Aggregation in Value-Based Argumentation Frameworks

Grzegorz Lisowski, Sylvie Doutre, Umberto Grandi

Aggregation in Value-Based Argumentation Frameworks

Grzegorz Lisowski
University of Warwick
United Kingdom

Sylvie Doutre
University of Toulouse
France

Umberto Grandi
University of Toulouse
France

Value-based argumentation enhances a classical abstract argumentation graph - in which arguments are modelled as nodes connected by directed arrows called attacks - with labels on arguments, called *values*, and an ordering on values, called *audience*, to provide a more fine-grained justification of the attack relation. With more than one agent facing such an argumentation problem, agents may differ in their ranking of values. When needing to reach a collective view, such agents face a dilemma between two equally justifiable approaches: aggregating their views at the level of values, or aggregating their attack relations, remaining therefore at the level of the graphs. We explore the strenghts and limitations of both approaches, employing techniques from preference aggregation and graph aggregation, and propose a third possibility aggregating rankings extracted from given attack relations.

1 Introduction

The strength of arguments plays an important role in establishing the outcome of a discussion. Strong arguments have a stronger impact on the interlocutors than arguments perceived as weak. In particular, a strong argument can be accepted even if it is undermined by insignificant issues. The methods of formal modelling of the impact of the arguments' strength have been intensively studied in the abstract argumentation literature (see e.g. [26, 12]). In particular, the problem of the perceived strength of arguments is crucial in a multi-agent environment.

It is far from being clear how a consensus can be reached among multiple agents disagreeing about which arguments are stronger than others. This observation connects to an extensive discussion on methods of finding a consensus among agents who disagree about the aspects of argumentation they participate in (e.g. [13, 3, 9]).

One of the approaches to this problem emphasizes the role of *values* to which arguments appeal in determining their strength. In value-based argumentation, pioneered by Bench-Capon [4] and whose motivation traces back to Perelman [29], the concept of abstract argumentation due to Dung [11] is extended to incorporate the information on the values associated with the particular arguments. More precisely, following Dung's model, arguments are conceived as points in the graph, linked by an *attack* relation, which allows for designing rationality constraints (semantics) on acceptable sets of arguments. Additionally, each argument is assigned a unique value[1]. Further, in this approach, the strength of arguments is induced by an agent's preference over values that arguments appeal to: an attack from an argument appealing to a weak value on an highly valued argument is blocked. In this way, a *defeat graph* based on an individual view on the hierarchy of values is induced. The value-based argumentation approach has been further studied and extended in various ways (e.g. [5]).

In this paper we consider a value-based argumentation framework, and several agents, each of them with its own preference over values. In this multi-agent context, a consensus is sought. Observe that

[1] The generalisation of the model by allowing arguments to appeal to multiple values has been studied by, e.g., Kaci and van der Torre [23]. For the sake of simplicity, in the current paper we will work with the simple scenario in which an argument is assigned a unique value.

doi:10.4204/EPTCS.297.20

the considered consensus is not directly about the choice of acceptable arguments, what is known in the literature as the semantics [11], but rather about the attack graph itself and its justification in terms of preference over values.

Note that there are two intuitive manners of aggregating the possibly conflicting views of the agents. The first method focuses on the individual ordering over values, and aggregates them into a collective ordering over values that in turns induces an attack relation over labelled arguments. Off-the-shelf techniques from *preference aggregation* and social choice theory in general can be used to find suitable aggregators (see, e.g., [2, 32, 21, 6]). The second approach to this problem is to aggregate directly the agents' attack graphs, searching for a collective justification in the form of an ordering over the values a posteriori. Techniques in *graph aggregation* [18] and *belief merging* [19] have already proven useful in various settings related to argumentation [9, 10, 7].

The aim of the current paper is to investigate the properties of the two approaches in the context of value-based argumentation, their benefits and limitations. The two possibly conflicting approaches are reminescent of the discursive dilemma in the literature on judgment aggregation [15, 22], where aggregating the views of a group of agents on the premises supporting a given statement might conflict with the aggregated view on the statement itself, posing a serious challenge to the search for justifications in collective reasoning. This point relates also to the recent line of research on explainability of automated decision-making, which deals with designing methods to explain to a human user automated decisions (see, e.g. [27]). Argumentation theory is already being employed in this context (e.g [20]).

