Truth-Tracking Evaluation in Opinion-Based Argumentation

Juliete Rossie, Jérôme Delobelle, Sébastien Konieczny, Srdjan Vesic

Truth-Tracking Evaluation in Opinion-Based Argumentation

Juliete Rossie[1], Jérôme Delobelle[2], Sébastien Konieczny[1], Srdjan Vesic[1]

[1]CRIL, CNRS, Univ. Artois, Lens, France
[2]Université Paris Cité, LIPADE, F-75006 Paris, France
{rossie,konieczny,vesic}@cril.fr, jerome.delobelle@u-paris.fr

Abstract

Truth-tracking in collective reasoning systems is a core challenge in domains such as e-democracy, online deliberation, and citizen opinion polling. Our prior work introduced Opinion-Based Argumentation (OBA), a framework modeling both voting and argumentation, along with collective opinion semantics (COS) designed to select sets of arguments that are mutually coherent and aligned with agents' votes. In this paper, we first formally define the truth-tracking problem within OBA. We then introduce VAST, a comprehensive evaluation framework to systematically assess the epistemic adequacy of COS. Our empirical analysis, conducted using VAST, demonstrates substantial variation in their truth-tracking performance across diverse deliberative conditions.

GitHub — https://github.com/JulieteRossie/VAST

Introduction

In high-stakes domains such as civic technology and misinformation detection[1], AI-driven systems are increasingly tasked with a central objective of collective reasoning: identifying outcomes that correspond to a "ground truth" within a debate. This objective aligns with the classical epistemological goal of "truthlikeness"(Popper 1962). This process is called truth-tracking: the reliable recovery of a ground truth from heterogeneous and potentially unreliable votes (Hartmann and Sprenger 2012). Although some methods have been proposed to aggregate votes on arguments into collective decisions, their practical deployment is hindered by the absence of a formalization of the truth-tracking problem in argumentation and a standardized methodology for evaluating its efficacy. This gap is evident in real-world systems, e.g., platforms such as Kialo support structured argumentation and collect user votes on arguments, but function as visualization and discussion tools without implementing reasoning mechanisms. Methods for deriving collective outcomes remain largely theoretical. Empirical evaluations are limited: some focus on computational scalability (Ganzer-Ripoll et al. 2019), others test on generated argumentation frameworks (AFs) without vote profiles (Bernreiter et al. 2024), or offer axiomatic analyses without empirical validation (Rossie et al. 2024). As a result, a central question remains open: Can these methods reliably recover a ground truth under heterogeneous voting conditions?

To formally represent arguments and their interactions, numerous approaches have extended Dung's framework (Dung 1995). For instance, in Opinion Based Argumentation (OBA) (Rossie et al. 2024), agents vote on arguments, and the goal is to derive set of arguments that reflect a collective outcome, this defines Collective Opinion Semantics (COS). However, the truth-tracking efficacy of these semantics remains largely untested. Within the OBA framework, the ground truth is represented as a set of acceptable arguments and the objective is to assess how well COS can determine this truth by leveraging both the AF and the agents' votes, whose reliability may vary, over the arguments.

To investigate this question, We begin by formally defining the problem of truth-tracking in epistemic argumentation. Then, to evaluate the existing approaches, we introduce the **V**oting and **A**rgumentation **S**emantics for **T**ruth-tracking (VAST) framework, a reproducible methodology for evaluating the epistemic reliability of collective reasoning systems. VAST combines formal truth-tracking metrics with a controlled synthetic environment, enabling systematic assessment of a method's ability to recover a known ground truth under varying parameters. We conduct an empirical study of how factors, like extension-based semantics, graph types, vote reliability, or number of extensions, affect truth-tracking. We expose key epistemic limitations in existing COS methods, including labelling-based aggregation (Caminada and Pigozzi 2011), approval-based social AFs (Bernreiter et al. 2024), collective satisfaction semantics and attack removal strategies (Rossie et al. 2024), showing their strengths and weaknesses under different scenarios. Finally, we demonstrate that integrating AFs with votes significantly improves truth-tracking accuracy.

Preliminaries

An AF is a pair $\mathcal{F} = \langle \mathcal{A}r, att \rangle$, where $\mathcal{A}r$ is a finite set of arguments and $att \subseteq \mathcal{A}r \times \mathcal{A}r$ is the attack relation (Dung 1995). Extension-based semantics are used to identify sets of arguments, termed extensions, that represent collectively coherent viewpoints. These semantics are defined based on foundational properties. A set $\mathcal{E} \subseteq \mathcal{A}r$ is conflict-free if there are no $(x, y) \in att$ with $x, y \in \mathcal{E}$. An argument $x \in \mathcal{A}r$ is acceptable w.r.t. $\mathcal{E}$ if for all $y \in \mathcal{A}r$ s.t.

[1]e.g. Kialo (https://www.kialo.com/), "parlement et citoyens" (https://purpoz.com/), and Full Fact (https://fullfact.ai/)

$(y, x) \in att$, there exists $z \in \mathcal{E}$ with $(z, y) \in att$. A set $\mathcal{E}$ is: admissible if it is conflict-free and each $x \in \mathcal{E}$ is acceptable w.r.t. $\mathcal{E}$; complete (co) if it is admissible and contains all arguments acceptable w.r.t. $\mathcal{E}$; preferred (pr) if it is a $\subseteq$-maximal admissible set. We denote the set of extensions under $\sigma \in \{\texttt{co}, \texttt{pr}\}$ as $\mathcal{E}_\sigma(\mathcal{F})$. An alternative to extension semantics is the labelling approach (Caminada 2006), where each argument is assigned a label denoting acceptance, rejection, or undecided status. There exists a formal correspondence between extension-based and labelling-based semantics (Caminada 2006). We refer the reader to (Baroni, Caminada, and Giacomin 2011) for a complete overview.

Opinion Based Argumentation

OBA extends abstract argumentation by incorporating a tuple of votes representing individual stances on arguments (Rossie et al. 2024). A voter assigns to an argument one of three values: 1 (accept), -1 (reject), or 0 (abstain).

