Reassessing the Impact of Reading Behaviour in Online Debates Under the Lens of Gradual Semantics

Jordan Thieyre[1,*], Aurélie Beynier[1], Nicolas Maudet[1] and Srdjan Vesic[2]

[1] *LIP6 - CNRS, Sorbonne Université, 4 place Jussieu, F-75005 Paris, France*
[2] *CRIL - CNRS - Univ. Artois*

Abstract

While it is unrealistic to assume users of online debate platforms to read and interpret all the arguments available, it is important to understand how positions will emerge on the basis of a fraction of those arguments. What arguments exactly will be accessed by users depend on assumptions on the platform design or on the readers' behaviours. Young et al. were the first to explore this question and report results in the context of an underlying extension-based semantics. We undertake a similar study in the context of gradual semantics, using a more comprehensive set of metrics, testing a larger number of behaviours, and come to different conclusions. We show in particular that a reading behaviour balancing supports and attacks provides interesting results.

Keywords

Online debates, reading behaviour, bipolar gradual argumentation frameworks

1. Introduction

The Internet allows people to express their opinions by participating in online discussions, sometimes involving many users and comments. These debates can provide a wealth of relevant information for users curious about the subjects under discussion. The main obstacle that such users may encounter is the sheer volume of comments exchanged, making it difficult for a human being to read all the arguments of a debate and to assess the relevance of each point of view in a reasonable amount of time. Indeed, under time constraints, the reading behaviour, i.e. the way users read the arguments of the debate, affects the subset of arguments they are exposed to, and consequently, their assessment of the acceptability or strength of those arguments. In this paper, we explore several "natural" reading behaviours and experimentally investigate how these behaviours influence the user's perception of a debate.

Online debates often take the form of an initial topic, to which many participants have responded with comments, comments which themselves have responses, and so on. These debates can therefore be represented in the form of graphs, more specifically trees, where the nodes are the comments and the edges are the attack or support relationships between two

SAFA'24: Fifth International Workshop on Systems and Algorithms for Formal Argumentation, September 17, 2024, Hagen, Germany

[*] Corresponding author.

jordan.thieyre@lip6.fr (J. Thieyre); aurelie.beynier@lip6.fr (A. Beynier); nicolas.maudet@lip6.fr (N. Maudet); vesic@cril.fr (S. Vesic)
0009-0006-8975-8169 (J. Thieyre); 0000-0002-3812-6755 (A. Beynier); 0000-0002-4232-069X (N. Maudet); 0000-0002-4382-0928 (S. Vesic)

comments. In short, these debates are well suited to the use of argumentation theory. In this paper, we focus on the Kialo platform[1] and we note that the number of arguments can be high. It is unlikely that an average user will have the time to read all the arguments.

The question investigated here is to study how the order in which the arguments are considered can affect the view readers have about the acceptability (as defined in the extension-based semantics introduced by Dung [1]) of the most central arguments in an online debate. In the first attempt to answer this question, Young et al. [2] presented a study of "comment sorting policies"[2] on Kialo data to determine how the order of presentation of arguments dynamically affects the acceptability status. As Kialo debates contain attacks and supports, Young et al. use a flattening[3] approach, borrowed from Cayrol and Lagasquie-Schiex [3], to transform scraped debates into attack-only argumentation graphs in order to readily use extension-based semantics. Then, for each debate, for each of the four reading behaviour they have introduced, they compute the grounded extension of the debate graph and the grounded extension of the n first arguments read following the considered reading behaviour. Young et al. then calculate a distance between the two obtained extensions.

We take inspiration from this study, but we depart from the methodology used by Young et al. for the following reasons. First, the flattening approach they use might lead to information loss. Let us illustrate this on an example with x being the central proposition attacked by one argument a and supported by two arguments b and c. The algorithm used by Young et al. to flatten this graph results here in ignoring the supports (they will be deleted). In our work, we propose a completely distinct way to take into account the supports (namely by directly using bipolar gradual semantics, and not a flattening which is inappropriate in the case of online debates) to take them into account properly.

Another issue with the work of Young et al. is that they use grounded semantics. As previously emphasized, extension-based semantics suffer from a lack of smoothness [4]. It is undesirable if a single argument results in drastic changes in the acceptability degrees. This *Christmas tree* behaviour where arguments suddenly change their acceptability status when adding a new argument is not intuitive in the context of online debates where one needs more robustness. As a consequence, even though we borrow the methodology proposed by Young et al. [2], we use a model we believe to be more adapted, namely that of bipolar gradual argumentation semantics. It addresses both problems mentioned above. We use several state-of-the-art gradual semantics to conduct our experiments. Besides these key differences, our approach departs from the work of Young et al. in the following manner: (1) we use a more diverse set of metrics to assess our results; (2) we focus on the evaluation of the most central arguments (as opposed to all the arguments of the debate); and (3) we introduce and study many more reading behaviours.

Recently, behavioural studies have become more popular as a way to assess formal argumentation frameworks [5]. Most of these studies focus on the reasoning perspective, testing

[1] www.kialo.com

[2] Young et al. study "comment sorting policies", but we take a more user-oriented perspective and talk about "reading behaviours", thus acknowledging that the order in which arguments are read is not only determined by the designer's choice of how they are presented on the platform. In practice both aspects interplay, but we keep things simple in the context of this paper.

