\mathcal{G}$, denoted by $f(\mathcal{G}, \delta)$, as $(Args, \tau', Att, Supp)$, where for $x \in Args$:

$$\tau'(x) = \begin{cases} \delta(x), & \text{if } x \in \text{ddom}(\delta); \\ \tau(x), & \text{otherwise.} \end{cases}$$

We illustrate the concept with our running example.

Example 3

In Figure 1, we can see the following strength changes and their applications:

- $\delta'_{\mathcal{G}} = \{(a, 2), (e, 3)\}$*; applied to $\mathcal{G}$ we obtain $f(\mathcal{G}, \delta') = \mathcal{G}'$.*
- $\delta''_{\mathcal{G}} = \{(a, 3)\}$*; applied to $\mathcal{G}$ we obtain $f(\mathcal{G}, \delta'') = \mathcal{G}''$.*
- $\delta^*_{\mathcal{G}} = \{(a, 2), (e, 4)\}$*; applied to $\mathcal{G}$ we obtain $f(\mathcal{G}, \delta^*) = \mathcal{G}^*$.*

Now we can define which strength changes amount to SXs. For this, we assume a preorder $\preceq$ on our set of arguments *Args*, representing the desired final strength ordering. For $x, y \in Args$ s.t. $(x, y) \in \preceq$, we use the shorthand $x \preceq y$, and if and only if it then does not hold that $y \preceq x$, we may use $x \prec y$.

Definition 5 (Final Strength Change Explanation (SX))

A strength change $\delta_{\mathcal{G}}$ of $\mathcal{G}$ is a final Strength change eXplanation *(SX) of $\preceq$ w.r.t. σ and the* mutable argument set $M \subseteq Args$ *iff it holds that* $\text{ddom}(\delta_{\mathcal{G}}) \subseteq M$ *and $f(\mathcal{G}, \delta)$ satisfies $\preceq$ w.r.t. σ. We denote the set of all SXs of $\preceq$ w.r.t. σ and M by $SX^{\sigma}_{\mathcal{G},M}(\preceq)$.* $\{x | x \in Args, \exists y \in Args : (x \preceq y)$ *or* $(y \preceq x)\}$ *is called the* topic set*, denoted by $T(\preceq)$.*

Whenever we have SXs w.r.t. σ and *Args*, we may drop the subscript M and denote them by $SX^{\sigma}_{\mathcal{G}}(\preceq)$; analogously we may then call $\delta_{\mathcal{G}} \in SX^{\sigma}_{\mathcal{G}}(\preceq)$ an "SX of $\preceq$ w.r.t. σ".

Example 4

We continue Example 3 and let $\preceq$ be the reflexive closure of $\{(b, c)\}$. *We can verify that $\delta'_{\mathcal{G}}$, $\delta''_{\mathcal{G}}$, and $\delta^*_{\mathcal{G}}$ are all SXs of $\preceq$ w.r.t. our naive semantics σ and $M = \{a, e\}$ as we achieve the desired ordering via $\sigma_{\mathcal{G}'}(c) > \sigma_{\mathcal{G}'}(b)$, $\sigma_{\mathcal{G}''}(c) > \sigma_{\mathcal{G}''}(b)$, and $\sigma_{\mathcal{G}^*}(c) > \sigma_{\mathcal{G}^*}(b)$.*

The empty strength change $\delta_\emptyset$ is an SX only if a given QBAG already satisfies the desired final strength ordering.

Lemma 4.1. *$\forall M \subseteq Args$, it holds that $\delta_\emptyset \in SX^{\sigma}_{\mathcal{G},M}(\preceq)$ iff $\mathcal{G}$ satisfies $\preceq$ w.r.t. σ.*

Note that the proofs of lemmata (as well as of complexity results) have been relegated to the technical appendix[2].

5 OPTIMAL SXS

Strength change explanations should not modify the original QBAG more than necessary. This desideratum gives rise to our notion of *approximate* and *optimal* SXs. As a prerequisite, we define the *amount of change* of an SX as the sum of its arguments' deltas to their initial strengths in the QBAG.

Definition 6 (Amount of Change)

Given a strength change $\delta_{\mathcal{G}}$, we call $\|\delta_{\mathcal{G}}\| = \sum_{x \in \text{ddom}(\delta_{\mathcal{G}})} |\delta_{\mathcal{G}}(x) - \tau(x)|$ the amount of change of $\delta_{\mathcal{G}}$.

An SX is an ϵ-approximate SX if it changes the initial strengths of arguments by not more than ϵ more than necessary. Hence, a 0-approximate SX is *optimal*.

Definition 7 (Approximate and optimal SX)

$\delta_{\mathcal{G}} \in SX^{\sigma}_{\mathcal{G},M}(\preceq)$ is an ϵ-approximate SX *of $\preceq$ w.r.t. σ, and $M \subseteq Args$ iff there exists no $\delta'_{\mathcal{G}} \in SX^{\sigma}_{\mathcal{G},M}(\preceq)$ s.t. $\|\delta'_{\mathcal{G}}\| < \|\delta_{\mathcal{G}}\| - \epsilon$ (with $\epsilon \in \mathbb{R}_{\geq 0}$). $\delta_{\mathcal{G}}$ is an* optimal SX *iff it is a 0-approximate SX. We denote the set of all ϵ-approximate SXs of $\preceq$ w.r.t. σ and M by $SX^{\sigma*}_{\mathcal{G},M}(\preceq, \epsilon)$.*

Again, whenever we have ϵ-approximate SXs w.r.t. σ and *Args*, we may drop the subscript M and simply denote them by $SX^{\sigma*}_{\mathcal{G}}(\preceq, \epsilon)$; analogously we may then call $\delta_{\mathcal{G}} \in SX^{\sigma*}_{\mathcal{G}}(\preceq, \epsilon)$ an "ϵ-approximate SX of $\preceq$ w.r.t. σ".

