Computational Models of Argument
C. Reed et al. (Eds.)

doi:10.3233/FAIA240322

An Empirical Study of Quantitative Bipolar Argumentation Frameworks for Truth Discovery

Nico POTYKA [1] and Richard BOOTH
Cardiff University
ORCiD ID: Nico Potyka https://orcid.org/0000-0003-1749-5233, Richard Booth https://orcid.org/0000-0002-6647-6381

Abstract. Truth discovery networks evaluate the trustworthiness of sources (e.g., websites) and their claims (e.g., the severity of a virus). Intuitively, the more trustworthy the sources of a claim, the more believable the claim and vice versa. Singleton noted that bipolar abstract argumentation could be a natural way to reason about these networks. We explain how this idea can be implemented naturally by quantitative bipolar argumentation frameworks (QBAFs) that we call TD-QBAFs. While most applications of QBAFs result in a (nearly) acyclic structure, TD-QBAFs have bi-directional edges and can feature complex cycles. The stability (convergence behaviour) of QBAFs in cyclic graphs is currently not well understood. While pathological examples of divergent QBAFs have been constructed, the problems seemed unlikely to occur in practice. However, convergence problems seem to be the rule rather than the exception for TD-QBAFs. We demonstrate how common QBAF semantics can fail to converge for very simple TD-QBAFs and discuss some of the potential causes. While this shows limitations of existing semantics, we also discuss how some previously proposed ideas can be used to mitigate the problems and demonstrate their effectiveness empirically.

Keywords. Quantitative Bipolar Argumentation, Convergence, Applications

1. Introduction

There is an increasing amount of information on the web and different actors make various claims which are often incompatible. Low-quality information sources may mistakenly provide erroneous data for topics on which they lack expertise, or malicious sources may try to deliberately deceive. Thus, trying to establish the *true* values associated to various objects, and working out which sources are *trustworthy* becomes a major concern. These two things are clearly inter-related. The more trustworthy a source is, the more believable its claims are. Conversely, a source that claims believable facts should be judged to be more trustworthy. *Truth discovery* [1] is concerned with rating and ranking a set of sources and facts given, as input, a set of claims reported by the sources. Application areas include real-time traffic navigation [2], drug side-effect discovery [3],

[1]Corresponding Author: Nico Potyka, potykan@cardiff.ac.uk.

and social sensing [4]. Algorithms for truth discovery have been based, for example, on iterative methods [5], neural networks [6,7], and voting [8].

Singleton and Booth [9,10] recently presented a general framework to compare different truth discovery approaches formally. Its central ingredient is a *truth discovery operator* that takes as input any *truth discovery network* (a graph-theoretic representation of the sources and claims) and provides as output an assignment of real numbers to the sources (intuitively interpreted as *trust scores*) and to the possible claims (*believability scores*). Singleton [11] noticed a possible way in which truth discovery networks can be interpreted as *bipolar argumentation frameworks*. Since the acceptability of a claim lends support to the statement "*s* is trustworthy", and vice versa, for every source *s* that makes that claim, sources and claims could be represented as arguments that *support* each other, while contradicting claims *attack* each other.

Singleton's idea can be naturally implemented by *quantitative* bipolar argumentation frameworks (QBAFs) [12] that we call *TD-QBAFs*. General QBAFs have various applications, but their graphical structure is often acyclic due to a clear causal relationship between the arguments. Examples include product recommendation [13], review aggregation [14], stance aggregation [15] and explaining neural networks [16]. One notable exception that features cycles are the PageRank argumentation frameworks (PRAFs) studied in [17]. PRAFs can be used to explain the PageRank of websites. The problem setting bears some resemblance to our setting, but whereas arguments (sources) in PRAFs can only support each other, arguments (sources and claims) in TD-QBAFs can both support and attack each other. To the best of our knowledge, TD-QBAFs are the first application of QBAFs that features complex cycles that contain both attack and support edges. TD-QBAFs are therefore not only an interesting application of QBAFs but also an ideal playground to study the stability of QBAF semantics.

In this paper, we take a somewhat unusual approach and start the investigation of TD-QBAFs empirically. The reason is that when we started studying TD-QBAFs, we noticed that common semantics fail to converge for a large number of examples. Since stability is at the core of computing trust values/rankings with TD-QBAFs, our focus in this paper will be on better understanding the problems and how they can be resolved. However, to evidence the practical usefulness of TD-QBAFs, we will also make some comments about formal properties and demonstrate that QBAFs will correctly identify the true claims if they are identifiable (there is a sufficient number of correct sources).

2. Background

2.1. Truth Discovery Networks (TDNs)

We start by revisiting the idea of a truth discovery network (TDN) from [9,10].