Definition 1 (Votes). *Let $\mathcal{F} = \langle \mathcal{A}r, att \rangle$ be an AF. Votes on $\mathcal{A}r$, denoted as $\mathcal{V}_{\mathcal{A}r} = \langle v_1, \ldots, v_n \rangle$, represent the system's votes. Each vote $v_i \in \mathcal{V}_{\mathcal{A}r}$ is a function $v_i : \mathcal{A}r \to \{-1, 0, 1\}$ which assigns a value for each argument in $\mathcal{F}$, indicating the voters' stances. Given $x \in \mathcal{A}r$, we denote $v^+(x) = \{v_i \in \mathcal{V}_{\mathcal{A}r} \mid v_i(x) = 1\}$, $v^o(x) = \{v_i \in \mathcal{V}_{\mathcal{A}r} \mid v_i(x) = 0\}$ and $v^-(x) = \{v_i \in \mathcal{V}_{\mathcal{A}r} \mid v_i(x) = -1\}$ the set of votes assigning 1, 0 or -1 respectively to x.*

Definition 2 (OBAF). *An Opinion Based Argumentation Framework (OBAF) is a pair $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ where $\mathcal{F} = \langle \mathcal{A}r, att \rangle$ is an AF and $\mathcal{V}_{\mathcal{A}r}$ are the votes on $\mathcal{A}r$.*

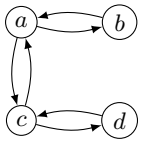
Collective Opinion Semantics

Collective Opinion Semantics (COS) refers to a family of semantics that assigns to an OBAF a set of extensions derived from the votes.

Definition 3 (COS). *Let $\mathcal{O} = \langle \langle \mathcal{A}r, att \rangle, \mathcal{V}_{\mathcal{A}r} \rangle$ be an OBAF. A COS is a function $\texttt{COS} : \mathcal{O} \to 2^{2^{\mathcal{A}r}}$.*

In the remainder of this sub-section, we present the definitions of several instantiations of COS from the literature on which we apply our evaluation framework. We focus on methods that aggregate individual opinions on arguments into collective outcomes. Works that deal with AF aggregation (Delobelle et al. 2016; Dickie et al. 2024; Chen and Endriss 2018), bipolar argumentaion (Ganzer-Ripoll et al. 2019; Irwin, Rago, and Toni 2022), quantitative argumentation (Rago and Toni 2017), or gradual semantics (Leite and Martins 2011; de Tarlé, Bonzon, and Maudet 2022) do not align with the evaluation conducted in this paper. For full descriptions, we refer the reader to the original sources. While these methods were proposed in various contexts, their truth-tracking performance is untested. Our work systematically evaluates them, addressing a key gap in the literature.

Labelling Aggregation This approach, introduced by Pigozzi and Caminada (2011), combines individual labellings into a collective labelling using skeptical (so), credulous (co), and super credulous (sco) aggregation. These labellings are post-processed into extensions via labelling-to-extension mappings ($Lab2Ext$). The COS based on this method is defined as follows, where $Vote2Lab$ is a function that transforms a vote into a labelling. We focus on $\texttt{COS}^{sco}$.



$\mathcal{V}_{\mathcal{A}r}$	a	b	c	d
$v1$	1	-1	0	-1
$v2$	1	0	0	-1
$v3$	1	-1	-1	0
$v4$	-1	0	1	-1

Figure 1: An Opinion Based Argumentation framework

Definition 4 ($\texttt{COS}^{LA}$). *Let $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ be an OBAF with $\mathcal{V}_{\mathcal{A}r} = \langle v_1, \ldots, v_n \rangle$. The COS based on labellings aggregation operator $LA \in \{so, co, sco\}$ is:*

$$\texttt{COS}^{LA}(\mathcal{O}) = \{Lab2Ext_{\mathcal{F}}(\ell) \mid \ell \in LA_{\mathcal{F}}(\{Vote2Lab(v_1), \ldots, Vote2Lab(v_n)\})\}$$

Approval Ballots Social AFs Bernreiter et al. (2024) introduce ABSAFs, where voters submit approval ballots over arguments. Formally, an ABSAF is a tuple $\mathcal{A}b = (\mathcal{F}, N, \bar{A})$, with N the set of voters and $\bar{A} = (A(i))_{i \in N}$ the approved arguments per agent i. Approval ballots specify support but leave ambiguity over unapproved arguments, which may reflect either rejection or neutrality. They use the representation score $rep_i(\mathcal{E}) = \frac{|\mathcal{E} \cap A(i)|}{|A(i)|}$, measuring the overlap between a candidate extension and a voter's approved set. Agent-level scores are aggregated using Ordered Weighted Averaging (OWA), with the utilitarian ($\vec{w}_u = (1, \ldots, 1)$) and the egalitarian ($\vec{w}_e = (1, 0, \ldots, 0)$) instances, or MaxCov, which selects extensions that fully satisfy the largest number of agents. The adaptation of the ABSAF method within COS is defined as follows:

Definition 5 ($\texttt{COS}_\sigma^{AB,rep}$). *Let $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ be an OBAF with $\mathcal{V}_{\mathcal{A}r} = \langle v_1, \ldots, v_n \rangle$ and σ be an extension-based semantics. An ABSAF is defined as $\mathcal{A}b = (\mathcal{F}, \{1, \ldots, n\}, \langle Vote2Bal(v) | v \in \mathcal{V}_{\mathcal{A}r} \rangle)$. The COS based on the representation operators rep is:*

$$\texttt{COS}_\sigma^{AB,rep}(\mathcal{O}) = \begin{cases} OWA_\sigma^{\vec{w},rep}(\mathcal{A}b) & \text{for } AB \in \{u, e\} \\ MaxCov_\sigma^{rep}(\mathcal{A}b) & \text{for } AB \in \{mc\} \end{cases}$$

Collective Satisfaction Semantics CSS selects extensions based on their agreement with the opinions expressed in a vote profile (Rossie et al. 2024). These operators involve comparing an extension $\mathcal{E}$ with a vote $v \in \mathcal{V}_{\mathcal{A}r}$. To do so, $\mathcal{E}$ is represented as a vector $Vec_{\mathcal{E}} : \mathcal{A}r \to \{-1, 1\}$, where $Vec_{\mathcal{E}}(x) = 1$ if $x \in \mathcal{E}$ and -1 otherwise. Three scoring functions are defined: $\mathcal{S}_v(\mathcal{E}) = |\{x \in \mathcal{A}r \mid v(x) = Vec_{\mathcal{E}}(x)\}|$ (to measure the agents' satisfaction with the outcome); $\mathcal{D}_v(\mathcal{E}) = -|\{x \in \mathcal{A}r \mid v(x) = -Vec_{\mathcal{E}}(x)\}|$ (idem but for the dissatisfaction); and $\mathcal{U}_v(\mathcal{E}) = \mathcal{S}_v(\mathcal{E}) + \mathcal{D}_v(\mathcal{E})$ (utility approach combining the previous two). To aggregate scores across all voters, aggregation methods $\otimes \in \{\Sigma, \min, lx\}$ (lx is leximin) are applied to a selected scoring function $\mathcal{M} \in \{\mathcal{S}, \mathcal{D}, \mathcal{U}\}$: $d^{\otimes}_{\mathcal{V}_{\mathcal{A}r}}(\mathcal{E}) = \otimes_{v \in \mathcal{V}_{\mathcal{A}r}} \mathcal{M}_v(\mathcal{E})$. Finally, CSS selects the extension(s) with maximum scores.