[3] Flattening is the procedure used to delete supports, preferences or other data and obtain a vanilla argumentation graph that is somehow equivalent to the initial graph.

in particular how argumentation semantics match observed human behaviour. On the other hand, the outcomes of gradual semantics have been compared [6], but only using synthetic (and complete) data. Finally, incomplete argumentation frameworks have been studied [7, 8], but from a theoretical perspective and in the extension-based framework. We stress again that we do not take the (best case) perspective of a designer trying to optimize the sequence of arguments to be read, or the (worst case) adversarial perspective of a reader who could manipulate the system. Instead, we study "natural" reading behaviours to see how they affect some global metrics evaluating the distance to some ground-truth under complete information.

The remainder of this paper is as follows. In Section 2 we present Kialo and the data. Section 3 provides the background on gradual semantics. We detail the reading behaviours and the metrics used in Section 4 and Section 5. The experimental results are reported in Section 6.

The full code together with a notebook allowing to explore many other parameters, semantics and reading behaviours is available on our Git[4] repository.

2. The Kialo Dataset

2.1. Presentation of Kialo

Kialo is an online platform for structured debates. When a user initiates a debate, she puts an initial claim which stands for the central question. Users can then add claims that respond directly to the question if it is closed-ended, or to one of the existing alternatives if the question is open-ended, or to another claim already in the discussion. Claims are classified as "PRO" and "CON" depending on whether they support or attack the claim to which they are attached. The moderation of the platform allows moderators to rephrase claims, to move or merge claims or to delete claims that do not fulfill the requirements of the platform. Finally, users can vote on the claims by choosing a score from 0 to 4 (integers only).

Example 1 (A debate of Kialo). *One of Kialo's biggest debates (in terms of number of claims) is "The Ethics of Eating Animals: Is Eating Meat Wrong?". The central proposition debated is "Humans should stop eating animal meat." An example of a "PRO" claim for this proposition is "Eating meat, in the majority of cases, involves the cruel and immoral treatment of animals." An example of a "CON" claim is "The taste of meat is delicious and brings many people pleasure in a manner that vegetarian food cannot fully imitate." This claim received 185 votes, distributed as follows: 66 users voted for 0, 31 users for 1, 26 users for 2, 17 users for 3, and 45 users for 4.*

Kialo debates are "clean" thanks to the moderation carried out by users: there are no insults, claims must be concise, be based on logic and facts, and present a single point related to the theme of the debate. Claims can therefore be considered as arguments. Furthermore, claims must not duplicate other claims. Finally, Kialo has been designed in such a way that the most general claims are the closest to the central question, allowing users to dive into more details as they move down the graph.

[4] https://gitlab.com/jthieyre/reassessing-the-impact-of-reading-behaviour-in-online-debates-under-the-lens-of-gradual-semantics

2.2. Scraping and Resulting Dataset

Firstly, we used Kialo's API (Application Programming Interface) to retrieve the list of public debate IDs. These IDs are then used to ask the API for the data relating to the debates, i.e. the ID, the text, the votes, the date of creation and of last modification of the claim, and the ID of its author; the ID, the type of relation (attack or support), the date of creation and of last modification of the relation between this claim and another, and the ID of its author. Of this large amount of data, we have only kept the IDs, dates and votes of the claims, and the dates and relationships between claims with their types, to protect the privacy of users as much as possible. In total, we recovered data from 2,959 public debates. We analysed the data collected and observed that a large proportion of the data in each debate corresponded to "archived" data, in other words, outdated claims or relationships. Using a depth-first search (DFS) algorithm, we were able to isolate the claims that corresponded to those displayed by the Kialo front-end from those that are archived. The outcome is 377,182 claims spread over 2,959 cleaned debates.