Definition 1 (TDN). *A TDN is a quadruple* $N = (\mathscr{S}, \mathscr{O}, \mathscr{D}, \mathscr{R})$ *consisting of a finite set of* sources $\mathscr{S}$*, a finite set of* objects $\mathscr{O}$*, a set* $\mathscr{D} = \{D_o\}_{o \in O}$ *of* domains *of the objects, where each* D_o *is a finite set of possible values for object o. We let* $V = \bigcup_{o \in O} D_o$*. Then* $\mathscr{R} \subseteq \mathscr{S} \times \mathscr{O} \times V$ *is a set of* reports *such that* (i) *for each* $(s,o,v) \in \mathscr{R}$*, we have* $v \in D_o$*, and* (ii) *if* $(s,o,v) \in \mathscr{R}$ *and* $(s,o,v') \in \mathscr{R}$*, then* $v = v'$*.*

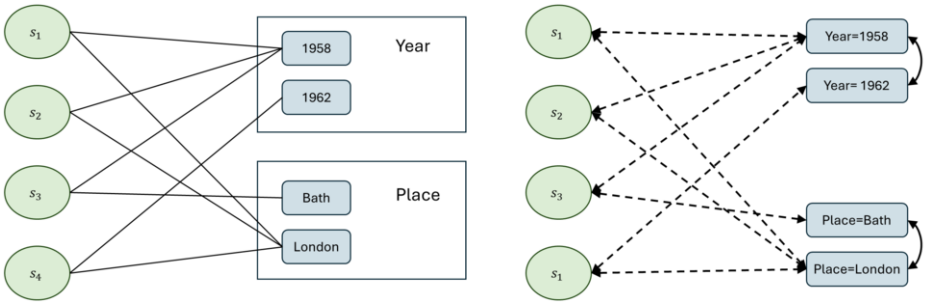

Figure 1. Example of a TDN (left) and a QBAF (right). Sources are shown in green, claims about objects in blue. The QBAF is the TD-QBAF associated with the TDN and all source and claim arguments have base scores 0.5 and 0, respectively.

A *claim* in a TDN N is a pair $c = (o, v)$, where $o \in O$ and $v \in D_o$. We write $\text{obj}(c) = o$, $\text{val}(c) = v$ in this case, and let C denote the set of all possible claims in a network N, i.e. $C = \{(o, v) \mid o \in O, v \in D_o\}$.

We can represent a TDN pictorially as a bipartite graph, with sources on one side and claims on the other, and an edge between each source and every claim it makes. For example, Figure 1 depicts a TDN with $\mathscr{S} = \{s_1, s_2, s_3, s_4\}$, $\mathscr{O} = \{Year, Place\}$, $D_{Year} = \{1958, 1962\}$, $D_{Place} = \{Bath, London\}$. $\mathscr{R}$ contains eight reports (one for each pair of source and claim), including, e.g., $(s_1, Year, 1958)$ and $(s_4, Place, London)$. We can see that s_1 and s_2 are in total agreement, while s_3 and s_4 are in total disagreement. Both s_3 and s_4 agree with s_1 and s_2 on one object but not the other.

The over-arching aim of truth discovery is to estimate the true value of each object, on the basis of the reports. To this end, we define *TD operators*, which, for any given TDN, return a real-numbered value for each source and claim in the network.

Definition 2 (TD Operator)**.** *A* TD operator *is a function* $T : \mathscr{S} \cup C \to \mathbb{R}$.

The works in [9,10] were concerned, to a large part, on defining and investigating different *axioms* which could be placed on the TD operators. These axioms were less focussed on regulating the actual output numerical *scores* of the various sources and claims in a TDN than on the *rankings* of trustworthiness (between sources) and believability (between claims) that the scores induced. For instance the *Coherence* axioms attempted to reflect the mutual interdependency between the source rankings and claim rankings (roughly, e.g., if the sources of claim c_1 are ranked higher than those of c_2 then c_1 should be ranked more believable than c_2), while *Symmetry* captured some notion of invariance of the output rankings under taking isomorphisms of the TDN (for example, in the TDN of figure 1, s_1 and s_2 play identical roles in the network and so should be ranked equally trustworthy). We refer to [9,10] for a more detailed discussion.

2.2. *Quantitative Bipolar Argumentation Frameworks (QBAFs)*

QBAFs [12,18] are abstract argumentation formalisms [19] that consider both attack and support [20] relationships between arguments. They associate every argument with a base score that can be seen as an a priori belief in the argument. The main computational problem is to assign a final strength to every argument relative to the base score and the

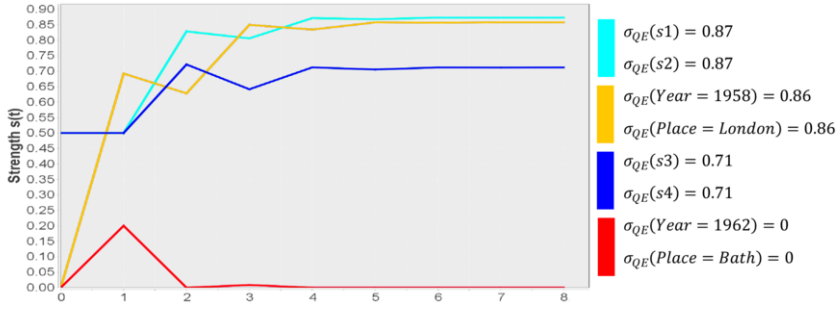

Figure 2. Evolution of strength values under quadratic energy semantics for the TD-QBAF in Figure 1. X-axis shows iteration, Y-axis shows strength values in the iteration.

final strength of attackers and supporters. In general, strength values can be from some arbitrary domain D. For concreteness, we will focus on $D = [0,1]$ here.