Our contribution After setting the stage for aggregating the views of agents on values and attacks over arguments, we analyze the arguably simpler method of aggregating directly the individual attack relations. As a first impossibility result we show that, if a number of natural properties are expected for the graph aggregation rule, then it is not always possible to extract a collective justification for the aggregated attack relation. We then move to the study of the aggregation of orderings over values. As there are several orderings over values that can justify a given attack relation, we investigate whether the result of the aggregation is stable with respect to this choice. We show a second impossibility result, again under standard axiomatic requirements on the aggregation procedure. As a side result, we characterize frameworks in which the ordering over values justifying a given attack relation is unique. We conclude then by providing an alternative aggregation procedure lying between the two studied approaches. To do so, we make use of the value-based framework, the context of our problem, to define an aggregation procedure over preferences on values that are suitably extracted from the individual attack relations, showing that it can ensure the same axiomatic properties of the preference aggregation rule chosen.

Related Work It is worth noting that the methods of obtaining a consensus structure of argumentation based on values has been studied before. For instance, Modgil [28] explored a modes for meta-argumentation about the conflicts between arguments which captures also argumentation about preferences over values. Furthermore, Airiau *et al.*[1] studied methods of checking if a set of argumentation frameworks can be seen as a set of defeat graphs reflecting disagreement on the importance of values. This is the closest related work to this paper, and we will refer to it further in the following sections. Finally, preference aggregation in the context of value-based argumentation has been previously proposed in the context of value-based argumentation by Pu *et al.*[30].

Outside the realm of value-based argumentation, a number of approaches have recently considered the problem of aggregation in multiagent argumentation. Coste-Marquis *et al.* [8] were the first to tackle this problem, proposing distance-based methods for the aggregation of argumentation structures. Dunne

et al. [14] later proposed an axiomatic study of the aggregation of argumentation frameworks, later expanded by the work of Delobelle *et al.* [10]. Closest to our analysis are the works of Awad *et al.* [3] and Chen and Endriss [7], which focuses on several problems related to the collective rationality of the aggregation process, i.e., whether properties satisfied by the input attack relations are preserved in the collective outcome.

This paper expands previous work by the authors [25], building on the Master thesis of Lisowski [24].

Paper structure The paper is organized as follows. In Section 2 we give the basic definitions of value-based argumentation, preference and graph aggregation, and list a number of desirable properties for the aggregation process. Further, Section 3 describes the properties of the approaches based on aggregating submitted graphs and on preference orderings. In Section 4, we analyze the combined approach. Finally, in Section 5 we provide conclusions and suggestions for further development.

2 The setting

The model we study in this paper is the *value-based argumentation setting*, due to Bench-Capon [4]. This formal framework extends Dung's abstract argumentation model [11]. Hence, the basic notion employed in the studied setting is of an *argumentation framework*. It is a directed graph, in which vertices correspond to arguments, and edges to the attack relation, capturing conflicts between arguments.

Definition 1. *An* argumentation framework *(AF) is a pair $AF = \langle A, \rightarrow \rangle$, where A is a set of arguments and $\rightarrow \subseteq A^2$ is the attack relation. We denote the fact that $\langle a,b \rangle \in \rightarrow$ as $a \rightarrow b$. $\langle a,b \rangle \notin \rightarrow$ is denoted $a \not\rightarrow b$.*

In order to capture how arguments appeal to values, argumentation frameworks are extended to *value-based argumentation frameworks*. These are labeled directed graphs, in which labels correspond to values that arguments relate to, an argument relating to one value only. For the sake of simplicity we assume that every argument is associated with a single value.

Definition 2. *A* value-based argumentation framework *(VAF) is a tuple $VAF = \langle A, \rightarrow, V, val \rangle$, where A is a set of arguments, $\rightarrow \subseteq A^2$ is an attack relation, V is a set of values and $val : A \rightarrow V$ is a function assigning values to arguments.*

Then, we can define how an agent's preferences over values determine the *strength of values* from the perspective of a particular agent. To achieve that, agents are allowed to express preference orderings over values. Following the literature on value-based argumentation, such a preference ordering is called an *audience*.