Example 5

Consider again the desired ordering $\preceq$ (Example 4), the mutable set $M = \{a, e\}$, and the strength changes $\delta'_{\mathcal{G}}$, $\delta''_{\mathcal{G}}$, and $\delta^*_{\mathcal{G}}$ (Example 3). We can observe that:

- *$\delta'_{\mathcal{G}}$ and $\delta''_{\mathcal{G}}$ are 1-approximate SXs of $\preceq$ w.r.t. our naive semantics σ and M. In the case of $\delta'_{\mathcal{G}}$, we change the initial strengths of* a *to* 2 *and of* e *to* 3 *(in sum, a change of* 2*). Given our wiggle room of $\epsilon = 1$, we cannot change the initial strengths of* a *and* e *substantially less to achieve the desired ordering: while we could change the initial strength of* a *just marginally more and abstain from changing* e*, the change to* a *would then amount to > 1, which is greater than $2 - \epsilon$. In the case of $\delta''_{\mathcal{G}}$, we clearly need to change the initial strength of at least one argument and we cannot change the initial strength of* a *substantially less (a change to at least marginally greater than* 2 *is required).*
- *In contrast, $\delta^*_{\mathcal{G}}$ is not a* 1*-approximate SX, as the strength change* $(\{a, 2.99\})$ *(a slightly smaller change than $\delta''_{\mathcal{G}}$) achieves the desired ordering by changing the initial strength of* a *substantially (by more than* 1*) less.*

If a desired strength ordering is already satisfied, only the empty strength change is an optimal SX.

Lemma 5.1. *If $\mathcal{G}$ satisfies $\preceq$ w.r.t. σ then $SX^{\sigma*}_{\mathcal{G}}(\preceq, 0) = \{\delta_\emptyset\}$.*

6 (NON-)EXISTENCE OF SXS

Finding SXs is a difficult problem. Note that SXs may not exist. For example, if the mutable arguments cannot reach our topic arguments, given modular semantics we cannot achieve a desired ordering by changing their initial strengths. Below, we analyse some basic properties w.r.t. the existence and non-existence of SXs.

[2]Available in the version at https://arxiv.org/abs/2603.00008.

As a prerequisite, we introduce some additional gradual semantics principles. The first one is a variant of directionality (Principle 1) pertaining to the existence of arguments rather than edges.

Principle 4 (Strong Directionality)
A gradual semantics σ satisfies the strong directionality principle *iff for all $\mathcal{G} = (Args, \tau, Att, Supp)$, $x \in Args$, and $\mathcal{G}' := \mathcal{G} \downarrow_{Args \setminus Args'}$ s.t. $Args'$ cannot reach x it holds that $\sigma_{\mathcal{G}}(x) = \sigma_{\mathcal{G}'}(x)$.*

Modular semantics satisfy strong directionality.

Lemma 6.1. *Every modular semantics satisfies strong directionality.*

Next, we consider a variant of monotonicity [5] that explicitly assumes an initial strength difference (the general principle is not sufficiently explicit about this case for our purposes).

Principle 5 (Weak Monotonicity)
A gradual semantics σ satisfies the weak monotonicity principle *iff for all $\mathcal{G} = (Args, \tau, Att, Supp)$, $x, y \in Args$, the following statements hold if $\mathcal{R}^-(x) \supseteq \mathcal{R}^-(y)$ and $\mathcal{R}^+(x) \subseteq \mathcal{R}^+(y)$:*

(1) if $\tau(x) \leq \tau(y)$ then $\sigma(x) \leq \sigma(y)$;
(2) if $\sigma(y) < \sigma(x)$ then $\tau(y) < \tau(x)$.

Intuitively, one would expect that many modular semantics satisfy weak monotonicity: initially weaker arguments with strictly less (or the same) attackers and more (or the same) supporters should be finally weaker as well.

Proposition 6.2. *DFQuAD, EB, and QE semantics satisfy weak monotonicity.*

Proof. Consider an aggregation function α that is either *Product*, as applied by DFQuAD semantics, or *Sum*, as applied by EB and QE semantics (cf. Tables 1 and 2). Given two arguments x and y s.t. $\mathcal{R}^-(x) \supseteq \mathcal{R}^-(y)$, and $\mathcal{R}^+(x) \subseteq \mathcal{R}^+(y)$ and their strength and relationship vectors s_x, v_x and s_y, v_y, respectively, it holds that $\alpha_{v_x}(s_x) \leq \alpha_{v_y}(s_y)$. Given this and if $\tau(x) \leq \tau(y)$ (Principle 5, Condition 1), it follows for an influence function ι that is either *Linear(1)* (for DFQuAD semantics), *EulerBased* (for EB), or *2-Max(l)* (for QE) that $\iota_{\tau(x)}(\alpha_{v_x}(s_x)) \leq \iota_{\tau(y)}(\alpha_{v_y}(s_y))$; conversely, because $\alpha_{v_x}(s_x) \leq \alpha_{v_y}(s_y)$, if $\sigma(y) < \sigma(x)$ (Condition 2), this can only be achieved by differences in initial strengths, i.e., given $\tau(y) < \tau(x)$. □

Below, we assume our final strength function σ is based on a modular gradual semantics and we only consider QBAGs that do not have undefined final strengths given σ. We first give several conditions under which we cannot find SXs, given the desired ordering $\preceq$ is currently not satisfied, i.e., we assume that $\mathcal{G}$ does not satisfy $\preceq$ w.r.t. σ.

If two arguments whose relative final strengths need to change cannot be reached by the set of mutable arguments, then we cannot find an SX.