Definition 6 (COS). *Let $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ be an OBAF with $\mathcal{F} = \langle \mathcal{A}r, att \rangle$. Let σ be an extension-based semantics, $\otimes \in \{\Sigma, min, leximin\}$ and $\mathcal{M} \in \{\mathcal{D}, \mathcal{S}, \mathcal{U}\}$. The* COS *is* $\mathtt{COS}_{\sigma}^{\mathcal{M},\otimes}(\mathcal{O}) = \underset{\mathcal{E} \in \mathcal{E}_\sigma(\mathcal{F})}{\operatorname{argmax}}(d^{\otimes}_{\mathcal{V}_{\mathcal{A}r}}(\mathcal{E}))$.

Attack Removal Semantics The last approach, called Attack Removal Semantics (ARS) (Rossie et al. 2024), differs from the preceding ones by directly modifying the attack relation of an AF based on scores derived from the votes. First, following (Leite and Martins 2011), each argument is assigned a score using an opinion aggregation function τ_ϵ, parameterized by $\epsilon \geq 0$ where $\tau_\epsilon(x) = 0$ if $|v^+(x)| = |v^-(x)| = 0$ and $\tau_\epsilon(x) = \frac{|v^+(x)|}{|v^+(x)|+|v^-(x)|+\epsilon}$ otherwise. Then, the idea is to remove attacks from an argument whose score is lower than the score of the argument it attacks, as in preference-based frameworks (Amgoud and Cayrol 2002). Finally, an extension-based semantics is applied to the modified AF.

Definition 7 ($\mathtt{COS}^{\mathtt{AR}}$). *Let $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ be an OBAF. Let σ be an extension-based semantics and τ be an opinion aggregation function. The* COS *based on attack removal is* $\mathtt{COS}^{\mathtt{AR}}_{\sigma,\tau}(\mathcal{O}) = \mathcal{E}_\sigma(\mathcal{F}_\tau)$ *where* $\mathcal{F}_\tau = \langle \mathcal{A}r, att^* \rangle$ *is the AF associated to a triplet* $\langle \langle \mathcal{A}r, att^* \rangle, \mathcal{V}_{\mathcal{A}r}, \succeq^\tau_{\mathcal{O}} \rangle$ *with*

- $att^* = \{(x, y) \mid (x, y) \in att \text{ and } x \succeq^\tau_{\mathcal{O}} y\}$;
- $\succeq^\tau_{\mathcal{O}}$ *is the total preorder on* $\mathcal{A}r$ *such that* $x \succeq^\tau_{\mathcal{O}} y$ *iff* $\tau(x) \geq \tau(y)$.

Epistemic Evaluation Framework

Consider a criminal trial where jurors must determine which of several suspects is guilty, based on conflicting pieces of evidence. Each juror evaluates a set of arguments: e.g. the arguments in Figure 2. Under preferred semantics, this AF yields three extensions representing incompatible conclusions: suspect X is guilty ($\{a, d\}$), suspect Y is guilty ($\{b, c\}$), and no guilt inferred ($\{b, d\}$). The epistemic goal is to identify the extension corresponding to the factual state, not to model consensus or compromise. This motivates combining argumentation with voting: each juror votes on arguments based on their beliefs, and a COS is applied to approximate the ground-truth extension. The task is two-fold: (i) assess whether this approach can reliably track truth given partial, potentially noisy judgments, and (ii) evaluate the added value of the AF structure beyond the votes. This instantiates a central problem in Knowledge Representation: formal collective epistemic inference under structural constraints. Despite its relevance, existing work lacks a principled, reproducible evaluation framework, relying instead on ad hoc or non-generalizable benchmarks.

Truth-Tracking as an Epistemic Evaluation Task

In this section, we provide a concrete definition of the truth-tracking problem in argumentation and justify the framework adopted. Our evaluation framework is based on the OBAF formalism and its family of COS. This choice is motivated by three factors. First, OBAFs model voter beliefs via tri-valued votes, encompassing simpler voting models like approval ballots. Second, COS provides a uniform abstraction over aggregation methods, yielding extensions that preserve Dung-style AF structure. Third, the framework has been axiomatically analyzed, ensuring formal coherence and enabling systematic semantic comparison. Alternative frameworks, such as ABSAFs, restrict inputs to positive approval sets and lack tri-valued generality. Ganzer-Ripoll et al.'s framework (2019) operates under widely different structural constraints (e.g., restricted graph classes or bipolarity) and produces argument sets not necessarily extensions, limiting suitability for truth-tracking.

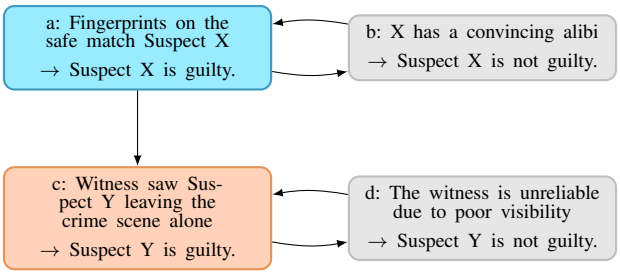


Figure 2: Example of an AF of a criminal trial – blue: X guilty, orange: Y guilty, grey: no one guilty.

We adopt an epistemic interpretation of argumentation:

Definition 8 (Truth-Tracking Task). *Given an OBAF $\mathcal{O} = \langle \mathcal{F}, \mathcal{V}_{\mathcal{A}r} \rangle$ and an extension-semantics σ:*

- $\mathcal{E}_{true} \in \mathcal{E}_\sigma(\mathcal{F})$ *is a unique ground truth extension.*
- *The truth-tracking task is to evaluate the ability of a* COS *to recover* $\mathcal{E}_{true}$ *given* $\mathcal{O}$.