Definition 3. *Let $VAF = \langle A, \rightarrow, V, val \rangle$. An* audience *$P$ is a linear ordering*[2] *over V. We denote that a value $v_1 \in V$ is more preferable than a value $v_2 \in V$ for P as $v_1 \succ_P v_2$.*

Then, the way in which agents perceive the strength of arguments influences the way in which they perceive the structure of argumentation. Intuitively, an agent might disregard the fact that a weak argument undermines a strong argument, as it is not important enough to be considered a legitimate reason to reject a vital point. Formally, we say that an argument *a defeats* an argument b, if a attacks b and either b's value is weaker than a's, or a and b appeal to the same value. It is worth noting that as we require audiences to be linear, this means that a must be stronger than b.

[2] A linear ordering is an irreflexive, transitive and complete binary relation.

Definition 4. *Let $VAF = \langle A, \rightarrow, V, val \rangle$ be a VAF and P be an audience. Then, we say that an argument a* defeats *an argument b* for *P (we denote it as $a \rightarrow_P b$) iff $a \rightarrow b$ and it is not the case that $val(b) \succ_P val(a)$.*

Further, given an initial *VAF* and an audience, we can consider an argumentation framework based on the set of arguments present in *VAF*, and the defeat relation. Such an argumentation framework is called a *defeat graph*.

Definition 5. *Let $VAF = \langle A, \rightarrow, V, val \rangle$ and P be an audience. The* defeat graph *of VAF* induced by *P is an argumentation framework $AF = \langle A, \rightarrow_P \rangle$.*

Let us illustrate the defined notion with an example, adapted from [1]:

Example 2.1. *Consider a debate regarding the possible ban of diesel cars, aimed at the reduction of air pollution in big cities. The following arguments are included in the discussion:*

- *A - Diesel cars should be banned.*
- *B - Artisans, who should be protected, cannot change their cars as it would be too expensive for them.*
- *C - We can subsidize electric cars for artisans.*
- *D - Electric cars, which could be a substitute for diesel, require too many new charging stations.*
- *E - We can build some charging stations.*
- *F - We cannot afford any additional costs.*
- *G - Health is more important than economy, so we should spend whatever is needed for fighting pollution.*

Further, it can be noticed that these arguments appeal to certain values. In particular, arguments A, G appeal to environmental responsibility (ER), B, C to social fairness (SF), F to economic viability (EV) and D, E - to infrastructure efficiency (IE).

These arguments are represented on the graph with a mapping of values depicted on Figure 1. For each argument, the first element of its description is its name, and the second one is the name of the value it appeals to.

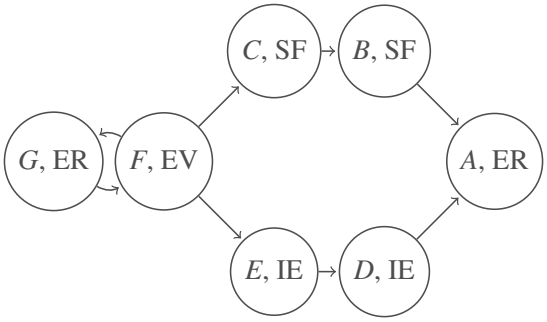

Figure 1: Value-based argumentation framework VAF of Example 2.1; each node is labeled with its name and the name of the value it appeals to; the arrows correspond to the attack relation.

Let us now consider the structure of this discussion from the perspectives of two experts of a decision-making jury, that should decide on whether Diesel cars should be banned or not.

For Expert 1, economic viability is the most important. She ranks infrastructure efficiency lower, but higher than social fairness. Environmental responsibility is the least important for her. Then, from her

point of view attacks in which the attacker appeals to a less important value than the attacked argument are disregarded. Taking her preferences into account, the structure presented in Figure 2 (a) is obtained, after the elimination of disregarded attacks. Let us now consider another expert of the jury, who believes that economic viability is the most important value. Expert 2 ranks environmental responsibility second, and social fairness third. Finally, she considers infrastructure efficiency as the least important. From her perspective, the structure of successful attacks is much different, as indicated in Figure 2 (b).