Proposition 6.3. *Given a modular semantics σ it holds that $SX^{\sigma}_{\mathcal{G},M}(\preceq) = \emptyset$ if $\exists x, y \in Args$ s.t. $x \preceq y$ but $x \not\xi^{Args}_{\mathcal{G}} y$ and M cannot reach $\{x, y\}$.*

Proof. Consider $x, y \in Args$ s.t. $x \preceq y$ but $x \not\xi^{Args}_{\mathcal{G}} y$ (as assumed by the proposition). Observe that every modular semantics σ satisfies strong directionality (Lemma 6.1). Thus, because M cannot reach $\{x, y\}$, for every $\delta_{\mathcal{G}}$ s.t. $\{z \mid (z, s) \in \delta_{\mathcal{G}}\} \subseteq M$, for every $S \subseteq Args$ it must hold that $x \not\xi^{S}_{f(\mathcal{G},\delta)} y$ and consequently $\delta_{\mathcal{G}} \notin SX^{\sigma}_{\mathcal{G},M}(\preceq)$, whence $SX^{\sigma}_{\mathcal{G},M}(\preceq) = \emptyset$. □

An immediate consequence is that we cannot find an SX if no mutable argument can reach any of the topic arguments.

Corollary 6.4. *Given a modular semantics σ, $SX^{\sigma}_{\mathcal{G},M}(\preceq) = \emptyset$ holds if M cannot reach $T(\preceq)$ and $\mathcal{G}$ does not satisfy $\preceq$ w.r.t. σ.*

Proof. As $\mathcal{G}$ does not satisfy $\preceq$ w.r.t. σ it must hold that $\exists(x, y) \in Args$ s.t. $x \preceq y$ but M cannot reach $\{x, y\}$ and for every $S \subseteq Args$ it holds that $x \not\xi^{S}_{\mathcal{G}} y$. Hence, the proof follows directly from Proposition 6.3. □

Assuming our gradual semantics satisfies weak monotonicity, we cannot find an SX, either, if two topic arguments whose relative final strengths need to *inverse* are not mutable arguments and have the same attackers and supporters.

Proposition 6.5. *Given a semantics σ that satisfies weak monotonicity it holds that $SX^{\sigma}_{\mathcal{G},M}(\preceq) = \emptyset$ if $\exists x, y \in Args$ s.t. $x \prec y$ but $x \not\xi^{Args}_{\mathcal{G}} y$, $x, y \notin M$, and $\mathcal{R}^-(x) = \mathcal{R}^-(y)$, as well as $\mathcal{R}^+(x) = \mathcal{R}^+(y)$.*

Proof. Consider $x, y \in Args$ s.t. $x \preceq y$ but $x \not\xi^{Args}_{\mathcal{G}} y$, as well as $x, y \notin M$ and $\mathcal{R}^-(x) = \mathcal{R}^-(y)$, as well as $\mathcal{R}^+(x) = \mathcal{R}^+(y)$ (as assumed in the proposition). This means that $\sigma(x) > \sigma(y)$ must hold (as implied by $x \not\xi^{Args}_{\mathcal{G}} y$). Because σ satisfies weak monotonicity, it must hold that $\tau(x) > \tau(y)$ (as x and y share all attackers and supporters). Consequently, for any $\delta_{\mathcal{G}}$ s.t. $\{z \mid (z, s) \in \delta_{\mathcal{G}}\} \subseteq M$ it must hold for $\mathcal{G}' := f(\mathcal{G}, \delta)$ that $\sigma_{\mathcal{G}'}(x) \geq \sigma_{\mathcal{G}'}(y)$ (otherwise, we would again violate weak monotonicity). Therefore, it holds that $\delta_{\mathcal{G}} \notin SX^{\sigma}_{\mathcal{G},M}(\preceq)$ and thus $SX^{\sigma}_{\mathcal{G},M}(\preceq) = \emptyset$. □

We now move on to some cases where we can guarantee that SXs exist. First, if all topic arguments are mutable and have neither attackers nor supporters, we can achieve the desired ordering by modifying their initial strengths directly, assuming our semantics satisfies stability.

Proposition 6.6. *Given a gradual semantics σ satisfying stability, $SX^{\sigma}_{\mathcal{G},T}(\preceq) \neq \emptyset$ if $\forall x \in T(\preceq)$ it holds that $\mathcal{R}^-(x) = \mathcal{R}^+(x) = \emptyset$ and $x \in M$.*

Proof. Because σ satisfies stability and $\forall x \in T(\preceq)$ it holds that $\mathcal{R}^-(x) = \mathcal{R}^+(x) = \emptyset$, for every $\mathcal{G}' = (Args, \tau', Supp, Att)$ for every initial strength function τ' it must hold that $\tau'(x) = \sigma'_{\mathcal{G}}(x)$. We can hence achieve a mapping $\delta_{\mathcal{G}} : T(\preceq) \to \mathbb{R}$ s.t. $\forall y, z \in T(\preceq)$ it holds that $\delta_{\mathcal{G}}(y) \leq \delta_{\mathcal{G}}(z)$ iff $y \preceq z$, thus achieving that $f(\mathcal{G}, \delta)$ satisfies $\preceq$; then, by definition of an SX (Definition 5), it must hold that $\delta_{\mathcal{G}} \in SX^{\sigma}_{\mathcal{G},M}(\preceq)$, i.e., $SX^{\sigma}_{\mathcal{G},M}(\preceq) \neq \emptyset$; intuitively: as all topic arguments are mutable arguments without external influence, we can change their initial strengths directly s.t. we achieve the desired ordering $\preceq$. □

Similarly, we can guarantee the existence of SXs if all topic arguments are mutable arguments that cannot reach each other and we can achieve zero influence of all incoming attackers and supporters, by changing mutable arguments other than the topic arguments.