Definition 3 (QBAF)**.** *A QBAF is a quadruple* $Q = (\mathscr{A}, \text{Att}, \text{Sup}, \beta)$ *consisting of a set of arguments* $\mathscr{A}$*, two binary relations* Att *and* Sup *called attack and support and a function* $\beta : \mathscr{A} \to [0,1]$ *that assigns a* base score $\beta(a)$ *to every argument* $a \in \mathscr{A}$*.*

Figure 1 shows, on the right, a QBAF that encodes Singleton's intuition of the TDN on the left. We denote support relationships by dashed and attack relationships by solid edges. The base scores of the source and claim arguments are 0.5 and 0, respectively. In order to assign a final strength to the arguments, QBAF semantics commonly consider an iterative procedure. The strength of every argument is initialized with its base score. Then the following two steps are repeated until the strength values converge:

Aggregation: use an *aggregation function* to aggregate the strength values of attackers and supporters.

Influence: use an *influence function* to set the new strength to a value based on the base score and the aggregate.

The aggregation function is typically monotonically decreasing (increasing) with respect to the strength of attackers (supporters). Popular instantiations are weighted sums as used for the Euler-based [21], quadratic energy [22] and MLP-based [23] semantics or a product-based definition as used in Df-QuAD [24]. The influence function is typically monotonically increasing with respect to the base score and the aggregate. Roughly speaking, influence functions combine the base score and the aggregate such that they fall into the strength domain ($D = [0,1]$) again. For example, Df-QuAD's product aggregation function yields an aggregate between -1 and 1 while sum-based aggregation functions yield an unbounded result. To illustrate the process, Figure 2 shows the evolution of the strength values of arguments for the QBAF from Figure 1 under quadratic energy semantics. From a ranking perspective, sources 1 and 2 (light blue) are strongest followed by the claims $Year = 1958$ and $Place = London$ (yellow). The sources 3 and 4 (dark blue) are ranked lower because their support is weaker. The remaining two claims (red) are ranked last. Notably, their strength is 0. The reason for this is that their attackers (counterclaims) are stronger than their supporters (weaker sources). Hence, their final strength should not be larger than their base score, which is already 0.

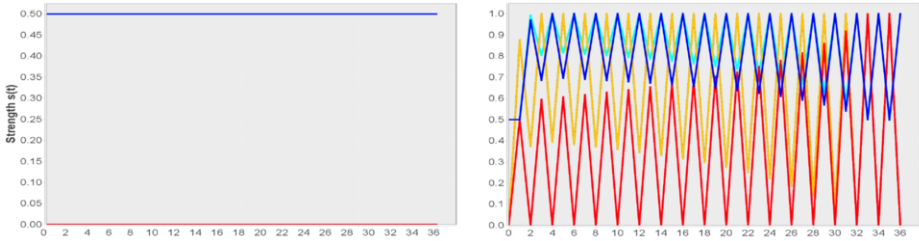

Figure 3. Evolution of strength values under Euler- and MLP-based semantics (left) and Df-Quad semantics (right) for the TD-QBAF in Figure 1.

When the update procedure of a typical semantics converges, it converges to a fixed-point of the update function (the composition of aggregation and influence function) [25,26] and the properties of semantics are usually studied by studying the properties of these fixed-points. It is interesting to note that the axioms proposed in [9,10] are closely related to the properties of QBAF semantics. For example, the *Coherence* axioms mentioned before are closely related to *Monotonicity* properties of QBAF semantics [12] and the *Symmetry* axiom corresponds to the *Anonymity* property of QBAF semantics [21]. We leave a deeper discussion of the exact relationship to future work and focus on the more fundamental question of convergence of semantics here because convergence is required to obtain well-defined strength values.

3. QBAFs for Truth Discovery (TD-QBAFs)

Singleton proposed encoding TDNs as bipolar argumentation frameworks to reason about them [11], but didn't study this idea in more detail. Here, we will extend the ideas to QBAFs to directly obtain a TD operator from the final strength values of the arguments associated with sources and claims. Singleton suggested to introduce one argument for every source and claim. For each pair of contradictory claims (they claim different values for the same object), a bi-directional attack relationship is introduced between the claims. For every report, a bi-directional support relationship is introduced between the source and the claim. In our setting, we need to assign base scores to the arguments. We assign 0.5 to sources (apriori, we are ignorant about the trustworthiness of sources) and 0 to claims (we do not believe anything without evidence).