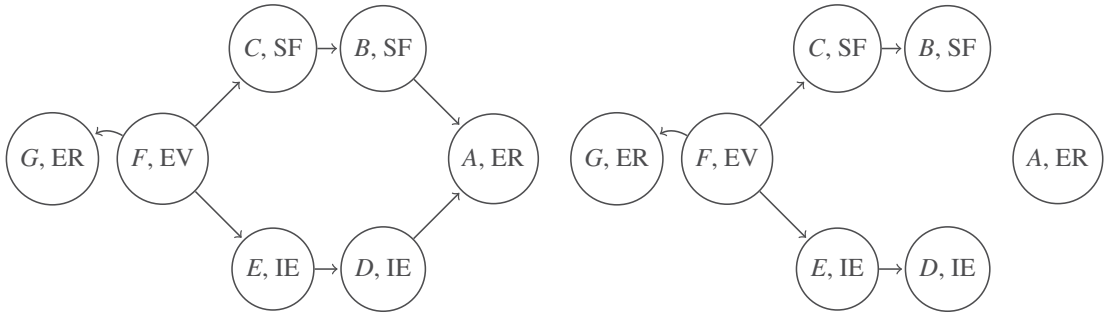

Figure 2: Defeat graphs based on (a) Expert 1's (EV $\succ$ IE $\succ$ SF $\succ$ ER) and (b) Expert 2's (EV $\succ$ ER $\succ$ SF $\succ$ IE) audiences. In the graphs, nodes are labeled with names of arguments and names of values they appeal to. The arrows correspond to the induced defeat relations based on the experts' audiences.

It is worth observing that in the current paper we assume for the sake of simplicity that agents have a certain *common ground*. Namely, they agree on the set of arguments and on the values that they appeal to. This assumption is indeed restrictive: in many cases participants of a dispute disagree on these parameters. We aim at addressing these issue in future work.

2.1 Aggregating preferences

A natural way of retrieving a collective view on the structure of argumentation taking into account individual opinions on the importance of values is to employ a *preference aggregation* approach. This method considers a profile of preference orderings, corresponding to individuals' opinions, and provides a single, collective preference ordering. We will denote the set of individuals as $\mathcal{N} = \{1, \ldots, n\}$.

Definition 6. *Let Pref* $= \{\succ_1, \ldots, \succ_n\}$ *be a set of preference orderings over a set V. A profile of* preference orderings *is a tuple* $\boldsymbol{P} = \langle \succ_1, \ldots, \succ_n \rangle$ *consisting of the elements of Pref.*

Definition 7 (Preference Aggregation Rule)**.** *Let* $V = \{v_1, \ldots, v_n\}$ *be a set of options and* $\mathscr{P}$ *be the set of all linear orderings over V. Then a* preference aggregation rule *is a function* $F : \mathscr{P}^m \to \mathscr{P}$*. We denote the set of agents supporting* $v_i \succ v_j$ *in a profile* $\boldsymbol{P}$ *as* $N_{\boldsymbol{P}}^{v_i \succ v_j}$.

This method provides a straightforward way of dealing with disagreements with respect to views on the hierarchy of values. All agents start first by submitting their preference orderings over values. Further, the orderings are aggregated with employment of a chosen preference aggregation function, resulting in a collective preference ordering ($\succ_{coll}$). Finally, by applying this collective preference ordering over values to a considered *VAF*, a collective defeat graph AF_{coll} is obtained. Figure 3 depicts this process.

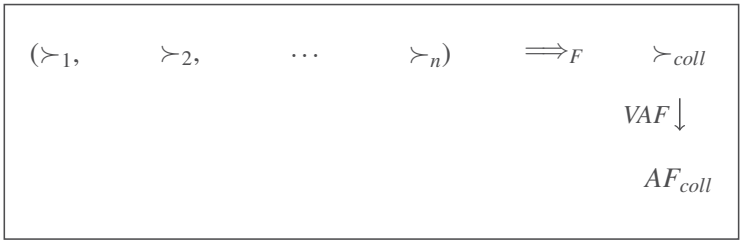

Figure 3: Preferences over values are aggregated into a collective ordering on values, which in turn induces an attack graph once a *VAF* is fixed.

A simple example of a preference aggregation rule, which we will use to illustrate the proposed mechanism, is the *Borda* rule. Let us first define the *rank* of an option from the perspective of the given preference ordering.