Proposition 6.7. *Given a gradual semantics σ satisfying balance, it holds that $SX^{\sigma}_{\mathcal{G},M}(\preceq) \neq \emptyset$ if $T(\preceq) \subseteq M$, $\forall x, y \in T(\preceq)$ s.t. $x \neq y$ it holds that x cannot reach y, and there exists a strength change $\delta_{\mathcal{G}}$ s.t. $ddom(\delta_{\mathcal{G}}) \subseteq M \setminus T(\preceq)$ and $\forall z \in Args$ s.t. $\exists x \in T(\preceq) : z \in \mathcal{R}^-(x) \cup \mathcal{R}^+(x)$ it holds that $\sigma_{f(\delta,\mathcal{G})}(z) = 0$.*

Proof. Because σ satisfies balance and there exists a strength change $\delta_{\mathcal{G}}$ s.t. $\forall z \in Args$ s.t. $\exists x \in T(\preceq) : z \in \mathcal{R}^-(x) \cup \mathcal{R}^+(x)$ it holds that $\sigma_{f(\delta,\mathcal{G})}(z) = 0$, for this strength change $\delta_{\mathcal{G}}$, $\forall x \in T(\preceq)$ it also holds that $\sigma_{f(\delta,\mathcal{G})}(x) = \tau(x)$. Because it also holds that $ddom(\delta_{\mathcal{G}}) \subseteq M \setminus T(\preceq)$, we can apply another strength change $\delta'_{\mathcal{G}}$ s.t. $\forall x \in T(\preceq)$ it still holds that $\sigma_{f(\delta',f(\delta,\mathcal{G}))}(x) = \tau(x)$ and in addition such that any total preorder $\preceq$ (on $T(\preceq)$, obviously), can be achieved by assigning final strengths (in $\mathbb{R}$) to all arguments in $T(\preceq)$ accordingly, analogous to how we can achieve this for Proposition 6.6. □

7 HEURISTIC SEARCH

Experimental Setups. We conduct experiments on *layered* acyclic QBAGs that we call *MLP-like QBAGs* because they feature a feed-forward structure like Multi-Layer Perceptrons (MLPs). In a layered QBAG, arguments can be partitioned into layers, such that only (and all) arguments in the first layer have no parents; arguments in the second layer then have parents only in the first layer and children in the third layer, and so forth. Only (and all) nodes in the final layer do not have children. Layered argumentation graphs are common in applications of CA, both generally [19, 21, 31] and specifically for weighted argumentation variants such as QBAGs [4, 8, 27].

While QBAGs in many real-world scenarios are naturally acyclic, extending our approach to cyclic QBAGs remains future work. Since there are no public benchmark datasets for QBAGs, we use synthetic graphs to evaluate the performance. We next distinguish two MLP types. The first type is randomly generated and may not exhibit a solution. The second type is constructed with additional constraints to guarantee the existence of a solution and we refer to them as *constrained QBAGs*. For both types, we consider four different structures that vary in the number of layers and arguments per layer: $[8, 32, 16, 3]$, $[8, 32, 16, 8]$, $[8, 64, 16, 8, 3]$ and $[8, 64, 16, 8, 8]$. For example, $[8, 32, 16, 3]$ represents 8 arguments in the first layer, 32 in the second, 16 in the third, and 3 in the final layer. We refer to the layers between the first and the last as *intermediate layers*. Arguments are assigned random base scores uniformly sampled from $[0, 1]$. Edges are added between all arguments in adjacent layers, making the QBAGs fully connected between consecutive layers. Each edge is independently labelled as either attack or support with equal probability. The topic arguments are set as those in the final layer, and the desired ordering follows the decreasing strengths of these arguments. To reduce the effect of randomness, we create 100 QBAGs for each structure. Finally, we use DF-QuAD semantics for evaluation due to its wide applicability (cf. [8, 15]).

In the constrained QBAG setting, the same structural templates are reused but with additional constraints to ensure that the SXs are guaranteed to exist. Let the layers be denoted by L_1 (the fist layer) to L_n (the final layer). Similarly, we set arguments in L_n as the topic arguments, which are mutually independent of each other. Our focus is on decreasing the strength of arguments in layer L_{n-1} to 0, so that they have no influence on the final layer. Then, a valid SX can be obtained if the algorithm successfully identifies a decreasing ordering of base scores in L_n. To this end, we make the arguments in L_{n-1} immutable and assign them with small random base scores uniformly sampled from $[0, 0.1]$, so that their strength can be more easily decreased to 0 by attackers from L_{n-2}. Furthermore, we enforce that L_{n-2} contains only attack relations targeting L_{n-1}, and L_{n-3} contains only support relations targeting L_{n-2}. This design ensures that the strengths of arguments in L_{n-2} can be maximised through supports from L_{n-3}, enabling them to strongly attack and minimise the arguments in L_{n-1}.

Objective Function and Optimisation Setups. Suppose the final layer arguments are denoted by $a_1, a_2, \cdots, a_n (n > 1)$ with a desired ordering $\sigma(a_1) \geq \sigma(a_2) \geq \cdots \geq \sigma(a_n)$. To find an ordering by local search, we need an objective function that decreases with respect to the number of order-constraint violations. We adopt the ReLU cost function $cost(\sigma) := \sum_{i<j} max(0, (\sigma(a_j) - \sigma(a_i)))$.

We employ the gradient descent algorithm with Adam optimiser to minimise the cost, with a maximum of 100 iterations. While it may converge to local minima, it can serve here as a proof of concept enabling our heuristic search. To evaluate the optimisation results, we first check *validity*, i.e., whether the final ranking exactly matches the desired ordering. Additionally, we employ two standard ranking correlation metrics: *Kendall's* τ and *Spearman's* ρ ranking correlation (cf. [17]). These metrics directly capture ordering quality. Both metrics range from -1 (reverse order) to 1 (equal order), with higher scores indicating better alignment with the desired ordering. We also report the average runtime across all QBAGs, as well as the average absolute base score difference (per argument) for those QBAGs for which we successfully identify the desired ordering.