Definition 4 (QBAF induced from a TDN)**.** *The QBAF induced from the TDN* $N = (\mathscr{S},\mathscr{O},\mathscr{D},\mathscr{R})$ *is defined as* $Q = (\mathscr{A},\text{Att},\text{Sup},\beta)$*, where* $\mathscr{A} = \mathscr{S} \cup \{(o,v) \mid \exists s \in \mathscr{S} : (s,o,v) \in \mathscr{R}\}$*,* $\text{Att} = \{(c,c') \in \mathscr{A}^2 \cap C^2 \mid \text{obj}(c) = \text{obj}(c'), \text{val}(c) \neq \text{val}(c')\}$*,* $\text{Sup} = \{(s,(o,v)),((o,v),s) \mid (s,o,v) \in \mathscr{R}\}$*.* $\beta(s) = 0.5$ *for all* $s \in \mathscr{S}$ *and* $\beta(c) = 0$ *for all* $c \in C$*.*

Figure 1 shows, on the right, the TD-QBAF associated with the TDN on the left. We already showed the strength values under quadratic energy semantics in Figure 2. Figure 3 shows, on the left, the evolution of strength values under Euler- and MLP-based semantics. Both Euler- and MLP-based semantics are unable to adjust base scores 0 or 1 and are therefore not particularly interesting for our encoding. On the right, we can see the strength values under Df-QuAD semantics. We were surprised to find that they start cycling after about 33 iterations. While [25] already demonstrated that semantics

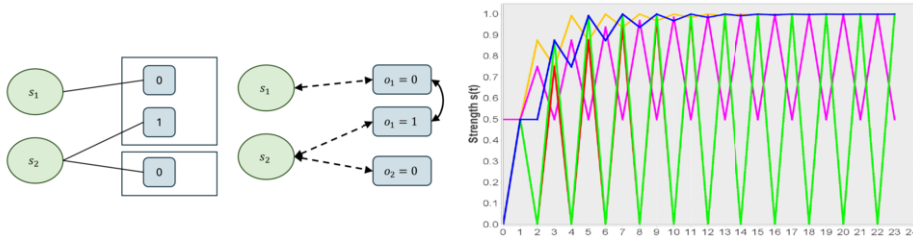

Figure 4. A TDN (left) and corresponding TD-QBAF (right) causing convergence problems for Df-QuAD.

can fail to converge, the examples were carefully designed for this purpose and have a complicated structure that is unlikely to occur in practice. The TD-QBAF in Figure 2 is relatively simple. Similar to the examples from [25], it also features a relatively large degree of symmetry, but we will see in the next section that QBAF semantics can fail to converge for even simpler examples of TD-QBAFs.

4. Stability of TD-QBAFs

When constructing simple examples of TD-QBAFs, we were surprised to find that convergence problems occured quite frequently. We conjecture that one reason is that all edges in TD-QBAFs are bi-directional, which can cause a high degree of symmetry. In Figure 4, we show the simplest example that we found that causes the Df-QuAD semantics to cycle. Notably, its graphical structure is just a bi-directional chain.

In order to evaluate how likely convergence problems are for TD-QBAFs, we created a random generator, which roughly works as follows:

1. Create $\frac{n}{2}$ sources and $\frac{n}{2}$ objects.
2. Every source has a *correctness probability* that determines the probability that one of its claims is correct. The probability is chosen uniformly at random from the interval $[0.5, 1]$.
3. For every object, we create a domain of size between 2 and 4 chosen uniformly at random. We assume that the first value is the correct value.
4. For every source, we iterate over all claims. With probability 0.5, the source will make a claim about the object. The correctness probability of the source determines the probability that the claim is correct. If the claim is not correct, the value is chosen uniformly at random from the remaining values in the domain.

To understand the convergence behaviour better, we created 100 TD-QBAFs for each $n = 10, 20, \ldots, 80$. Since the strength values typically converge quickly when they converge, we set an iteration limit of 100 iterations. It can happen that the strength values did not start cycling but simply did not converge within the iteration limit. To take account of this if the limit is reached, we take the last vector of strength values v and perform two more iterations to obtain the next two strength vectors v_1 and v_2. We call the fraction $\rho = \frac{\|v-v_1\|}{\|v-v_2\|}$ the *Divergence Ratio*. If the strength values did not cycle, we should have $\|v - v_1\| \approx \|v - v_2\|$ and ρ should be approximately 1. If the strength values started to cycle (with period 2), we should have $\|v - v_2\| \approx 0$ and ρ should be very large or even ∞ if $v = v_2$. We found that in many cases ρ was ∞ and, in all cases where the

iteration limit was reached, ρ was larger than 90, that is, $\|v - v_1\| > 90 \cdot \|v - v_2\|$, which is strong evidence that the strength values were oscillating. Since there is a chance that the oscillations fade away when the divergence ratio is smaller than ∞ and Df-QuAD failed to converge for almost all examples, we tested it again with an iteration limit of $10,000$ but obtained the same convergence percentage and the minimal divergence ratio increased to more than 10^4.