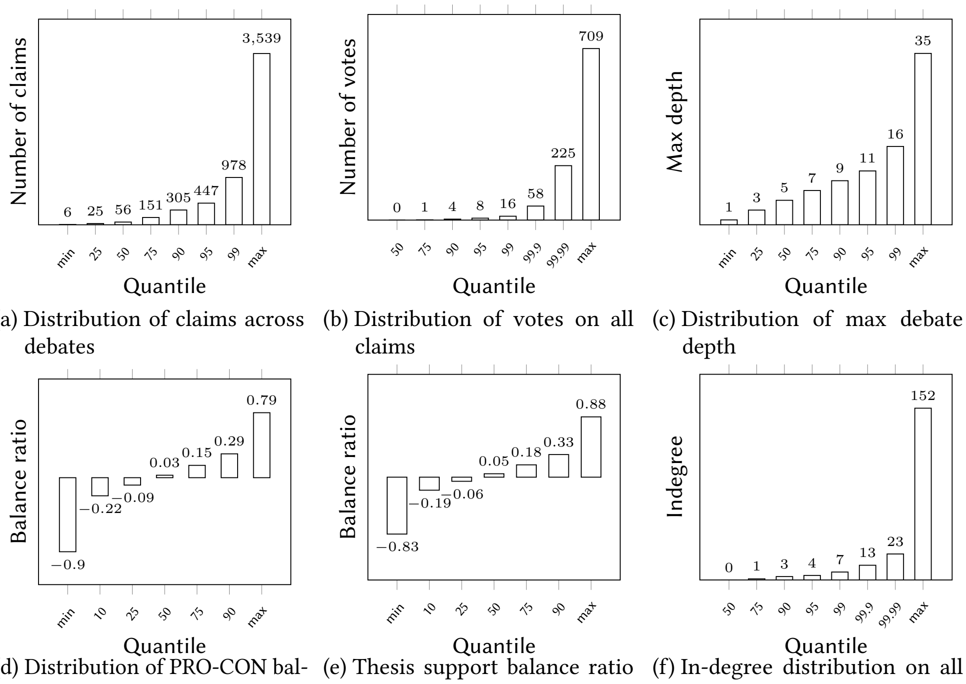
2.3. First analysis of debates

The distribution of claims per debate is illustrated in Figure 1a and shows that only a small number of debates have a significant number of claims. This is consistent with the distribution of maximum depths depicted in Figure 1c, which shows that half of the debate graphs have a maximum depth of 5 or less. All of Kialo's claims received a total of 590,261 votes. The distribution of votes by claim is depicted in Figure 1b and shows a very uneven distribution. The closed-question debates were relatively balanced, as shown in Figures 1d and 1e: Figure 1d gives the difference between the number of PRO claims and the number of CON claims, as a proportion of the number of claims in the debate. For example, 25% of debates have more CON claims than PRO claims, so that the difference between the two is at least 9% of the total number of claims in these debates. Figure 1e depicted the distribution of the difference between the number of claims which contribute to support the central question and the number of claims which help to attack it, as a proportion of the number of claims in the debate. In the same way as illustrated in Figure 1d, we can see that 10% of debates have a greater number of claims supporting the central proposition than attacking it, so that the difference between the two is at least 33% of the number of claims in these debates. Figure 1f illustrates the distribution of incoming degrees, and shows that at least half of the claims have neither an attacker nor a supporter.

3. Background

3.1. Bipolar gradual semantics

As stated in Section 1, the Kialo debates are well suited to the use of argumentation theory. Abstract argumentation theory was first introduced in [1]. The idea is to model argumentative debates using graphs: the nodes represent the arguments and the edges the attack relations between arguments. Initially, an argument can be "accepted" or "rejected" depending on the arguments that attack it. Since then, a large number of papers have enriched the modeling possibilities. The most advanced approaches today include gradual bipolar argumentation



(a) Distribution of claims across debates

(b) Distribution of votes on all claims

(c) Distribution of max debate depth

(d) Distribution of PRO-CON balance ratio across debates

(e) Thesis support balance ratio distribution across debates

(f) In-degree distribution on all claims

Figure 1: Statistical descriptions of Kialo dataset

frameworks [9, 10, 11, 12]: the arguments are still represented by nodes, but the edges represent support relationships as well as attack relationships. The arguments are associated with two real numbers: the first can be interpreted as representing the initial score (or acceptability) of an argument, i.e. when it is considered on its own, the second can represent the final score (or acceptability), i.e. when it is considered with the arguments that attack and/or support it. The "functions" that enable the final score to be calculated from the initial score and the relations between arguments are called bipolar gradual semantics. We could not study every bipolar gradual semantics from the literature here. We have chosen to study three of the most prominent ones: Discontinuity Free Quantitative Argumentation Debate (*DF-QuAD*), Euler-based Semantics and Quadratic Energy Model (*QuEM*). These semantics are diverse as they satisfy different sets of principles [11], as we will discuss later. The approach developed here is designed to handle any bipolar gradual semantics. Our objective is to empirically study reading behaviours and draw conlcusions that are as general as possible, regardless of the semantics used.

Definition 1. *A Bipolar Argumentation Framework (BAF) is a triple* (A, R^-, R^+) *such that:*

- A *is a finite set of arguments,*
- $R^- \subseteq A \times A$ *is an acyclic binary relation on* A *describing attack relations,*

- *$R^+ \subseteq A \times A$ is an acyclic binary relation on A describing support relations.*

$R^-(a)$ hence denotes the set of direct attackers of an argument a and $R^+(a)$ its set of direct supporters. In the following, we will generalize the equations by using the notation R^* which can designate either R^+ or R^-.

The first semantics we studied, *DF-QuAD* [10], was introduced to overcome the discontinuities of the *QuAD* semantics [9]. A single function is used to define the strength of the attackers and supporters separately, and the function for the final score is simplified (compared to *QuAD*). In the following s_i denotes the initial score of an argument and s_f its final score.