Algorithm 1 illustrates the iterative heuristic search, which consists of three main steps. Since a valid solution may not always exist, a maximum number of iterations is set to prevent infinite loops. First, the algorithm computes the ReLU cost. If the cost equals 0, indicating that a valid solution has been found, the algorithm returns the base score function; otherwise, it proceeds to the next step. Second, the gradients of the cost function w.r.t. each mutable argument is computed and stored. Finally, the base scores are updated based on their corresponding gradients, with the dynamic learning rate α provided by the Adam optimiser. If no solution is found, we return `null`.

Let us formally observe the time complexity of Algorithm 1.

Proposition 7.1. *Let n be the number of topic arguments, K the maximum number of iterations, and $N = |Args| + |Att| + |Supp|$. For acyclic QBAGs, the time complexity of Algorithm 1 is $O(K \cdot (|M| \cdot N + n^2))$.*

We have seen that SXs may or may not exist. Deciding their existence and finding an optimal SX are challenging problems: Even in acyclic, layered QBAGs, the impact of one argument's initial strength on the final strength of another argument may not be monotonic. E.g., increasing an argument's initial strength by a value of, assume, 0.1, may have a positive impact on the final strength of another argument and further increasing the initial strength (e.g. by 0.11 instead of just 0.1) may then have a negative impact (cf. Figure 1 in [13]). To address this challenge pragmatically, we design and implement a local search (gradient descent) algorithm

Algorithm 1 Heuristic Search

Input: QBAG $\mathcal{G} = (Args, \tau, Att, Supp)$, semantics σ, learning rate α, mutable set $M \subseteq Args$, desired ordering $\sigma(a_1) \geq \cdots \geq \sigma(a_n)$
Parameter: Perturbation value ε, maximum iterations K
Output: Updated τ (which also is an SX)

$\nabla cost = \{\}$ # *gradient dictionary*
for $k = 1$ to K **do**
  # *1. compute cost*
  compute $\sigma(a)$ for all $a \in Args$
  $cost \leftarrow \sum_{1 \leq i < j \leq n} \max(0, \sigma(a_j) - \sigma(a_i))$
  if $cost = 0$ **then**
    return τ # *solution found*

  # *2. compute gradients*
  for a in M **do**
    $\tau(a) \leftarrow \tau(a) + \varepsilon$ # *perturb* $\tau(a)$
    compute $\sigma(a)$ for all $a \in Args$
    $cost' \leftarrow \sum_{1 \leq i < j \leq n} \max(0, \sigma(a_j) - \sigma(a_i))$
    $\nabla cost[a] \leftarrow (cost' - cost)/\varepsilon$
    $\tau(a) \leftarrow \tau(a) - \varepsilon$ # *restore* $\tau(a)$

  # *3. update base scores*
  for a in M **do**
    $\tau(a) \leftarrow \max(0, \min(1, \tau(a) - \alpha \cdot \nabla cost[a]))$
return `null` # *No solution found*

that tries to find SX by minimising the violation of order constraints. Strictly speaking, since our algorithm takes gradient information into account, it can only be applied if the strength function is differentiable. While this is the case for acyclic QBAGs, the strength function for cyclic QBAGs is not necessarily differentiable, and even if it was, it would be difficult to derive a closed-form solution for the partial derivatives. In principle, one could replace the partial derivatives with difference quotients in this case. Still, since the majority of QBAG applications results in acyclic graphs, we focus on this case.

Results and Analysis. Table 3 shows the experimental results. The third column shows the results for the constrained QBAGs. Our algorithm consistently finds SXs, achieving 100% validity, which results in the best Kendall and Spearman correlation. The runtime increases with both the number of intermediate layers and the number of topic arguments involved in the desired ordering. The average absolute base score differences are larger for those QBAGs with more topic arguments.

The remaining three columns show results of the random QBAGs under different configurations of mutable arguments. Theoretically, if all arguments are mutable, SXs always exist by directly assigning decreasing base scores to the topic arguments and zero to all others, thereby nullifying any undesired influence. Our experimental results (shown in the last column) confirm that the algorithm reliably identifies SXs under this condition. Although the validity reaches 99% for the final configuration, the algorithm successfully could find a solution for the previously failed case after increasing the number of iterations to 1000. However, in cases with only partially mutable arguments, our algorithm does not always succeed. This may be attributed to several possible factors: SXs may not exist, the number of iterations may be insufficient, or the algorithm may have converged

Table 3: Average validity, Kendall & Spearman correlation, runtime (in seconds), and absolute base score difference (per argument) over 100 MLP-like QBAGs with varying structures.

Structure	Metric	Constrained L_{n-1} fixed	First mutable	Interm. mutable	All mutable
[8,32,16,3]	Validity	**100%**	0%	83%	**100%**
	Kendall	1.00	-0.24	0.78	1.00
	Spearman	1.00	-0.24	0.78	1.00
	Runtime	0.03	1.09	0.28	0.03
	$\lvert\Delta$ BS$\rvert$	0.01	NA	0.30	0.15
[8,32,16,8]	Validity	**100%**	0%	33%	**100%**
	Kendall	1.00	-0.02	0.62	1.00
	Spearman	1.00	-0.03	0.68	1.00
	Runtime	0.11	7.46	1.01	0.11
	$\lvert\Delta$ BS$\rvert$	0.06	NA	0.39	0.27
[8,64,16,8,3]	Validity	**100%**	3%	87%	**100%**
	Kendall	1.00	-0.19	0.89	1.00
	Spearman	1.00	-0.20	0.90	1.00
	Runtime	0.08	3.57	0.70	0.08
	$\lvert\Delta$ BS$\rvert$	0.01	~ 0	0.08	0.04
[8,64,16,8,8]	Validity	**100%**	0%	24%	99%
	Kendall	1.00	0.02	0.54	0.99
	Spearman	1.00	0.03	0.61	0.99
	Runtime	0.34	3.99	3.19	0.40
	$\lvert\Delta$ BS$\rvert$	0.03	NA	0.12	0.10

to a local minimum, which is a known limitation of gradient-based methods. Despite these challenges, we observe a clear trend: as the number of mutable arguments increases—from only first layer mutable, to intermediate layers mutable, and finally to all layers mutable—the validity, Kendall, and Spearman correlation scores improve consistently, which aligns with our expectation[3]. As for the absolute base score difference, we observe that configurations with more topic arguments require larger adjustments when the number of mutable arguments is fixed.