The assumption of a unique ground truth $\mathcal{E}_{\text{true}}$ is consistent with standard epistemic models, which posit a single factual state of the world (Tarski 1956; Popper 1962). The assumption reflects the nature of the underlying task as epistemic (e.g. medical diagnosis, scientific assessment, or the resolution of a specific, dichotomous question) rather than normative or preference-based. The single-truth model provides a well-defined benchmark for evaluation. This assumption does not constrain the framework, as it can be extended to multi-truth scenarios by generating votes that reflect a distribution over multiple ground-truth extensions.

Baselines To assess the performance of COS methods in identifying $\mathcal{E}_{\text{true}}$, we introduce two baselines that represent opposing strategies. The first, denoted "All extensions", considers the enitre set of extensions $\mathcal{E}_\sigma(\mathcal{F})$ under the selected Dung semantics σ, independent of the vote profile $\mathcal{V}_{\mathcal{A}r}$. This baseline quantifies the performance of a non-selective strategy. It establishes a lower bound, testing whether COS yields any improvement over accepting all semantically valid outcomes. The second baseline, the Majority Rule, represents the converse strategy: it operates exclusively on the vote profile, selecting arguments supported by a majority and entirely disregarding the AF. The core objective is to evaluate the use of argumentation in the truth tracking task.

Definition 9 (Majority Rule). *Let $\mathcal{V}_{\mathcal{A}r}$ be the vector of votes on the argument $\mathcal{A}r$. The majority rule is:*
Maj$(\mathcal{V}_{\mathcal{A}r}) = \{a \mid a \in \mathcal{A}r \text{ and } |v^+(a)| > |v^-(a)|\}$

Evaluation Metrics To evaluate the epistemic adequacy of COS methods, it is necessary to evaluate their ability to recover the ground-truth extension. Let us recall that the output $\texttt{COS}(\mathcal{O})$ may be a single extension or a set of extensions, so the objective is to measure how closely this output approximates $\mathcal{E}_{\text{true}}$. We employ two metrics for this purpose: (1) Accuracy metric (AM): This metric measures exact recovery. It serves as a stringent criterion for reliability, capturing the method's ability to identify the true state unambiguously. (2) Similarity metric (SM): This metric evaluates epistemic proximity by quantifying partial overlap or argument-level similarity. SM provides a necessary alternative evaluation especially when aggregation yields multiple extensions or exact recovery is infeasible due to disagreement, but approximating $\mathcal{E}_{\text{true}}$ remains the valid epistemic goal. The epistemic validity of our metrics, AM and SM, is grounded in their ability to measure distinct but complementary aspects of truth-tracking.

Definition 10 (AM)**.** *Let $\mathcal{F}$ be an AF, σ an extension based semantics, $\mathcal{E} \subseteq \mathcal{E}_\sigma(\mathcal{F})$ and $\mathcal{E}_{true} \in \mathcal{E}_\sigma(\mathcal{F})$ the ground truth extension of $\mathcal{F}$. The accuracy metric is the evaluation function defined as follows:*

$$AM_{\mathcal{E}_{true}}(\mathcal{E}) = \begin{cases} 0 & \text{if } \mathcal{E}_{true} \notin \mathcal{E}, \\ \frac{1}{|\mathcal{E}|} & \text{otherwise.} \end{cases}$$

To instantiate the SM, we employ the double skeptical extension-aggregation function DS, which jointly considers argument inclusion and exclusion across all extensions returned by COS. Specifically, DS retains only those arguments that appear in all returned extensions and excludes those absent from all extensions, thereby producing a representative set for comparison with the ground truth.

Definition 11 (DS)**.** *Let $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_n\}$ be some extensions and $\mathcal{A}r$ the list of associated arguments. The double skeptical value of an argument $x \in \mathcal{A}r$ is defined as:*

$$DS_{\mathcal{E}}(x) = \begin{cases} 1 & \text{if } x \in \cap_{i=1}^{n} \mathcal{E}_i, \\ -1 & \text{if } x \notin \cup_{i=1}^{n} \mathcal{E}_i, \\ 0 & \text{otherwise.} \end{cases}$$

With abuse of notation, $DS_{\mathcal{E}}(\mathcal{A}r)$ represents the (vector of) arguments corresponding to the resulting extension.

Example 1. *Let us consider a set of extensions such that $\mathcal{E} = \{\{a, b\}, \{a, c\}\}$ and the ground truth extension $\mathcal{E}_{true} = \{a, d\}$, where $|\mathcal{A}r| = 4$. Then $DS_{\mathcal{E}}(\mathcal{A}r) = [1, 0, 0, -1]$.*

Similarity Sim computes alignment between two vectors for a given argument.

Definition 12 (Similarity)**.** *Let $\mathbf{e}_1 = [e_1, e_2, \ldots, e_{|\mathcal{A}r|}]$, $\mathbf{e}_2 = [e_1, e_2, \ldots, e_{|\mathcal{A}r|}] \in \{-1, 0, 1\}^{|\mathcal{A}r|}$ be two vectors. The similarity is such as:*

$$Sim_{\mathbf{e}_1,\mathbf{e}_2}(x) = \begin{cases} 1 & \text{if } \mathbf{e}_1(x) = \mathbf{e}_2(x), \\ -1 & \text{if } \mathbf{e}_1(x) = -\mathbf{e}_2(x) \text{ and } \mathbf{e}_1(x) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Using these functions, we define SM. This metric allows for a graded evaluation that rewards partial alignment when full recovery is not achieved.

Definition 13 (SM)**.** *Let $\mathcal{F} = \langle \mathcal{A}r, att \rangle$ be an AF with $\mathcal{A}r = \{a_1, \ldots, a_n\}$, σ an extension based semantics, $\mathcal{E} \subseteq \mathcal{E}_\sigma(\mathcal{F})$, and $\mathcal{E}_{true} \in \mathcal{E}_\sigma(\mathcal{F})$. The similarity metric is the evaluation function defined as follows:*
$SM_{\mathcal{E}_{true}}(\mathcal{E}) = \sum_{i=1}^{n} Sim_{Vec(\mathcal{E}_{true}), DS_{\mathcal{E}}(\mathcal{A}r)}(a_i)$.

VAST Benchmark

Our benchmark, **V**oting and **A**rgumentation **S**emantics for **T**ruth-tracking (**VAST**), consists of synthetically generated OBAFs. Its objective is to evaluate COS that combine argumentation with voting to produce extension-based outcomes. VAST enables systematic assessment of truth-tracking performance across a range of deliberative conditions. All code and data are publicly available (see abstract).