Notation 1. *Let P be a linear order over some set V. We denote as $rank_P(v)$ the position of the option v in the ordering P. Formally, $rank_P(v) = |\{v' \in V | v \succ_P v'\}|$.*

Then, given a profile $\mathbf{P}$ of preference orderings P_i over set V, $Borda(\mathbf{P}) = \succ_{coll}$, such that $v_i \succ_{coll} v_j$ iff $\sum_{P_i \in \mathbf{P}} rank_{P_i}(v_i) > \sum_{P_i \in \mathbf{P}} rank_{P_i}(v_j)$ for every $v_i, v_j \in V$, combined with a tie-breaking rule to obtain a strict ranking in case of equality of the Borda score.[3] Let us illustrate how a collective defeat graph is computed using the Borda rule on our running example.

Example 2.2. *(Continuation of Example 2.1) Let us consider an additional expert, Expert 3. Let us present her audience, and let us recall the audiences of the other two experts. These three experts form a panel **P**.*

- *Expert 1: $EV \succ IE \succ SF \succ ER$*
- *Expert 2: $EV \succ ER \succ SF \succ IE$*
- *Expert 3: $SF \succ ER \succ EV \succ IE$*

*Let us now calculate the result of the Borda rule for **P**. The scores are: ER: 4, EV: 7, IE: 2, SF: 5. So, Borda(**P**)= $EV \succ SF \succ ER \succ IE$. The defeat graph for this ordering is presented in Figure 4; this is the collective defeat graph for the panel.*

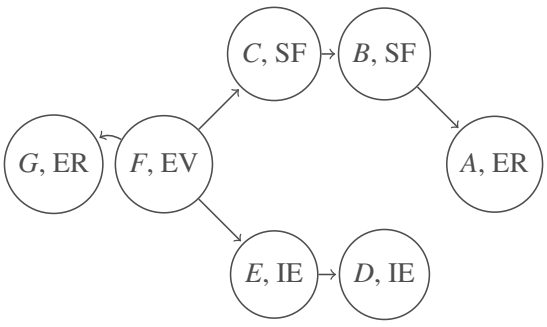

Figure 4: Collective defeat graph for the panel $\mathbf{P}$, under the Borda rule.

[3] We refer to $\sum_{P_i \in \mathbf{P}} rank_{P_i}(v_i)$ as to the *score* of v_i.

2.2 Aggregating defeat graphs

In performing preference aggregation over values, the previous section assumed that these individual preferences are known, as an input data of the search for a collective position. However, it may happen that what is known from the agents is the way they see arguments and their relationships, that is, what their defeat graph is, and not what their preferences over values are. For instance, the experts may not want to reveal what their inner preferences are, but just present how they see the resulting situation in terms of arguments and attacks. In such contexts, the search for a collective position can be done by aggregating the defeat graphs.

An intuitive requirement for this process is that the resulting collective defeat graph be *justifiable* with respect to a value-based argumentation framework the input defeat graphs are based on: that is, there should exist an ordering of values that, when applied to the initial known or unknown value-based framework, produces the collective defeat graph.

If graph aggregation is fully justified when the preferences over values are not known, one may also think of using this technique to obtain a collective view when they are known. The two techniques, graph aggregation and preference aggregation, will be compared in further sections. For now, let us present graph aggregation.

A *graph aggregation* rule is a function taking a profile of graphs as an input and providing a single graph. It is worth noting that in the considered setting we only take into account profiles of graphs sharing the set of vertices and require that the collective graph is also based on this set.

Definition 8 (Graph Aggregation Rule)**.** *Let A be a set of arguments and Graphs be the set of all argumentation frameworks based on A. Then a* graph aggregation rule *is a function $F : Graphs^m \to Graphs$. For any pair of arguments $a, b \in A$, we denote the set of agents supporting $a \to b$ in a profile $\boldsymbol{AF}$ as $N^{a\to b}_{\boldsymbol{AF}}$.*

The aggregation process is depicted on Figure 5.

$$(AF_1, \quad AF_2, \quad \cdots \quad AF_n) \quad \Longrightarrow_F \quad AF_{coll}$$

Figure 5: Individual graphs are aggregated into a collective graph.