Our experiments demonstrate that, while the general problem is challenging, our algorithm can reliably identify SXs in some scenarios where we can guarantee the existence of SXs. Accordingly, future research towards more applied directions, e.g. by utilising SXs for MLP debugging, can be considered promising.

8 RELATING SXS TO INVERSE & COUNTERFACTUAL PROBLEMS

SXs are closely related to the *inverse problem* as introduced by [22], as well as to the related *strong counterfactual problem* [32] that is defined as a stepping stone to counterfactual explanations for QBAGs. In this section, we will show the following: (i) Every solution of an inverse problem is also an SX; note that the reverse is not the case as SXs can specify specific sets of topic and mutable arguments, and can start off with arbitrary initial strengths assignments; (ii) Strong counterfactual problems and their solutions can be reduced to SXs; again the reverse is not the case, as SXs cover preferences over arbitrary many arguments in a QBAG.

To be able to formally integrate our explanations into the body of related work, we introduce some additional definitions, starting with the *inverse problem*, that given an argumentation framework

[3]Note: the technical appendix contains results with an additional experimental setting where both first and intermediate layers are mutable; we also report results for EB and QE semantics.

without initial strengths seeks to identify an initial strength assignment that achieves a desired final strength-based ordering of the arguments. Note that [22] defines the inverse problem for attack-only instead of bipolar argumentation frameworks and semantics; for the sake of conciseness, we generalise immediately to QBAGs.

Definition 8 (Inverse Problem)
An inverse problem *with respect to a gradual semantics σ is a 4-tuple $I = (Args, Att, Supp, \preceq)$, where Args is a set of arguments, $Att, Supp \subseteq Args \times Args$, and $\preceq \subseteq Args \times Args$. $\preceq$ is called the* desired ordering. *A solution of the inverse problem I is an initial strength function $\tau : Args \rightarrow \mathbb{I}$ s.t. $\{(x, y) \mid x, y \in Args, \sigma(x) \leq \sigma(y)\} = \preceq$.*

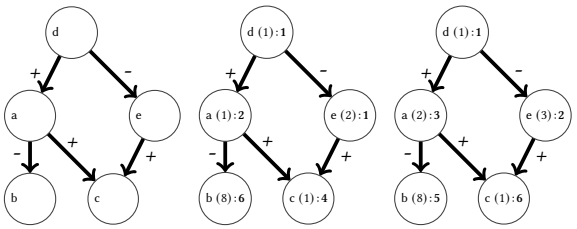
We denote the class of inverse problems by $\mathcal{I}$.

A somewhat similar problem has been introduced as a prerequisite of an argumentation-based XAI approach. The *strong counterfactual problem* describes, given a QBAG and a topic argument of that QBAG, the identification of an initial strength function that achieves a specific desired final strength of the topic argument [32].

Definition 9 (Strong Counterfactual Problem)
The strong counterfactual problem with respect to an argumentation semantics σ is a 3-tuple $C = (\mathcal{G}, x, s)$, where $\mathcal{G} = (Args, \tau, Att, Supp)$ is a QBAG, $x \in Args$, $s \in \mathbb{I}$ as well as $s \neq \sigma_{\mathcal{G}}(x)$. The solution of the strong counterfactual problem C is an initial strength function $\tau' \neq \tau$ such that, given $\mathcal{G}' = (Args, \tau', Att, Supp)$, it holds that $\sigma_{\mathcal{G}'}(x) = s$.

The following example illustrates the two *problems*.

Example 6
Consider $\mathcal{G}$ in Figure 2.1[4]*, with arguments $Args = \{a, b, c, d, e\}$, $Att = \{(a, b), (d, e)\}$, and $Supp = \{(a, c), (d, a), (e, c)\}$. With the sequence $\preceq^* = \langle d, e, a, b, c\rangle$ giving rise to the corresponding total order $\preceq$*[5]*, we have the inverse problem $I = (Args, Att, Supp, \preceq)$. Given $\mathcal{G}^*$ in Figure 2.2, $C = (G^*, c, 6)$ is a strong counterfactual problem. The initial strength function seen in $\mathcal{G}'$ (Figure 2.3), i.e., $\tau' = \{(a, 2), (b, 8), (c, 1), (d, 1), (e, 3)\}$, is a solution of I, and of C.*



2.1: $\mathcal{G}$ 2.2: $\mathcal{G}^*$ 2.3: $\mathcal{G}'$

Figure 2: Inverse and strong counterfactual problems ($\mathcal{G}$, with desired total order $\langle d, e, a, b, c\rangle$, and $\mathcal{G}^*$, with topic c and desired strength 6, respectively) and their solutions (in $\mathcal{G}'$).

To show that SXs generalise inverse problems and their solutions, we first introduce a function that assigns an arbitrary value $s \in \mathbb{I}$ as the initial strength to all arguments of an inverse problem.