Analysis of the Existing Datasets

Ganzer-Ripoll et al. (2019) introduced a dataset of bipolar frameworks generated as directed acyclic graphs, with synthetic labelling profiles created by randomly assigning labels (in, out, undec) to each argument. The dataset was designed to evaluate the computational performance of their semantics under varying debate sizes and participation levels. However, it is not suitable for evaluating truth-tracking.

Bernreiter et al. (2024) introduced a dataset modeling voters divided into "parties" aligned with different ground truths. While foundational, it has notable limitations. First, vote generation employs a Mixed Mallows Model (MMM) that produces ballots symmetrically distributed around the ground truth with noise controlled by a dispersion parameter. This fails to represent more complex, "real-world" scenarios characterized by distinct belief factions centered on non-ground-truth alternatives. Such an assumption is inherently favorable to average aggregation methods and trivializes the truth-detection problem. Second, no explicit evaluation metric is provided. The results rely on the rep operator, which introduces bias by assessing methods using the operator underlying their design. Moreover, this operator cannot distinguish partial agreement with the ground truth, risking misleading conclusions in practical settings. Third, the dataset provides only generated AFs but not the votes. This omission hinders reproducibility, as vote generation is runtime-dependent and sensitive to parameterization. Finally, the experimental evaluation underrepresents structural variability, as it contains only AFs with exactly 10 preferred extensions generated via the Barabási-Albert model. Additionally, it only uses 50 AFs and 20 ballots per AF, a scale that may insufficiently capture scenario diversity.

AF Generation

We use three random graph models: the Barabási-Albert model (1999), the Watts-Strogatz model (1998), and the Erdös-Rényi model (1959). This choice ensures distinct structural properties, thereby enhancing the representational variety of our benchmark. For all generation models, only AFs with a minimum of two preferred extensions are retained to ensure epistemic relevance. The benchmark is based on synthetically generated data, an approach consistent with standards in argumentation research and compe-

titions[2]. This methodology is essential for enabling controlled, reproducible evaluation. Furthermore, this approach is necessitated by the fact that real-world datasets with a known, objective ground-truth extension do not currently exist. The framework is, however, modular and designed to incorporate such real-world AFs and vote profiles as they become available. In VAST, AFs are generated using AFBenchGen2 (Cerutti, Giacomin, and Vallati 2016) for the Barabási-Albert and Watts-Strogatz models, as well as the NetworkX[3] implementation of the Erdös-Rényi model. We generated 726 AFs which includes 300 Barabási-Albert (BA) and 300 Erdös-Rényi (ER) AFs using parameters like the number of arguments $n_a \in \{5, 10, 15, 20, 30, 40\}$, the probability of cycles and of attacks p_c, $p_a \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ (10 AFs per parameter pair). For the Watts-Strogatz model (WS), we generate 126 AFs with $n_a \in \{5, 10\}$, $p_c, p_r \in \{0.2, 0.5, 0.8\}$, and $k \in \{2, 4, \ldots, |\mathcal{A}r| - 1\}$. We refer the reader to (Cerutti, Giacomin, and Vallati 2016) for the meaning of the parameters.

Vote Generation

We now proceed to construct the OBAFs using the previously built AFs. This requires two steps: (1) determining the ground truth, and (2) generating a set of votes that are more or less aligned with the ground truth.
(1) Given an AF $\mathcal{F} = \langle \mathcal{A}r, att \rangle$, the ground truth $\mathcal{E}_{\text{true}}$ is randomly selected from the set of extensions of $\mathcal{F}$ w.r.t. a extension-based semantics $\sigma \in \{\texttt{co}, \texttt{pr}\}$. This models a scenario in which one of the admissible positions is, in fact, the correct one, while avoiding bias toward any particular extension. This is why our AFs have at least two extensions.
(2) Regarding the generation of votes, we impose two restrictions for standardization and comparability. First, we exclude abstentions (i.e., vote with value 0), yielding only acceptance and rejection. This avoids disadvantaging methods like $\texttt{COS}^{AB}$, given the ambiguity of non-accepted arguments, which may indicate rejection or neutrality. Second, we relax consistency constraints, allowing agents to accept conflicting arguments. This choice simplifies generation and reflects realistic cognitive inconsistencies in collective settings. We first define the individual reliability $\rho_i \in [0, 100]$ of a vote as its alignment with the ground truth $\mathcal{E}_{\text{true}}$, measured by the normalized similarity score (Definition 12). Generating a single vote is equivalent to uniformly sampling at random from the set of all possible votes matching a ρ score. The internal sampling procedure is one any selecting votes following these rules: If $\rho_i = 100$, the vote matches $\mathcal{E}_{\text{true}}$; if $\rho_i = 0$, it is fully opposed. For intermediate values, ρ_i is mapped to a target similarity score $= \lfloor (2 \times |\mathcal{A}r| \times \frac{\rho}{100}) - |\mathcal{A}r| \rfloor$. For our experiments, we generate groups of 100 votes for each of the 726 AFs. Each group is defined by a target mean reliability $\rho \in \{0, 10, \ldots, 90, 100\}$. To construct a group, votes are generated with varied ρ_i such that the average reliability of the 100 votes in the group is exactly the target mean ρ. We test both preferred and complete semantics, resulting in a total of 15,972 OBAFs. This vote generation model is distinct from standard noise models (e.g., MMM), which typically inject noise into $\mathcal{E}_{\text{true}}$ via a function. Our method constructs a population of votes with a known, controlled average reliability. This approach enables reliability-driven deviations, ensuring that low-reliability votes are systematically distant rather than merely random, and provides precise control over the aggregate disagreement.

[2]https://www.argumentationcompetition.org/

[3]NetworkX is a widely used Python library for the creation, manipulation, analysis and study of graph-based data.

Results

In this section, we report the results of evaluating opinion semantics using the VAST benchmark. Performance is assessed using the AM and SM metrics (see Definitions 10 and 13). Results are shown in Figure 3, where shaded regions denote 95% confidence intervals. These intervals are computed via non-parametric bootstrapping by resampling the data with replacement, and reflect the variability of each method's performance across experimental conditions.