Definition 2 (DF-QuAD, Discontinuity Free QuAD). *Let f^* be the function used to calculate the strength of the attackers or supporters such that:*

$$\begin{aligned} f^* &\colon A \to [0,1] \\ f^*(a) &= 1 - \prod_{b \in R^*(a)} (1 - s_f(b)) \end{aligned}$$

where s_f is the function used to calculate the final score (see below), and f^* stands for f^+ (respectively f^-) if we consider the set of supporters (respectively attackers). Note that R^* stands for R^+ or R^- (but not both at the same time). By convention, if $R^*(a) = \emptyset$, $\prod_{b \in R^*(a)} (1 - s_f(b)) = 1$.
The function s_f computing the final score of an argument is defined as:

$$\begin{aligned} s_f &\colon A \to [0,1] \\ s_f(a) &= \begin{cases} s_i(a) - s_i(a) \cdot |f^+(a) - f^-(a)| & \text{if } f^-(a) \geq f^+(a) \\ s_i(a) + (1 - s_i(a)) \cdot |f^+(a) - f^-(a)| & \text{otherwise} \end{cases} \end{aligned}$$

The second semantics we studied is the *Euler-based* semantics [11]. In the *Euler-based* semantics, an energy E is calculated using the final scores of the attackers and supporters. This energy is then used to determine the final score of the considered argument. This semantics is defined as follows:

Definition 3 (Euler-based semantics). *The function s_f computing the final score of an argument is defined as:*

$$\begin{aligned} s_f &\colon A \to [0,1[\\ s_f(a) &= 1 - \frac{1 - s_i(a)^2}{1 + s_i(a) e^{E(a)}} \qquad \text{where } E(a) = \sum_{b \in R^+(a)} s_f(b) - \sum_{b \in R^-(a)} s_f(b) \end{aligned}$$

Finally, the last semantics we considered, is the *Quadratic Energy Model* (*QuEM*) [12]. This semantics uses the same definition of E as the *Euler-based* semantics but it proposes a different way of calculating the final score of an argument in order to guarantee a symmetric impact of attacks and supports. Like the *Euler-based* semantics, *QuEM* satisfies all the desirable properties identified by Amgoud and Ben-Naim.

Definition 4 (QuEM, Quadratic Energy Model). *Let $e \in \mathbb{R}$, the impact of e is given by h such that:*

$$h\colon \mathbb{R} \to [0,1]$$
$$h(e) = \frac{max(e,0)^2}{1+max(e,0)^2}$$

The function s_f used to calculate the final score of an argument is defined as:

$$s_f\colon A \to [0,1]$$
$$s_f(a) = \begin{cases} s_i(a) + (1 - s_i(a)) \cdot h(E(a)) & \textit{if } E(a) > 0 \\ s_i(a) - s_i(a)) \cdot h(-E(a)) & \textit{otherwise} \end{cases}$$

3.2. Score Initialization

All studied semantics need to define an initial score s_i for each argument. In this paper, we chose to leverage the votes on arguments provided by Kialo debates to initialize the scores. As mentioned before, on the Kialo platform, each participant can vote for an argument by choosing an integer score in $[0 \cdots 4]$. For each argument a, the votes are summarized by a vector $\mathbf{v}(a) \in \mathbb{R}^5$ of size 5 indicating how many participants voted for each score. We define the initial score of each argument a by performing a normalized weighted average as proposed by Yang et al. [2]. In the following, $\mathbf{w} \in \mathbb{R}^5$ denotes the vector of weights. The initial score of each argument is then calculated as follows:

$$s_i\colon A \times \mathbb{R}^5 \to [0,1]$$
$$s_i(a, \mathbf{w}) = \begin{cases} 0.5 & \text{if } \mathbf{v}(a) = [0,0,0,0,0] \\ \frac{\mathbf{v}(a)\cdot\mathbf{w}}{\sum_{k \in \mathbf{v}(a)} k} & \text{otherwise} \end{cases}$$

As explained by Young et al., we can interpret the five vote values as representing a "rational" reader's belief in the truth of a claim, with 0 for those thinking it is false and 4 for those believing it is true. Then, we use weights $\mathbf{w} = [0, 0.25, 0.5, 0.75, 1]$ for score initialization, where each step reflects an equal increase in confidence.

3.3. First observations about the impact of the semantics

The semantics presented in the previous section propose different ways to update the score of an argument. Before studying different reading behaviours, we wanted to emphasize how these different semantics affect the final scores obtained in debates. As discussed before, the axiomatic analysis of gradual semantics may provide precious insights. For instance, the notion of *open-mindedness* has been proposed by Potyka [13] to capture the fact that final scores can reach the bounds of their interval definition, regardless of their initial score. It was established that Euler-based and DF-QuAD do not satisfy this axiom, while QuEM does. In practice though, we observe that the score variation is higher with DF-QuAD than with QuEM on our Kialo debates. Figure 2 illustrates the distributions of the differences between the final score and

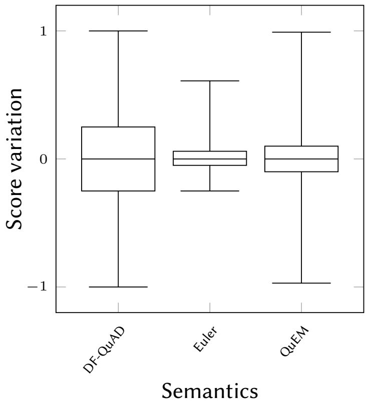


Figure 2: Score range by semantics

G_1: a

G_2: $b \longrightarrow a$

G_3: $c \longrightarrow b \longrightarrow a$

G_4: $d \longrightarrow c \longrightarrow b \longrightarrow a$

G_5: $e \longrightarrow d \longrightarrow c \longrightarrow b \longrightarrow a$

G_6: $f \longrightarrow e \longrightarrow d \longrightarrow c \longrightarrow b \longrightarrow a$