Definition 10 (Initial Strength Assignment Function)
The initial strength assignment function $\phi_s : \mathcal{I} \rightarrow \mathcal{Q}$, with $s \in \mathbb{I}$, takes an inverse problem $(Args, Att, Supp, \preceq) \in \mathcal{I}$ and returns a QBAG $(Args, \tau, Att, Supp) \in \mathcal{Q}$ s.t. $\tau = \{(x, s) | x \in Args\}$.

We can then show that the solution of an inverse problem is also an SX, assuming an inverse problem that is augmented with an initial strength assignment function assigning arbitrary initial strengths s to all arguments, and excluding initial strength assignments of arguments to s from the solution.

Proposition 8.1. *For every inverse problem $I = (Args, Att, Supp, \preceq)$, for every initial strength function τ that is a solution of I, for every $s \in \mathbb{I}$ it holds that $\tau \setminus \{x | x \in Args, (x, s) \in \tau\} \in SX^{\sigma}_{f(\phi_s(I),\tau),Args}$.*

Proof. By definition of an inverse problem and its solution, $f(\phi_s(I), \tau)$ satisfies $\preceq$. However, for some $(x, s') \in \tau$ it may hold that $s' = s$ and therefore (x, s') must not occur in a strength change. Hence, $\tau \setminus \{x | x \in Args, (x, s) \in \tau\} \in SX^{\sigma}_{\phi_s(I),Args}(\preceq)$. □

Similarly, we can show that strong counterfactual problems and their solutions can be reduced to SXs: a change to the initial strengths of arguments in a QBAG that leads to a desired final strength of a specific topic argument can be characterised by an SX, given we add a "dummy argument" to the QBAG that serves as a reference to the desired final strength of the topic argument. Here, we assume the gradual semantics satisfies the stability principle, which we claim is a common-sense desideratum.

Proposition 8.2. *Given a gradual semantics σ satisfying stability, for every strong counterfactual problem $C = (\mathcal{G} = (Args, \tau, Att, Supp), x, s)$, for every τ' that is a solution of C it holds that $\tau' \in SX^{\sigma}_{\mathcal{G}_y,Args}(\preceq)$, where $\mathcal{G}_y = (Args \cup \{y\}, \tau \cup \{(y, s)\}, Att, Supp)$, $y \notin Args$, and $\preceq = \{(x, y), (y, x)\}$.*

Proof. Because σ satisfies stability and $y \notin Args$, for $\mathcal{G}'_y = (Args \cup \{y\}, \tau' \cup \{(y, s)\}, Att, Supp)$ it must hold that $\sigma_{\mathcal{G}_y}(y) = \sigma_{\mathcal{G}'_y}(y) = s$ (note that y has neither attackers nor supporters). This means by definition of a strong counterfactual problem and its solution, we must have $\sigma_{\mathcal{G}'_y}(x) = \sigma_{\mathcal{G}'_y}(y)$. It follows that because $\mathcal{G}'_y = f(\mathcal{G}_y, \tau')$, it holds that $\tau' \in SX^{\sigma}_{\mathcal{G}_y,Args}(\preceq)$, with $\preceq = \{(x, y), (y, x)\}$, as achieved by $\sigma_{\mathcal{G}'_y}$. □

Let us expand on Example 6 to give an intuition of the results.

Example 7
Consider the previous inverse and strong counterfactual problems $I = (Args, Att, Supp, \preceq)$ and $C = (\mathcal{G}^* = (Args, \tau^*, Att, Supp), c, 6)$, respectively (cf. Figure 2), as well as their solution τ'. We observe that:

- *Given $\mathcal{G}_0 = (Args, \{(x, 0) \mid x \in Args\}, Att, Supp)$ it holds that $\tau' \in SX^{\sigma}_{\mathcal{G}_0,Args}(\preceq)$;*
- *Given $\mathcal{G}_y = (Args \cup \{y\}, \tau_y = \tau \cup \{(y, 6)\}, Att, Supp)$ it holds that $\tau' \setminus \tau^* \in SX^{\sigma}_{\mathcal{G}_y,Args}(\{(c, y), (y, c)\})$.*

9 CONCLUSIONS

We have introduced argumentative strength change explanations, as a potential foundation for argumentative XAI and contestable AI. Our explanations generalise solutions of previously studied problems in gradual argumentation. We have demonstrated some (non)existence results, as well as the empirical feasibility of finding explanations in relatively small, layered QBAGs, with some expected limitations. Future research may expand our investigations regarding theoretical existence and empirical find-ability of our strength change explanations, especially in large QBAGs, measure other characteristics of the explanations, such as simplicity and robustness, and apply the explanations to real-world contestability problems and datasets.

[4] Here, $\mathcal{G}$ is technically not a QBAG.
[5] I.e., $\preceq$ is the transitive and reflexive closure of $\{(d, e), (e, a), (a, b), (b, c)\}$.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] Gianvincenzo Alfano, Sergio Greco, Francesco Parisi, and Irina Trubitsyna. 2024. Counterfactual and Semifactual Explanations in Abstract Argumentation: Formal Foundations, Complexity and Computation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024*, Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco (Eds.). https://doi.org/10.24963/KR.2024/2

[2] Leila Amgoud and Jonathan Ben-Naim. 2018. Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning* 99 (2018), 39–55.

[3] Caren Al Anaissy, Jérôme Delobelle, Srdjan Vesic, and Bruno Yun. 2024. Impact Measures for Gradual Argumentation Semantics. *CoRR* abs/2407.08302 (2024). https://doi.org/10.48550/ARXIV.2407.08302 arXiv:2407.08302

[4] Hamed Ayoobi, Nico Potyka, and Francesca Toni. 2024. Argumentative Interpretable Image Classification. In *2nd International Workshop on Argumentation for eXplainable AI co-located with 10th International Conference on Computational Models of Argument (COMMA 2024), Hagen, Germany, September 16, 2024 (CEUR Workshop Proceedings, Vol. 3768)*. CEUR-WS.org, 3–15.