An example of an intuitive class of graph aggregation rules is the class of *quota rules*. There, an edge is included in the collective graph if a specified fraction of agents includes it in their submitted graphs.

Definition 9 (Quota rule)**.** *Let $\boldsymbol{AF}$ be a profile of argumentation frameworks. Then, F is a* quota rule *if there is $q \in [0,1]$ such that for any attack $a \to b$, $a \to b \in F(\boldsymbol{AF})$ iff $N^{a\to b}_{\boldsymbol{AF}} \geq \lfloor q * n \rfloor$, where n is the total number of voters.*

The most well-known quota rule is the *(weak) majority rule*, where $q = \frac{1}{2}$.

It is worth noting that in the current setting we are often assuming that agents submitting their graphs have a well defined hierarchy of values in mind. Indeed, we can often safely assume that agents submit graphs which are *justifiable*. A profile is justifiable if there is a single *VAF* such that all members of the profile are defeat graphs of *VAF*.[4]

Definition 10 (Justifiable Profiles)**.** *Let $\boldsymbol{AF}$ be a profile of graphs. $\boldsymbol{AF}$ is* justifiable *if there is a VAF such that for any $AF_i \in \boldsymbol{AF}$, AF_i is a single defeat graph of VAF.*

[4]Justifiability of profiles of graphs has been studied in depth in [1].

In the later part of the paper we will mainly focus on graph aggregation rules restricted to justifiable inputs. In such rules, input profiles are only limited to profiles of graphs such that there is a *VAF* such that all members of the profile are defeat graphs of *VAF*.

We will be searching for graph aggregation rules which guarantee that if all the graphs considered in an input are defeat graphs of some *VAF*, then so is the output. It is worth noting that this problem is a special case of the problem of *collective rationality*, studied intensively in the social choice literature, also in the context of the abstract argumentation (see, e.g [7, 31]). This problem relates to the issue of whether an aggregation rule makes sure that if all of the agents involved in a decision process are submitting an option which satisfies a certain rationality constraint, so does the outcome of the procedure. In the context of graph aggregation, collective rationality means that if all input graphs satisfy a given property, the output graph satisfies it as well.

We will refer to the collective rationality with respect to the property described before as to the *preservation of being a defeat graph*.

Definition 11 (Preservation of being a defeat graph)**.** *A graph aggregation rule preserves being a defeat graph of VAF if whenever for every $AF_i \in \boldsymbol{AF}$, AF_i is a defeat graph of VAF, so is $F(\boldsymbol{AF})$.*

We can now illustrate the application of graph aggregation rules on the running example.

Example 2.3. *(Continuation of Example 2.2). Consider now the defeat graphs induced by the experts' audiences as the input of the aggregation process. We will apply the majority rule to provide a common graph for the panel. Let us list out the attacks such that the majority of agents agree that they should hold: $F \to G$, $F \to C$, $F \to E$, $C \to B$, $E \to D$, $B \to A$.*

Note that these edges corresponds to the graph obtained while using the Borda rule in the Example 2.2. It is worth observing, however, that it is not necessarily always the case, as will be highlighted by the comparison between graph and preference aggregation conducted in the following sections.

2.3 Aggregation Axioms

In following sections we will often refer to *desirable properties* of the aggregation mechanisms under consideration. Here, we will define such properties (or *axioms*), both for preference and graph aggregation. The considered axioms are standard in computational social choice, and are explained for instance in [6] for preference aggregation and [17] for graph aggregation. The wording of the definitions has been changed for the ease of presentation.

2.3.1 Preference aggregation rule

Let us start with defining the desired properties for the preference aggregation approach. We will start with an informal description of the axioms, which will be followed by a formal definition.