	G_2	G_3	G_4	G_5	G_6
DF-QuAD	-0.250	0.125	-0.062	0.031	-0.016
Euler	-0.075	0.010	-0.001	0.000	0.000
QuEM	-0.100	0.031	-0.009	0.003	-0.001

Table 1: Variation of the final score of argument a

the initial score of all arguments in the Kialo dataset as a function of the semantics used, when these differences are different from 0. In fact, the majority of arguments keep the same score value. This is due to the fact that they have neither attacker nor supporter, as can be seen in Figure 1f. Besides that, Figure 2 shows that the *Euler-based* semantics modifies the score less than the other semantics, but is biased towards positive variation, while the other semantics show a symmetric behaviour. This asymmetric behaviour of the Euler-based semantics is in line with axiomatic analysis [12].

Secondly, to get a better intuition regarding the impact of debates' *depth* on the scores of arguments for each semantics, we illustrate how they vary on single-path attack-only debates of increasing lengths. As depicted in Table 1, we iteratively derive from an argumentation graph G_1 containing a single argument (a), 5 other graphs (G_2 to G_6) organized as a line. A node corresponds to an argument and arrows represent attack relations between two arguments. From each graph G_i, we build G_{i+1} by adding an attacking argument to the last added node. Table 1 reports the variation of the final score of argument a for the different graphs. Note that we assume that each argument has an initial score of 0.5. Graph G_1 is not mentioned since the score of a remains the same (0.5) for each semantics (a has no attack nor support). For all studied semantics, the variation of the final score of a decreases as the depth of the graph i.e. the length of the attack path, increases. Due to the way we extend the graphs, the sign of the variation is alternating: the last argument attacks a or attacks an attack on a. Nonetheless, we can see that the relation between the depth and impact of arguments is very different among semantics. We observe that the Euler-based semantics leads to almost no variation in the score of argument a already from depth 3. On the other hand, the largest variations are observed for DF-QuAD.

4. Reading Behaviours

Debates can involve a large number of arguments leading to an information overload, so that users are rarely able to read all the arguments. In this section, we introduce various debate reading behaviours that seem natural for humans. In particular, we assume that the reader starts with the central question being debated and the arguments directly related to it. As reported by [14] in the context of participatory platforms, stakeholders "are mainly looking for assessments and arguments related to their proposal". These proposals are typically the main alternatives under discussion which can be found at the first level.

Since Kialo debates are represented as trees, each behaviour is based on a tree traversal method combined with (or without) the exploitation of information about the arguments (number of votes, chronological order, PRO/CON classification).

We considered the following methods for traversing the debate tree:

- Depth First Search (**DFS**),
- Breadth First Search (**BFS**),
- Hybrid Traversal (***HT***): based on a DFS but forcing to explore all the child nodes of an argument before exploring in depth the sub-tree of one of these children,
- No traversal (***NT***), which does not exploit the structure of the tree.

We combine these traversal methods with a ranking method for ordering the nodes (i.e. arguments) within the *same level*:

- Chronological Order (***CO***): oldest to newest (based on the timestamps of the arguments),
- Descending *likes* (***DL***): arguments are ranked based on their initial score s_i,
- Descending *likes* with PRO-CON Diversity (***PCD***): the PRO argument with the highest *like* value is ranked first, then the CON argument with the highest *like* value, then the PRO argument with the second highest *like* value, and so on.

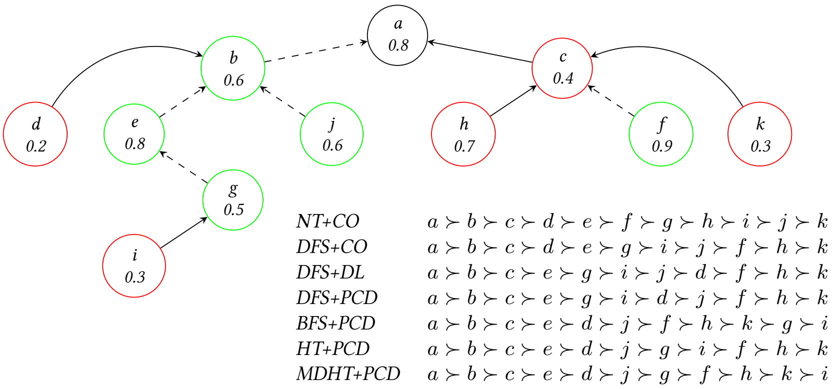
We cannot present all the combinations of traversing and ranking methods but we focus on the following reading behaviours[5]:

1. Behaviour ***NT+CO***: chronological order with no traversal,
2. Behaviour ***DFS+CO***: depth first search + chronological order,
3. Behaviour ***DFS+DL***: depth first search + descending *likes* (introduced as "likes" policy by [2]),
4. Behaviour ***DFS+PCD***: depth first search + PRO-CON diversity,
5. Behaviour ***BFS+PCD***: breadth first search + PRO-CON diversity,
6. Behaviour ***HT+PCD***: hybrid traversal + PRO-CON diversity,
7. Behaviour ***MDHT+PCD***: *HT* with maximum depth + PRO-CON diversity. A variant of *HT+PCD* where we first rank all the arguments whose depth is less or equal to 3 and then rank the remaining arguments. Indeed, we believe that a standard reader rarely ventures further than this depth. This intuition can be justified by the fact that 51% of votes are located at a depth of 3 or less.