The first two sections of Table 1 show our experimental results for Df-QuAD (DfQ) and the quadratic energy model (QE). We first show the percentage of TD-QBAFs for which the strength values converged. Next, we show the mean runtime[2] that was consistently below one millisecond. It is followed by the mean number of iterations for TD-QBAFs that converged. For the non-convergent examples, we report the minimal divergence ratio (the mean divergence ratio was ∞ because there was always at least one example where the strength values cycled perfectly). Finally, to give first evidence that TD-QBAFs are useful for truth discovery, we evaluate the *correct* evaluation of claims for convergent and non-convergent cases. We say that an object is *evaluated correctly* if the strength of the true value (the first value in our experiments) is at least as high as all other values. This value has to be considered with care because the true value may be *non-identifiable*, that is, there may be too many incorrect sources. Note that since the probability that a source makes a claim about an object is 0.5, the expected number of sources for the true claim is $\frac{n}{4}$. The expected correctness probability of a source is 0.75 (because it is chosen uniformly at random from $[0.5, 1]$). Since everything is sampled uniformly at random, we assume that the probability that the true value does not receive the majority of supporters (at least $\frac{n}{8}$) resembles a binomial distribution $B(\frac{n}{4}, 0.75)$. Under this assumption, the probability that the true source is non-identifiable should be about 15% for $n = 10$ and gradually go to 0 as n increases. This explains why the correctness probability is relatively low for $n = 10$. However, we can see that it quickly goes to 1 as the size of the TD-QBAF (and thus the chance that the correct claim is identifiable) increases.

We were surprised to find that Df-QuAD did not converge for any TD-QBAF of size $n \geq 20$. Interestingly, even though it did not converge, the states (at least the termination state) between which it cycled are reasonable states in that they rank the true claim stronger than its counterclaims. The correctness for the non-converged examples for the Quadratic Energy semantics is sometimes higher than the one for the convergent examples. However, we probably shouldn't draw the conclusion that oscillating states are better than fixed-points from this observation. We believe that a more plausible explanation is that for TD-QBAFs where the true claim is identifiable, there is a higher risk of oscillations (at least in our experiments). The reader may wonder if the unstable states can be seen as reasonable alternatives similar to extensions in classical argumentation. To see that this is not the case let us look at the unstable states of Df-QuAD in Figure 4. The strength of s_1 (violet) oscillates between 0.5 and 1 and the strength of its claim $o_1 = 0$ (green) between 0 and 1. However, s_1 takes its maximum (is accepted) when $o_1 = 0$ takes its minimum (is rejected) and vice versa, while we should expect that they take their maximum/minimum simultaneously in reasonable alternatives. This is indeed the reason why the semantics does not converge (the strength values have to be adapted to make the state more plausible).

[2] All experiments ran on a Windows 11 laptop with Intel i7-13700H 2.4Ghz processor and 16 GB RAM.

		10	20	30	40	50	60	70	80
DfQ	% Converged	14	0	0	0	0	0	0	0
	Runtime (ms)	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	Iterations (C)	9.5	-	-	-	-	-	-	-
	Min. Div. Ratio	$> 10^{15}$	> 91	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^4$	$> 10^4$
	Correctness (C)	0.93	-	-	-	-	-	-	-
	Correctness (N)	0.87	0.97	0.98	0.99	0.99	0.99	0.99	0.99
QE	% Converged	92	94	96	100	95	89	81	72
	Runtime (ms)	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	Iterations (C)	18	14	13	15	16	17	17	17
	Min. Div. Ratio	132	∞	∞	-	$> 10^{15}$	333	269	247
	Correctness (C)	0.8	0.92	0.95	0.97	0.99	0.99	0.99	0.99
	Correctness (N)	0.7	0.97	0.95	-	0.99	0.99	0.99	1
DfQ(2)	% Converged	97	9	76	32	16	5	8	3
	Runtime (ms)	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	Iterations (C)	21.7	29.3	31.2	30.1	36.1	61.6	95.5	84.6
	Min. Div. Ratio	22	5	21	19	18	17	17	19
	Correctness (C)	0.81	0.91	0.95	0.98	0.99	0.98	1	0.99
	Correctness (N)	0.47	0.93	0.96	0.98	0.99	0.99	0.99	0.99
DfQ(3)	% Converged	100	100	100	100	100	100	100	100
	Runtime (ms)	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	Iterations	9.4	13.5	15.2	18.5	19.5	20.7	20.2	20.3
	Correctness	0.8	0.92	0.95	0.98	0.99	0.99	0.99	0.99
QE(2)	% Converged	100	100	100	100	100	100	100	100
	Runtime (ms)	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	Iterations (C)	9.58	12.91	11.56	10.28	10.3	10.49	10.87	11.48
	Correctness	0.8	0.92	0.95	0.98	0.99	0.99	0.99	0.99
	S-Distance	1.05	1.01	1.00	1.07	1.22	1.32	1.44	1.53
cDfQ	% Converged	100	97	86	89	92	96	100	100
	Runtime (ms)	0.84	2.4	5.79	6.66	6.49	7.81	7.3	9.26
	Iterations (C)	11.94	22.82	31.71	23.38	16.81	17.43	15.32	15.17
	Min. Div. Ratio	-	0.5	0.5	0.5	0.5	0.5	-	-
	Correctness (C)	0.8	0.92	0.95	0.98	0.99	0.99	0.99	0.99
	Correctness (N)	-	0.87	0.93	0.97	0.99	0.98	-	-
cQE	% Converged	100	100	100	100	100	100	100	100
	Runtime (ms)	0.83	0.93	1.08	1.62	2.19	3.02	3.44	4.3
	Iterations	12.03	9.02	8.35	8.38	8.45	8.74	8.85	8.84
	Correctness	0.79	0.92	0.95	0.98	0.99	0.99	0.99	0.99
	S-Distance	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 1. Statistics for trials with TD-QBAFs of increasing size (100 TD-QBAFs per size): percentage of converging TD-QBAFs, Mean Runtime, Mean number of iterations until convergence (if converged), Minimal Divergence Ratio, Mean correctness for converged (C) and non-converged (N) frameworks. For QE, we also show the mean Euclidean distance between the strength values under the original semantics and its conservative (QE(2)) and continuous counterpart (cQE) for the convergent examples. Decimals are rounded to two digits, but no digit is rounded up to 1 (so 1 means exactly 100%).