Collective Opinion Semantics Figure 3(a) reports performance under the preferred semantics across 7,986 OBAFs, with vote reliability $\rho \in \{40, 50, \ldots, 100\}$. As expected, most methods show improved performance as ρ increases. The "All extensions" baseline yields approximately 30% accuracy (AM), and serves as a naive reference. $\texttt{CSS}^{\mathcal{U},\Sigma}_{\texttt{pr}}$ and $\texttt{COS}^{\texttt{AR}}_{\texttt{pr}}$ exhibit the fastest increase, reaching near-perfect scores from $\rho \geq 70\%$. $\texttt{CSS}^{\mathcal{U},lx}_{\texttt{pr}}$ follows closely but requires slightly higher ρ to converge. $\texttt{COS}^{AB,rep}_{\texttt{pr}}$, for $AB \in \{u, e\}$, improve more gradually and require higher ρ for optimal accuracy. In contrast, $\texttt{COS}^{sco}$ remains low across all non-maximal ρ and converges only at $\rho = 100\%$. The plot in Figure 3(c) uses the same parameters as Figure 3(a) but evaluates performance using the SM metric instead of AM. The same performance trends are observed, though methods tend to achieve higher scores with the SM metric. Finally, the plot in Figure 3(b) uses the complete semantics instead of the preferred semantics. All COS exhibit higher variability with the complete semantics. $\texttt{CSS}^{\mathcal{U},\Sigma}_{\texttt{co}}$ and $\texttt{CSS}^{\mathcal{U},lx}_{\texttt{co}}$ maintain their strong performance, achieving steep improvements for $\rho \geq 60$ % and stabilizing at 98 % rather than 100%. In contrast, other methods exhibit significant performance degradation when switching from preferred to complete semantics. They exhibit a 40 point drop in performance even at high ρ.

OBAF Parameters Figure 3(d) shows the impact of different graph types in the VAST benchmark under preferred semantics. For clarity, the plot includes three representative COS. The comparison reveals two distinct patterns. The first, illustrated by $\texttt{COS}^{\texttt{AR}}_{\texttt{pr}}$, exhibits low variability and stable performance across graph types, a trend also observed for utilitarian methods such as $\texttt{CSS}^{\mathcal{U},\Sigma}_{\texttt{pr}}$. In contrast, the second pattern, exemplified by $\texttt{CSS}^{\mathcal{U},lx}_{\texttt{pr}}$, displays high variability with performance varying significantly across graph types; this behavior is also found in some other methods like $\texttt{COS}^{e,rep}_{\texttt{pr}}$. Finally, $\texttt{COS}^{u,rep}_{\texttt{pr}}$ combines both patterns, with low variability between ER and WS graphs and high variability relative to BA. In addition to graph type, Figure 3(e) shows

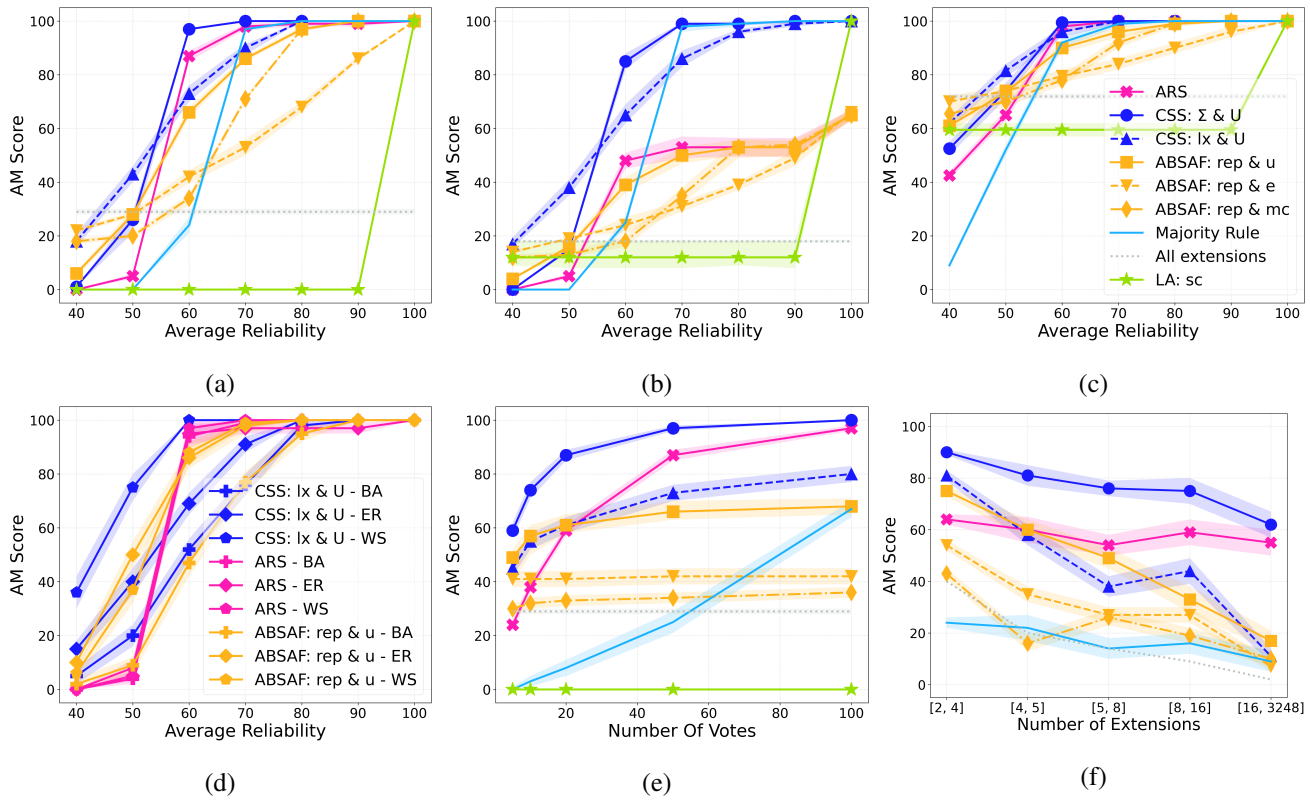


Figure 3: All plots are based on 7,986 OBAFs. (a) uses preferred semantics and reports AM scores; (b) uses complete semantics and reports AM scores; (c) uses preferred semantics ($\sigma = \texttt{pr}$) with SM scores (+legend). (d) compares graph types for CSS, ARS, and ABSAF under $\sigma = \texttt{pr}$. (e)–(f) vary, respectively, the number of votes and extensions, all with $\rho = 60\%$ and $\sigma = \texttt{pr}$.

the effect of the number of votes on the performance of different COS, i.e., the impact of varying the number of votes ($n_v \in \{5, 10, 20, 50, 100\}$) at a fixed reliability level $\rho = 60\%$. The plot reveals a steep performance increase for $\texttt{COS}^{\texttt{AR}}_{\texttt{pr}}$ and $\texttt{CSS}^{\mathcal{U},\Sigma}_{\texttt{pr}}$, with scores rising from 50% at $n_v = 5$ to nearly 90% at $n_v = 20$. Other methods show a slower rate of improvement as n_v increases. Finally, Figure 3(f) shows the effect of a high number of extensions and arguments on opinion semantics performance. $\texttt{CSS}^{\mathcal{U},\Sigma}_{\texttt{pr}}$ and $\texttt{COS}^{\texttt{AR}}_{\texttt{pr}}$ maintain stable performance, with a maximum decrease of 10 points. In contrast, other semantics exhibit an average performance decline of 40 points as the number of extensions increases.