[5] More reading behaviours are available on our Git repository.

Note that the behaviours *NT+CO*, *DFS+CO* and *DFS+DL* correspond to the comment sorting policies proposed in [2].

Example 2. *We illustrate the reading order of the arguments obtained by the different reading behaviours on the graph below. Argument a is the source node. Numerical values correspond to the like value of each node. Arguments are assumed to be labeled in chronological order (a is the oldest one and k is the newest). Solid edges represent attacks and dashed ones represent supports. PRO (resp. CON) arguments are depicted in green (resp. red).*



NT+CO	$a \succ b \succ c \succ d \succ e \succ f \succ g \succ h \succ i \succ j \succ k$
DFS+CO	$a \succ b \succ c \succ d \succ e \succ g \succ i \succ j \succ f \succ h \succ k$
DFS+DL	$a \succ b \succ c \succ e \succ g \succ i \succ j \succ d \succ f \succ h \succ k$
DFS+PCD	$a \succ b \succ c \succ e \succ g \succ i \succ d \succ j \succ f \succ h \succ k$
BFS+PCD	$a \succ b \succ c \succ e \succ d \succ j \succ f \succ h \succ k \succ g \succ i$
HT+PCD	$a \succ b \succ c \succ e \succ d \succ j \succ g \succ i \succ f \succ h \succ k$
MDHT+PCD	$a \succ b \succ c \succ e \succ d \succ j \succ g \succ f \succ h \succ k \succ i$

5. Metrics

The question is now how to compare the different reading behaviours presented in the previous section. Recall that the objective is to compare the situation where someone would have *entirely* read the debate, with the situation where someone has *partially* read the debate. In [2], in line with their choice of studying an extension-based semantics, Young et al. use the Jaccard coefficient over the set of accepted arguments obtained from the whole debate and set of accepted arguments obtained from the partially read debate. In our gradual setting, we will use the *Root-Mean-Square-Error*, *Kendall Tau* distance, and *Jaccard coefficient* in the same way to compare final scores of arguments. But before, we need to define some usefull notations. Let:

- $F = (A, R^-, R^+)$ be a *BAF*,
- $N = |A|$ the number of arguments,
- $n \in [1 \cdots N]$ the number of arguments read,
- $rb(F, n)$ a reading behaviour returning the ordered list of the n first read arguments of F,
- $s_f(a, rb(F, n))$ (respectively $s_f(a, rb(F, N))$) the final score of a when we consider the *partially read* graph (respectively the *entire* graph)
- $S_n = (s_f(a, rb(F, n))_{a \in rb(F,n)}$ the list of scores of the n first read arguments of F when we consider the *partially read* graph,
- $S_N = (s_f(a, rb(F, N))_{a \in rb(F,n)}$ the list of scores of the n first read arguments of F when we consider the *entire* graph.

In the following $S_n(i)$ (resp. $S_N(i)$) denotes the score of the i-th element of S_n (resp. S_N). We assume the i-th elements of S_n and S_N correspond to the same argument a. Note that s_f being recursive, its result depends on the size of the considered graph, but we keep only the results of n first arguments given by rb. We can now define the three considered metrics:

- the *Root-Mean-Square-Error* (RMSE) between the obtained scores:
$$RMSE = \sqrt{\sum_{a \in rb(F,n)} \frac{(s_f(a, rb(F,n)) - s_f(a, rb(F,N))^2}{n}}$$
- the *Kendall Tau* (KT) (normalized) distance between the rankings of the arguments induced by the scores, as was done in [6]:
$$KT = \frac{2|\{(i,j) : i < j, B(i,j) \vee C(i,j)\}|}{n(n-1)}$$
where $B(i,j) = (S_n(i) < S_n(j)) \wedge (S_N(i) > S_N(j))$ and $C(i,j) = (S_n(i) > S_n(j)) \wedge (S_N(i) < S_N(j))$. KT counts the number of pairwise disagreements between two ranking lists. Thus, $B(i,j)$ checks whether the i-th argument is ranked *before* the j-th in S_n, and *after* in S_N. Similarly, $C(i,j)$ checks whether the i-th argument is ranked *after* the j-th in S_n, and *before* in S_N.
- the *Jaccard coefficient* (JC) between the sets of accepted arguments, where arguments are considered as accepted in our setting when their score reaches a predefined threshold (set to 0.5 in this paper):
$$JC = \frac{|D_n \bigcap D_N|}{|D_n \bigcup D_N|} \quad \text{where} \quad \begin{array}{c} D_n = \{a \in rb(F,n) | s_f(a, rb(F,n)) > 0.5\} \\ D_N = \{a \in rb(F,n) | s_f(a, rb(F,N)) > 0.5\} \end{array}$$
JC measures similarity between D_n, the set of accepted arguments when we consider the *partially* read graph, and D_N, the set of accepted arguments when we consider the *entire* graph and when we keep only the n first arguments given by rb. Indeed, when a reader has seen n arguments, she cannot expect to have seen more than n accepted arguments. By doing this, JC is not biased when n is less than the final number of accepted arguments.