5. Improving Stability with Conservative Semantics

One way to improve the convergence guarantees of QBAF semantics is to make them more *conservative* [26]. For the modular semantics that we discussed here, this can be achieved by dividing the result of the aggregation function by a constant κ, which is called the *conservativeness* parameter [26]. Intuitively, this will make the semantics more conservative in the sense that it will hold on more strongly to the initial beliefs given by the base scores. The relevance of κ can be explained by the following result.

> [26, Corollary 3.5]: Consider a QBAF such that the indegree of every argument is at most D. Then the strength values are guaranteed to converge under
>
> - DF-Quad semantics with conservativeness $\kappa > D$,
> - quadratic energy semantics with conservativeness $\kappa > 2 \cdot D$.

Hence, in principle, for every finite QBAF, there is a semantics that guarantees convergence. However, the semantics may be very conservative. As discussed in [26], there is a tradeoff between conservativeness and *open-mindedness*, that is, a semantic's ability to move away from the initial beliefs (the base score). We will not go deeper into this discussion here because it is of less relevance when only focussing on the ranking of arguments (finding the most believable claim about an object). Let us also note that the result above is a sufficient and not a necessary condition. While a conservativeness parameter of 100 for Df-Quad guarantees that it will converge for all QBAFs with indegree less than 100, a significantly smaller parameter can be sufficient.

To understand the magnitude of the required conservativeness better, we repeated our experiments with conservativeness 2 and 3 and report the results in sections 3 - 5 in Table 1. Conservativeness 2 was sufficient for convergence of all examples under quadratic energy semantics and the number of convergent examples under Df-QuAD increased significantly. The minimal divergence ratio was also significantly smaller. We again increased the iteration limit for Df-QuAD to 10,000 to see if the non-convergent examples will converge eventually. While this increased the number of convergent examples, there was still a large number of non-convergent examples (up to 29%) and the minimal divergence ratio increased to ∞, so all non-convergent examples started indeed cycling. However, Df-Quad converges for all examples when we increase the conservativeness to 3. We can see in Table 1 that the conservative semantics still provide meaningful states in the sense that the true claims are ranked highest when they are identifiable. However, the actual strength values under conservative semantics typically differ from those under the original semantics (they tend to stay closer to the base scores). To quantify this, we also show the mean Euclidean distance between the original strength vectors and the conservative strength vectors (for examples that converged under the original semantics) in the QE(2) section (S-distance).

6. Improving Stability with Continuous Semantics

Making semantics more conservative improves stability but reduces open-mindedness. There is a less invasive method to improve stability that maintains the open-mindedness of semantics called *continuization* [22,26]. The term *continuization* is motivated by the following result: under mild conditions that are met for all semantics considered in this

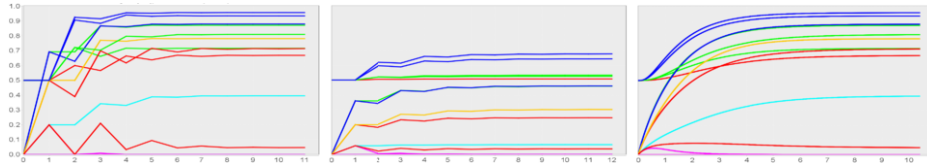

Figure 5. Discrete QE (left) vs. conservative QE(2) (middle) vs. continuous QE (right) semantics.