[5] Pietro Baroni, Antonio Rago, and Francesca Toni. 2018. How Many Properties Do We Need for Gradual Argumentation?. In *Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI 2018)*. AAAI Press, 1736–1743.

[6] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From Fine-Grained Properties to Broad Principles for Gradual Argumentation: A Principled Spectrum. *International Journal of Approximate Reasoning* 105 (feb 2019), 252–286. https://doi.org/10.1016/j.ijar.2018.11.019

[7] Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation* 6, 1 (2015), 24–49.

[8] Oana Cocarascu, Antonio Rago, and Francesca Toni. 2019. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 1261–1269. http://dl.acm.org/citation.cfm?id=3331830

[9] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *30th International Joint Conference on Artificial Intelligence*, Zhi-Hua Zhou (Ed.). IJCAI, Montreal, 4392–4399. https://doi.org/10.24963/ijcai.2021/600 arXiv:2105.11266

[10] Emmanuelle Dietz, Antonis C. Kakas, and Loizos Michael. 2022. Argumentation: A calculus for Human-Centric AI. *Frontiers Artif. Intell.* 5 (2022). https://doi.org/10.3389/FRAI.2022.955579

[11] Sylvie Doutre and Jean-Guy Mailly. 2018. Constraints and changes: A survey of abstract argumentation dynamics. *Argument & Computation* 9 (2018), 223–248. https://doi.org/10.3233/AAC-180425

[12] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 38, 5 (2024), 2770–2824.

[13] Timotheus Kampik, Nico Potyka, Xiang Yin, Kristijonas Cyras, and Francesca Toni. 2024. Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. *Int. J. Approx. Reason.* 173 (2024), 109255. https://doi.org/10.1016/J.IJAR.2024.109255

[14] Timotheus Kampik, Kristijonas Čyras, and José Ruiz Alarcón. 2023. Change in Quantitative Bipolar Argumentation: Sufficient, Necessary, and Counterfactual Explanations. *International Journal of Approximate Reasoning* (2023), 109066. https://doi.org/10.1016/j.ijar.2023.109066

[15] Neema Kotonya and Francesca Toni. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *6th Workshop on Argument Mining*. 156–166.

[16] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. 2024. Contestable AI Needs Computational Argumentation. (2024). https://doi.org/10.24963/KR.2024/83

[17] Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3, 3 (2009), 225–331.

[18] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 106 (April 2021), 25 pages. https://doi.org/10.1145/3449180

[19] Paulo Maio and Nuno Silva. 2011. A Three-Layer Argumentation Framework. In *Theorie and Applications of Formal Argumentation - First International Workshop, TAFA 2011. Barcelona, Spain, July 16-17, 2011, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 7132)*, Sanjay Modgil, Nir Oren, and Francesca Toni (Eds.). Springer, 163–180. https://doi.org/10.1007/978-3-642-29184-5_11

[20] Till Mossakowski and Fabian Neuhaus. 2018. Modular semantics and characteristics for bipolar weighted argumentation graphs. *arXiv preprint arXiv:1807.06685* (2018).

[21] Martin Nöllenburg, Christian Pirker, Anna Rapberger, Stefan Woltran, and Jules Wulms. 2024. Visualizing Extensions of Argumentation Frameworks as Layered Graphs. *CoRR* abs/2409.05457 (2024). https://doi.org/10.48550/ARXIV.2409.05457 arXiv:2409.05457

[22] Nir Oren, Bruno Yun, Srdjan Vesic, and Murilo S. Baptista. 2022. Inverse Problems for Gradual Semantics. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 2719–2725. https://doi.org/10.24963/IJCAI.2022/377

[23] Nico Potyka. 2018. Continuous Dynamical Systems for Weighted Bipolar Argumentation. In *Principles of Knowledge Representation and Reasoning: 16th International Conference*, Michael Thielscher, Francesca Toni, and Frank Wolte (Eds.). 148–157.

[24] Nico Potyka. 2019. Extending Modular Semantics for Bipolar Weighted Argumentation. In *18th International Conference on Autonomous Agents and MultiAgent Systems*, Noa Agmon, Edith Elkind, Matthew E. Taylor, and Manuela Veloso (Eds.). IFAAMAS, Montreal, 1722–1730.

[25] Nico Potyka. 2021. Interpreting Neural Networks as Quantitative Argumentation Frameworks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*. AAAI Press, 6463–6470.

[26] Nico Potyka and Richard Booth. 2024. Balancing Open-Mindedness and Conservativeness in Quantitative Bipolar Argumentation (and How to Prove Semantical from Functional Properties). In *International Conference on Principles of Knowledge Representation and Reasoning (KR 2024)*.

[27] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni. 2021. Argumentative explanations for interactive recommendations. *Artificial Intelligence* 296 (2021), 103506. https://doi.org/10.1016/j.artint.2021.103506

[28] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, Chitta Baral, James P. Delgrande, and Frank Wolter (Eds.). AAAI Press, 63–73. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874

[29] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martin Pereira-Farina. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[30] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36 (2021), e5. https://doi.org/10.1017/S0269888921000011

[31] Yilin Xia, Daphne Odekerken, Shawn Bowers, and Bertram Ludäscher. 2024. Layered Visualization of Argumentation Frameworks. In *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18-20, 2024 (Frontiers in Artificial Intelligence and Applications, Vol. 388)*, Chris Reed, Matthias Thimm, and Tjitze Rienstra (Eds.). IOS Press, 373–374. https://doi.org/10.3233/FAIA240346

[32] Xiang Yin, Nico Potyka, and Francesca Toni. 2024. CE-QArg: Counterfactual Explanations for Quantitative Bipolar Argumentation Frameworks. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*. 697–707. https://doi.org/10.24963/kr.2024/66