Example 3. *Let $s = \langle 0.8, 0.2, 0.7, 0.6, 0.1 \rangle$ and $s' = \langle 0.6, 0.4, 0.7, 0.1, 0.2 \rangle$ be two score vectors for some arguments a, b, c, d, e. The RSME between these two vectors s and s' of argument scores is 0.261. The induced rankings are respectively $a \succ c \succ d \succ b \succ e$ and $c \succ a \succ b \succ e \succ d$, yielding a KT distance of 0.3. Finally, the set of accepted arguments of s is $\{a, c, d\}$ while it is $\{a, c\}$ for s', hence the JC is 2/3.*

Furthermore, we will parametrize our metrics in such a way that the focus of the study can be put on specific arguments of the debates. For instance, top-k focused metrics are only concerned with the central question, together with arguments up to depth k. Technically, it suffices to weight the different metrics introduced above. As mentioned in Section 4, because stakeholders of participatory platforms "are mainly looking for assessments and arguments related to their proposal"[14], our interest will be the top-1, so in our metrics the root argument and the first level will have a weight of 1, and other arguments will have a weight of 0. But our pipeline can handle any weighting. On average, we found there are a bit fewer than 9 arguments when considering the top-1 of a debate.

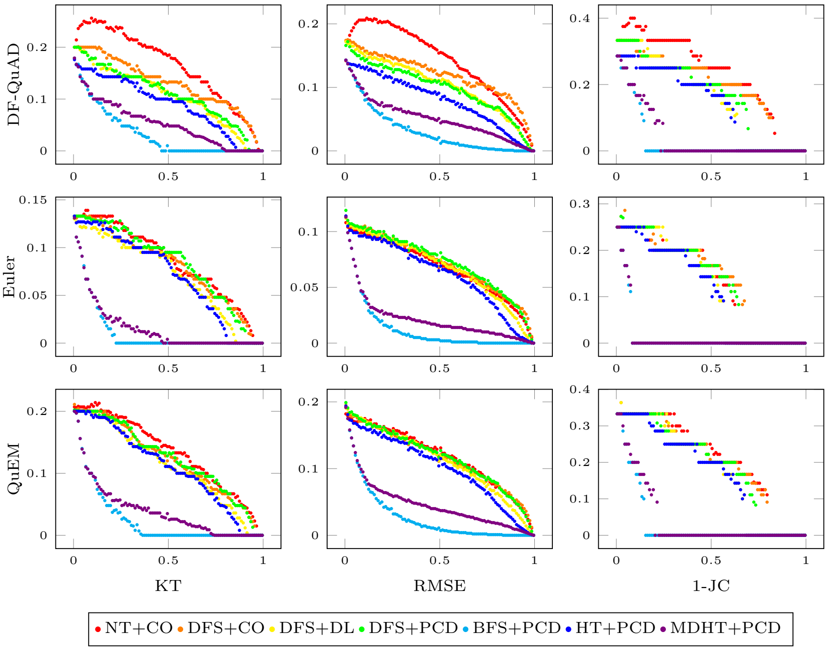


Figure 3: Simulation results on all debates

6. Experimental Results

Figure 3 shows the results of the simulations over all the debates. Each row corresponds to one of the semantics presented in Section 3, and each column to one of the metrics presented in Section 5. The x-axis corresponds to the fraction of the graph that was read, and the y-axis to the value of the corresponding metric. We emphasize that the scale of the y-axis may vary depending on the study considered. In order to help with reading the figure, we plot the values $(1 - JC)$ such that all other things being equal, the smaller the metric value, the better. Finally, each point represents the median of the results for a given reading behaviour.

Observation 1. *Whatever the metric or semantics used, behaviour BFS+PCD followed by MDHT+PCD outperform all the others.*

Observation 2. *Under BFS+PCD, the true ranking can be retrieved for all semantics when reading at least 17% in the best case and in the worst case 50% of the graph.*

Observation 3. *In trend, values of the metrics decrease for all the reading behaviours except for NT+CO under the DF-QuAD semantics for which the error first increases and then decreases.*

We have investigated the differences in errors between the behaviours and highlighted 3 main explanations:

1. Traversal direction: adding an argument in breadth rather than in depth leads to a stronger reduction of the median error. This can be observed when comparing *DFS+PCD*, *HT+PCD*, *MDHT+PCD* and *BFS+PCD*. *HT* has a behaviour close to a DFS, while *MDHT* is close to a BFS. This coincides with Figure 3 where we can see that the results of these behaviours can be ranked (from best to worst) as: *BFS+PCD* $\succ$ *MDHT+PCD* $\succ$ *HT+PCD* $\succ$ *DFS+PCD*. Finally, when we analyse the type of traversal that *NT+CO* produces at the start of reading, we realise that it is similar to a DFS.