Discussion

Graph Famillies In Figure 3(d), we observe that the sensitivity of COS to different graph types varies significantly. Notably, $\texttt{COS}^{\texttt{AR}}$ and utilitarian-based methods (like $\texttt{CSS}^{\mathcal{U},\Sigma}$ and $\texttt{COS}^{u,rep}$) exhibit robustness across graph families, whereas other methods demonstrate considerable sensitivity. Our experimental results indicate that AFs generated by the WS and ER models yield comparable outcomes when considering the entire dataset. This similarity arises from the fact that both algorithms generate AFs with a moderate number of extensions, even as the argument count increases. The performance degradation of egalitarian methods on larger AFs stems from their tendency to return more extensions as ρ decreases. Since the AM penalizes multiple extensions, this reduces their performance. Consequently, egalitarian methods perform better on WS AFs, which have the lowest average number of extensions among the VAST benchmark sampling algorithms. To further investigate, we conducted an experiment on BA AFs with $n_a = 5$, where the mean number of extensions is limited to 3.8. The results confirmed our initial findings, with egalitarian methods again demonstrating high effectiveness. Based on these observations, the egalitarian methods, particularly $\texttt{CSS}^{\mathcal{U},lx}$, outperform other methods when the number of extensions does not exceed 10. Whereas in scenarios where robustness to different graph types is needed, $\texttt{COS}^{\texttt{AR}}$ and $\texttt{CSS}^{\mathcal{U},\Sigma}$ outperform the other methods. The utilitarian-leaning COS exhibit this robustness because they are inherently more selective and tend to converge on a smaller set of consensus extensions, which gives them a significant advantage, especially with AM.

Extensions Semantics Comparing Figures 3(a) and 3(b) reveals significant variability in COS robustness across the considered extension semantics. CSS methods consistently achieve near-perfect scores in both cases, demonstrating stability regardless of the semantics. In contrast, other methods show substantial performance degradation when switching from preferred to complete semantics. We chose the pr and co semantics due to their widespread use and comple-

mentary insights. pr typically produce fewer, more selective extensions, whereas co allow extensions to be subsets of one another. This difference challenges both $\texttt{COS}^{AB}$ and $\texttt{COS}^{\texttt{AR}}$. While these methods achieve near-optimal performance (around 100%) under pr with $\rho \geq 80\%$, their scores drop significantly to approximately 70% with co (even at $\rho = 100\%$). The issue is more pronounced for BA AFs, where scores fall to around 40% due to the higher number of extensions produced. The primary difficulty for $\texttt{COS}^{AB}$ stems from the definition of the rep operator. The score is computed as the cardinality of the intersection between the vote and the extension divided by the cardinality of the vote. Thus, partial agreement with the ground truth can yield a high score, causing confusion between subsets and supersets of the actual ground truth. Similarly, $\texttt{COS}^{\texttt{AR}}$ experiences performance reduction with co. This is explained by $\texttt{COS}^{\texttt{AR}}$ modifying the original AF through attack removal, which can hinder correct ground truth identification, especially when the ground truth is a subset of another co extension. The effectiveness of CSS stems from a two-fold advantage: scoring functions and aggregation functions. First, the scoring functions, particularly utility, capture agreement, disagreement, and neutrality, thereby effectively representing fully the information contained in the votes. This comprehensive capture of opinion facilitates faster convergence to the ground truth, especially when the group reliability exceeds 50%. Second, the aggregation functions allow for flexible adaptation to different scenarios. Specifically, the sum and leximin operators offer complementary strengths: while leximin is highly effective even at low reliability when the number of extensions is limited, sum rapidly converges to the ground truth when reliability reaches 60% or higher. This versatility makes CSS robust across a range of contexts.

Collective Opinion Semantics One nuanced difference between aggregation metrics arises from AM and SM scores for egalitarian methods. The SM score appears more lenient toward indecision while being stricter on mismatched arguments, which explains why egalitarian methods tend to perform slightly better with SM compared to AM. This difference highlights the importance of selecting an appropriate metric when evaluating methods that inherently favor balanced extension distribution. Within the CSS methods, $\texttt{CSS}^{\mathcal{U},\Sigma}$ exhibits rapid performance improvement for $\rho \in \{50, 60\}$. Conversely, $\texttt{CSS}^{\mathcal{U},lx}$ outperforms $\texttt{CSS}^{\mathcal{U},\Sigma}$ when $\rho \leq 50\%$, after which $\texttt{CSS}^{\mathcal{U},\Sigma}$ surpasses it and attains the highest accuracy. The stable performance of $\texttt{CSS}^{\mathcal{U},lx}$ across ρ levels suggests robustness, especially under moderate ρ conditions. Similarly within $\texttt{COS}^{AB}$ methods, $\texttt{COS}^{u,rep}$ consistently outperform the egalitarian variants, $\texttt{COS}^{e,rep}$, for $\rho \geq 50\%$.Utilitarian COS methods capitalize on the majority signal, exhibiting rapid performance gains when $\rho > 50\%$. Conversely, egalitarian COS methods excel in low-reliability scenarios by minimizing worst-case disagreement. $\texttt{COS}^{mc,rep}$ performs poorly at low reliability but converges with other $\texttt{COS}^{AB,rep}$ for $\rho \geq 70\%$. Methods based on labelling aggregation (e.g. $\texttt{COS}^{sco}$) consistently underperform across almost all reliability levels. Thy exhibit an inability to resolve disagreement except under unanimous votes, which is unrealistic in practice. This limitation reduces their ability to track $\mathcal{E}_{\text{true}}$ when opinions conflict.