paper, the semantic's update function can be associated with a system of differential equations such that the fixed-points of the update function correspond to the equilibrium solutions of the system of differential equations [26, Proposition 4.1]. Since the semantical properties of QBAF semantics are studied over fixed-points, the continuized semantics satisfies the exact same properties as the original (discrete) semantics. In fact, in cases where we can prove convergence of discrete semantics, the fixed-point is unique and thus coincides with the equilibrium solution [26, Proposition 3.3 and 4.1]. What is more, the discrete update procedures that are used to compute the final strength values under discrete semantics can actually be seen as a naive algorithm to approximate an equilibrium solution of a system of differential equations [22]. This algorithm is called *Euler's method* with step size 1. In the context of differential equations, Euler's method would hardly ever be applied with step size 1 because this step size is so large that it is likely to cause stability problems. By decreasing the step size (say to 0.01), we continuize the process and improve stability. We refer to [26, Section 4] for a thorough discussion of the relationship between discrete and continuous semantics and to [22] (paragraph after Remark 1) for a discussion of alternatives to Euler's method that can find an equilibrium solution more efficiently.

To understand the behaviour of continuous semantics better, we repeated the experiments for the semantics' continuous counterparts and report the results in sections 6 and 7 of Table 1. While the continuous quadratic energy model converged for all examples, continuous Df-QuAD failed to converge within the iteration limit in some cases. However, we can see that the minimal divergence ratio is very small. We therefore increased the time limit to 500 and found that the semantics does indeed converge for all examples in less than 300 iterations. Recall that we increased the time limit for discrete Df-QuAD to 10,000 and only found that the divergence ratio increased significantly. This gives further evidence that continuization can increase the stability of QBAF semantics significantly. Indeed, all known divergence cases for discrete semantics can be solved by continuizing the semantics and there are no known examples where a continuous semantics cycles. However, it remains an open question if this is always the case. To strengthen our conjecture that continuization maintains the original semantics, we again show the mean Euclidean distance between the original strength vectors and the continuous strength vectors (for examples that converged under the original semantics) in the cQE section. As opposed to QE(2), the mean distance for cQE was consistently smaller than 0.01 evidencing that continuization indeed preserves the original semantics well. To illustrate this, we plot the evolution of strength values of the first TD-QBAF in our benchmark under discrete QE, QE(2) and continuous QE semantics in Figure 5. Even though continuization comes at a higher computational cost because it has to perform dozens of steps for each discrete step (iteration) that discrete semantics perform, the runtime still remains in the low millisecond range. As demonstrated in [22], even QBAFs with thousands of arguments can be evaluated in seconds under continuous semantics.

7. Conclusions and Future Work

Truth discovery is an interesting and natural application of QBAFs. While discrete semantics can suffer from stability problems in this setting, we can avoid these problems by using conservative and continuous semantics. All semantics seem to be able to identify the correct claims when they are identifiable. However, we believe that continuous semantics are the preferred solution for TD-QBAFs as they converge for all examples that we generated, do not require fine-tuning of the conservativeness parameter and seem to preserve the original semantics.

More analytical studies are necessary to understand the stability and the formal properties of TD-QBAFs better, but we believe that the experiments and the benchmark provided in this work are useful to enhance the understanding of TD-QBAFs and QBAFs in general. For example, [27] recently proposed using graph neural networks to approximate QBAFs, but were only able to evaluate their approach on acyclic graphs for lack of a benchmark containing cyclic QBAFs. TD-QBAFs provide a challenging and application-driven benchmark and the results for conservative and continuous semantics provide a strong baseline for evaluating the stability and semantical preservation of approximation approaches as proposed in [27]. Our benchmark and the source code for all experiments are available in the Java library Attractor[3] [28,29].

Our next step is to study TD-QBAFs more formally by analyzing which properties from [9,10] they satisfy when understanding their semantics as truth-discovery operators. It may also be interesting to look into other encodings that allow attacks of sources, for example, by adding attack relationships between sources that make contradictory claims. Furthermore, recently a number of interesting explanation approaches for QBAFs have been proposed [30,31,32,33] and it would be interesting to explore what additional insights they can give in the truth discovery setting.

References

[1] Gupta M, Han J. Heterogeneous Network-based Trust Analysis: A Survey. SIGKDD Explor Newsl. 2011;13(1):54-71. Available from: http://doi.acm.org/10.1145/2031331.2031341.

[2] Du Y, Sun YE, Huang H, Huang L, Xu H, Bao Y, et al. Bayesian Co-Clustering Truth Discovery for Mobile Crowd Sensing Systems. IEEE Transactions on Industrial Informatics. 2019:1-1.

[3] Ma F, Meng C, Xiao H, Li Q, Gao J, Su L, et al. Unsupervised Discovery of Drug Side-Effects from Heterogeneous Data Sources. In: SIGKDD 2017. KDD '17. New York, NY, USA: ACM; 2017. p. 967-76. Available from: http://doi.acm.org/10.1145/3097983.3098129.

[4] Zhang DY, Han R, Wang D, Huang C. On robust truth discovery in sparse social media sensing. In: Big Data 2016; 2016. p. 1076-81.

[5] Pasternack J, Roth D. Knowing What to Believe (when You Already Know Something). In: COLING 2010. ACL; 2010. p. 877-85.