2. Respecting the final attack and support balance: let a be one of the arguments considered by our metrics, i.e. with a weight different from 0. When we analyse the factors influencing the evaluation of the final score of a, we see that the balance between the attackers and the supporters of a plays an important role. Indeed, the final score depends on the number of attackers and supporters, and their respective scores. Therefore, any behaviour that leads to a balance, when n arguments have been read, that is too different from the one when all the arguments have been read, will produce a greater difference between the final scores of a, and therefore a greater error. In practice, this result can be observed when we compare *DFS+CO*, *DFS+DL* and *DFS+PCD*: the ranking method *PCD* produces a more faithful balance (by alternating between PRO and CON arguments and placing the strongest arguments first) throughout the reading, and so it performs better. For instance, at the maximum RMSE for *DF-QuAD* semantics, i.e. when a reader has read between 11% and 12% of the graph, 62% of the arguments added by *NT+CO* are PRO arguments compared with 54% for *BFS+PCD*. However, we did not observe any significant difference between the average scores for the arguments added by these two behaviours for this example.

3. The semantics used: although the relative performance is preserved from one semantics to another, we can see that the error values differ. In addition, we can see that the *DF-QuAD* semantics is more sensitive to the two error explanation factors that we described above, than the *Euler-based* and *QuEM* semantics.

One important finding is thus that *NT+CO* produces an unbalanced and in-depth reading of the debate. From a behavioural point of view, this can be explained by the fact that on Kialo, the first arguments are added by the creator of the debate who is more likely to be biased by her own opinion on the issue and therefore add, at the first stage of the debate, arguments that support the central question, and also arguments that support her other arguments. Oldest arguments are hence more likely to support the central question.

Even though our approach differs from [2] in many respects, it is instructive to compare our respective results. Recall that, contrary to us, they single out DFS as the best policy for their metric. We found out that the relative performances of the behaviours strongly depends on the weighting applied to the metrics. In our case of strong focus on the most central arguments, behaviours of type BFS perform best. But when no weighting is applied as in [2], we also obtain results showing that the DFS type behaviours are the ones that minimise the error.

7. Conclusion

In this paper we performed a new assessment of the impact of the behaviour of users when reading through online debates, following the methodology proposed in [2]. In a departure from this work, we explore a setting of gradual semantics natively designed for bipolar argumentation frameworks. We study a large number of natural behaviours and evaluate them through various metrics, focusing on the core arguments of the debate. Our results show differences with those of Young et al.: if your focus is on the core arguments of the debate, breadth-first reading gives a fairly accurate evaluation with as few as a third of the arguments read. It also illustrates the value of keeping a good diversity when skimming through debates. The complex interplay between the reading behaviours, the semantics used and the focus of the metrics picture a rich landscape which calls for further investigation, and one of our contributions is also to offer a fully available software environment to perform complementary studies. In future work, we plan to investigate whether our approach can provide guidance for comment sorting policies design in the case of arguments elicitated as important by the debate owner. While our metrics are flexible enough to address these cases, it may be challenging to come up with general guidelines when arguments of interest lie between the limit cases of all the arguments and only the most central arguments. We also intend to further explore our conjecture that the debate creator's bias strongly affects the initial chronological reading.

Acknowledgments

This work has been supported by ANR (french National Research Agency) project AGGREEY (ANR-22-CE23-0005).

References

[1] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artificial Intelligence 77 (1995) 321–358.

[2] A. P. Young, S. Joglekar, G. Boschi, N. Sastry, Ranking comment sorting policies in online debates, Argument & Computation 12 (2021) 265–285.

[3] C. Cayrol, M. C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 378–389.

[4] J. Leite, J. G. Martins, Social abstract argumentation, in: T. Walsh (Ed.), Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI/AAAI, 2011, pp. 2287–2292.

[5] F. Cerutti, M. Cramer, M. Guillaume, E. Hadoux, A. Hunter, S. Polberg, Empirical Cognitive Studies About Formal Argumentation, 2021, p. Chapter 14.

[6] E. Bonzon, J. Delobelle, S. Konieczny, N. Maudet, An empirical and axiomatic comparison of ranking-based semantics for abstract argumentation, Journal of Applied Non-Classical Logics 33 (2023) 328–386.

[7] D. Baumeister, M. Järvisalo, D. Neugebauer, A. Niskanen, J. Rothe, Acceptance in incomplete argumentation frameworks, Artificial Intelligence 295 (2021) 103470.

[8] J.-G. Mailly, Extension-based semantics for incomplete argumentation frameworks: properties, complexity and algorithms, Journal of Logic and Computation 33 (2023) 406–435.

[9] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argument & Computation 6 (2015) 24–49.

[10] A. Rago, F. Toni, M. Aurisicchio, P. Baroni, Discontinuity-free decision support with quantitative argumentation debates, in: Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning, KR 2016, AAAI Press, 2016, pp. 63–72.

[11] L. Amgoud, J. Ben-Naim, Evaluation of arguments in weighted bipolar graphs, in: A. Antonucci, L. Cholvy, O. Papini (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer International Publishing, Cham, 2017, pp. 25–35.

[12] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: M. Thielscher, F. Toni, F. Wolter (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, AAAI Press, 2018, pp. 148–157.

[13] N. Potyka, Open-mindedness of gradual argumentation semantics, in: N. Ben Amor, B. Quost, M. Theobald (Eds.), Scalable Uncertainty Management, Springer International Publishing, Cham, 2019, pp. 236–249.

[14] W. Aboucaya, Collaborative systems for large-scale online citizen participation, Phd thesis, 2023.