On the Usefulness of Argumentation Figure 3(d) shows a correlation between the number of votes and the truth-tracking performance of COS. The majority rule baseline exhibits a substantial increase in accuracy, starting from a score of 0% when the number of votes is 5 and reaching approximately 70% when it reaches 100. This significant rise indicates that even when $\rho = 60\%$, the accumulation of votes alone can substantially enhance accuracy. Any methods below this threshold do not adequatly use the AFs in addition to the votes. In the same plot, $\texttt{CSS}^{\mathcal{U},\Sigma}$ increases from 60% to 100% as the number of votes grows, while $\texttt{COS}^{\texttt{AR}}$ rises from approximately 25% to 98%. The contrast with the majority rule baseline is notable: although the majority rule reaches 70% at 100 votes, methods integrating AFs, such as $\texttt{CSS}^{\mathcal{U},\Sigma}$, achieve around 60% accuracy even with only 5 votes. This demonstrates the benefit of combining AF structure with vote aggregation to improve truth-tracking, especially when votes are few. Moreover, argumentation contributes significantly when vote quality is limited; for instance, at $\rho = 60\%$ in Figure 3(a), the majority rule attains 25% accuracy, whereas $\texttt{CSS}^{\mathcal{U},\Sigma}$ exceeds 95%. Additionally, these results show that both $\texttt{CSS}^{\mathcal{U},\Sigma}$ and $\texttt{COS}^{\texttt{AR}}$ effectively leverage the increasing number of votes, significantly outperforming other methods at the given reliability level. Their variability decreases as vote count rises, indicating improved stability in high-vote scenarios. In contrast, methods such as $\texttt{COS}^{AB}$ and $\texttt{CSS}^{\mathcal{U},lx}$ maintain more stable performance across vote counts but do not benefit from the larger voting pool, exhibiting relatively consistent yet lower AM scores. This limited scalability suggests reduced sensitivity to additional votes. Finally, most methods exhibit significant performance improvements only beyond a reliability level of 60% (Figure 3(a)). This shows that low to moderate reliability is insufficient for accurate truth-tracking, highlighting the need for both vote quantity and quality.

Conclusion

In this paper, we formalize the problem of truth tracking in argumentation, more specifically within the OBA framework and collective opinion semantics to determine the accepted arguments. We assess how reliably these semantics recover the ground-truth under heterogeneous votes. To this end, we introduced VAST, a reproducible framework that combines formal truth-tracking metrics with controlled generation of AFs and vote profiles. This enabled a systematic evaluation of multiple COS methods across semantics, graph types, voter reliabilities, and group sizes. The results show that argumentation-based aggregation outperforms direct voting under noisy or sparse conditions, and that utilitarian semantics are the most consistently robust. This work establishes a reproducible foundation for empirical research on epistemic adequacy in collective reasoning systems, supporting the design of trustworthy deliberative AI and future extensions to real-world decision-making contexts.

Acknowledgments

This work has benefited from the support of the ANR AGGREEY Project (ANR-22-CE23-0005) and the AI Chair BE4musIA of the French National Research Agency (ANR-20-CHIA-0028).

References

Amgoud, L.; and Cayrol, C. 2002. A Reasoning Model Based on the Production of Acceptable Arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3): 197–215.

Barabási, A.-L.; and Albert, R. 1999. Emergence of scaling in random networks. *Science*, 286(5439): 509–512.

Baroni, P.; Caminada, M.; and Giacomin, M. 2011. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4): 365–410.

Bernreiter, M.; Maly, J.; Nardi, O.; and Woltran, S. 2024. Combining Voting and Abstract Argumentation to Understand Online Discussions. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS'24)*, 170–179.

Caminada, M. 2006. On the Issue of Reinstatement in Argumentation. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA'06)*, 111–123. Springer.

Caminada, M.; and Pigozzi, G. 2011. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1): 64–102.

Cerutti, F.; Giacomin, M.; and Vallati, M. 2016. Generating structured argumentation frameworks: Afbenchgen2. In *In Proceedings of the Conference on Computational Models of Argument (COMMA' 16)*, 467–468.

Chen, W.; and Endriss, U. 2018. Aggregating Alternative Extensions of Abstract Argumentation Frameworks: Preservation Results for Quota Rules. In *Computational Models of Argument (COMMA)*, volume 305 of *Frontiers in Artificial Intelligence and Applications*, 425–436. IOS Press.

de Tarlé, L. D.; Bonzon, E.; and Maudet, N. 2022. Multi-agent Dynamics of Gradual Argumentation Semantics. In Faliszewski, P.; Mascardi, V.; Pelachaud, C.; and Taylor, M. E., eds., *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022)*, 363–371. Auckland (online), New Zealand: International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

Delobelle, J.; Haret, A.; Konieczny, S.; Mailly, J.-G.; Rossit, J.; and Woltran, S. 2016. Merging of Abstract Argumentation Frameworks. In *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (KR'16)*, 33–42.

Dickie, C.; Lauren, S.; Belardinelli, F.; Rago, A.; and Toni, F. 2024. Aggregating bipolar opinions through bipolar assumption-based argumentation. *Autonomous Agents and Multi-Agent Systems*, 39(1).

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2): 321–357.

Erdős, P.; and Rényi, A. 1959. On Random Graphs. I. In *Proceedings of Publ. Math. Debrecen*, volume 6, 290–297.

Ganzer-Ripoll, J.; Criado, N.; Lopez-Sanchez, M.; et al. 2019. Combining Social Choice Theory and Argumentation: Enabling Collective Decision Making. *Group Decision and Negotiation*, 28: 127–173.

Hartmann, S.; and Sprenger, J. 2012. Judgment aggregation and the problem of tracking the truth. *Synthese*, 187(1): 209–221.

Irwin, B.; Rago, A.; and Toni, F. 2022. Forecasting Argumentation Frameworks. 533–543.

Leite, J.; and Martins, J. 2011. Social Abstract Argumentation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*, 2287–2292.

Popper, K. R. 1962. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, England: Routledge.

Rago, A.; and Toni, F. 2017. Quantitative Argumentation Debates with Votes for Opinion Polling. In An, B.; Bazzan, A.; Leite, J.; Villata, S.; and van der Torre, L., eds., *PRIMA 2017: Principles and Practice of Multi-Agent Systems*, 369–385. Cham: Springer International Publishing. ISBN 978-3-319-69131-2.

Rossie, J.; Delobelle, J.; Konieczny, S.; Lens, C.; and Vesic, S. 2024. Collective Satisfaction Semantics for Opinion Based Argumentation. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning (KR' 24)*, 631–641.

Tarski, A. 1956. The Concept of Truth in Formalized Languages. In Tarski, A., ed., *Logic, semantics, metamathematics*, 152–278. Clarendon Press.

Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393: 440–442.