[6] Kotonya N, Toni F. Explainable Automated Fact-Checking for Public Health Claims. In: EMNLP 2020; 2020. p. 7740-54.

[7] Marshall J, Argueta A, Wang D. A neural network approach for truth discovery in social sensing. In: MASS 2017. IEEE; 2017. p. 343-7.

[8] Elsaesser Q, Everaere P, Konieczny S. S&F: Sources and Facts Reliability Evaluation Method. In: AAMAS 2023. ACM; 2023. p. 2778-80.

[3] https://sourceforge.net/p/attractorproject/

[9] Singleton J, Booth R. Towards an axiomatic approach to truth discovery. Autonomous Agents and Multi-Agent Systems. 2022;36(2):1-49. Available from: https://doi.org/10.1007/s10458-022-09569-3.
[10] Singleton J. Trustworthiness and Expertise: Social Choice and Logic-based Perspectives [PhD thesis]. Cardiff University; 2023.
[11] Singleton J. On the Link Between Truth Discovery and Bipolar Abstract Argumentation. Online Handbook of Argumentation for AI. 2020:43-6.
[12] Baroni P, Rago A, Toni F. How Many Properties Do We Need for Gradual Argumentation? In: AAAI 2018. AAAI Press; 2018. p. 1736-43.
[13] Rago A, Cocarascu O, Toni F. Argumentation-based recommendations: Fantastic explanations and how to find them. In: IJCAI 2018; 2018. p. 1949-55.
[14] Cocarascu O, Rago A, Toni F. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In: AAMAS 2019; 2019. p. 1261-9.
[15] Kotonya N, Toni F. Gradual Argumentation Evaluation for Stance Aggregation in Automated Fake News Detection. In: Workshop on Argument Mining; 2019. p. 156-66.
[16] Ayoobi H, Potyka N, Toni F. SpArX: Sparse Argumentative Explanations for Neural Networks. In: ECAI 2023. IOS Press; 2023. p. 149-56.
[17] Albini E, Baroni P, Rago A, Toni F. PageRank as an Argumentation Semantics. In: COMMA 2020. IOS Press; 2020. p. 55-66.
[18] Amgoud L, Doder D. Gradual Semantics Accounting for Varied-Strength Attacks. In: AAMAS 2019. IFAAMAS; 2019. p. 1270-8.
[19] Dung PM. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence. 1995;77(2):321-57. Publisher: Elsevier.
[20] Cayrol C, Lagasquie-Schiex MC. Bipolarity in argumentation graphs: Towards a better understanding. International Journal of Approximate Reasoning. 2013;54(7):876-99. Publisher: Elsevier.
[21] Amgoud L, Ben-Naim J. Evaluation of arguments in weighted bipolar graphs. In: ECSQARU 2017. Springer; 2017. p. 25-35.
[22] Potyka N. Continuous dynamical systems for weighted bipolar argumentation. In: KR 2018; 2018. p. 148-57.
[23] Potyka N. Interpreting neural networks as quantitative argumentation frameworks. In: AAAI 2021; 2021. p. 6463-70.
[24] Rago A, Toni F, Aurisicchio M, Baroni P. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In: KR 2016; 2016. p. 63-73.
[25] Mossakowski T, Neuhaus F. Modular semantics and characteristics for bipolar weighted argumentation graphs. arXiv preprint arXiv:180706685. 2018.
[26] Potyka N. Extending Modular Semantics for Bipolar Weighted Argumentation. In: AAMAS 2019. IFAAMAS; 2019. p. 1722-30.
[27] Anaissy CA, Suntwal S, Surdeanu M, Vesic S. On Learning Bipolar Gradual Argumentation Semantics with Neural Networks. In: ICAART 2024. SCITEPRESS; 2024. p. 493-9.
[28] Potyka N. A Tutorial for Weighted Bipolar Argumentation with Continuous Dynamical Systems and the Java Library Attractor; 2018. NMR.
[29] Potyka N. Attractor - A Java Library for Gradual Bipolar Argumentation. In: COMMA 2022. vol. 353. IOS Press; 2022. p. 369-70.
[30] Cyras K, Kampik T, Weng Q. Dispute Trees as Explanations in Quantitative (Bipolar) Argumentation. In: ArgXAI 2022. CEUR-WS.org; 2022. p. 1-12.
[31] Yin X, Potyka N, Toni F. Argument Attribution Explanations in Quantitative Bipolar Argumentation Frameworks. In: ECAI 2023. IOS Press; 2023. p. 2898-905.
[32] Kampik T, Cyras K, Ruiz Alarcón J. Change in Quantitative Bipolar Argumentation: Sufficient, Necessary, and Counterfactual Explanations. International Journal of Approximate Reasoning. 2023:109066. Available from: https://www.sciencedirect.com/science/article/pii/S0888613X23001974.
[33] Yin X, Potyka N, Toni F. Explaining Arguments' Strength: Unveiling the Role of Attacks and Supports. In: IJCAI 2024; 2024. p. to appear.