A preference aggregation function is *unanimous* if it never changes any ordering between options that all agents agree upon; the function is *anonymous* if it provides the same output regardless of the ordering of items in its input; *independent*, if the decision about the ordering of two values only depends on the way in which voters order this particular pair. In addition, we will strive to find rules which are not *dictatorial*. Formally:

Definition 12. *A preference aggregation function F is:*

- Unanimous*: if whenever in a profile of orderings $\boldsymbol{P}=\langle P_1, \ldots, P_n\rangle$ all voters submit that $v_i \succ v_j$, then $v_i \succ v_j$ in $F(\boldsymbol{P})$.*

- Anonymous*: if for any profile of orderings* $\boldsymbol{P}=\langle P_1,\ldots,P_n\rangle$*, any pair of items* v_i, v_j *and any permutation* $\pi : \mathscr{N} \to \mathscr{N}$*,* $a \to b \in F(\boldsymbol{P})$ *iff* $a \to b \in F(P_{\pi(1)},\ldots,P_{\pi(n)})$.
- Independent*: if it holds that for any pair of profiles of preference orderings* $\boldsymbol{P}$*,* $\boldsymbol{P}'$ *and any pair of values* $v_1, v_2 \in V$*, if* $N_{\boldsymbol{P}}^{v_1 \succ v_2} = N_{\boldsymbol{P'}}^{v_1 \succ v_2}$*, then* $v_1 \succ v_2 \in F(\boldsymbol{P})$ *iff* $v_1 \succ v_2 \in F(\boldsymbol{P'})$.
- Dictatorial*: if there is i such that for any* $\boldsymbol{P}=\langle P_1,\ldots,P_n\rangle$*,* $F(\boldsymbol{P}) = P_i$.

2.3.2 Graph aggregation rule

Let us now define corresponding axioms for graph aggregation. They are adopted from [17].

The *unanimity* axiom states that if all agents agree that some attack should be included in the collective graph, then it is. The *anonymity* condition expresses that a choice of attacks does not depend on the name of voters. *Independence* states that all attacks are treated equally in any profile of defeat graphs. We also additionally demand for a rule not to be *dictatorial*. Formally:

Definition 13. *A graph aggregation rule F is:*

- Unanimous: *if for every profile of argumentation frameworks* $\boldsymbol{AF}=\langle AF_1,\ldots,AF_n\rangle$*, if there is some pair of arguments* $a,b \in A$ *such that for every* $AF_i \in \boldsymbol{AF}$ $a \to_i b$*,* $a \to b \in F(\boldsymbol{AF})$.
- Anonymous: *if for every profile of argumentation frameworks* $\boldsymbol{AF}=\langle AF_1,\ldots,AF_n\rangle$*, any attacks* $a \to b$ *and any permutation* $\pi : \mathscr{N} \to \mathscr{N}$*,* $a \to b \in F(\boldsymbol{AF})$ *iff* $a \to b \in F(\boldsymbol{AF}_{\pi(1)},\ldots,\boldsymbol{AF}_{\pi(n)})$.
- Independent: *if for any pair of profiles of argumentation frameworks* $\boldsymbol{AF}$,$\boldsymbol{AF'}$*, if* $N_{a\to b}^{\boldsymbol{AF}} = N_{a\to b}^{\boldsymbol{AF'}}$*, then* $a \to b \in F(\boldsymbol{AF})$ *iff* $a \to b \in F(\boldsymbol{AF'})$.
- Dictatorial*: if there is i such that for any* $\boldsymbol{AF}$*,* $F(\boldsymbol{AF}) = AF_i$.

3 Impossibility Results

In this section we explore the limitation of the two aggregation approaches defined in Section 2, thus comparing aggregation at the level of values with aggregation at the level of attack graphs in the context of value-based argumentation.

3.1 Graph aggregation

We will commence with exploring the properties of the graph aggregation in the context of aggregating views on rankings of values.

Note that graph aggregation requires substantially less information than the preference aggregation based mechanism. Indeed, a graph aggregation rule can be used without specifying a context of a *VAF*. On the other hand, a preference aggregation can only be obtained when the values to which arguments appeal are specified. So, to use preference aggregation in the context of value-based argumentation we need to have the knowledge of both the profile of graphs and the *VAF*.

Unfortunately, as graph aggregation rules do not take the context of a *VAF* into account it might be the case that even though all agents participants of a discussion submit argumentation frameworks which are induced by their preferences over values, the collective graph is not justifiable by any preference ordering in the context of a given debate based on values. This is why the preservability of being a defeat graph is of high interest. In this section we will investigate the conditions which graph aggregation rules need to satisfy to preserve being a defeat graph.

Unfortunately, as we will see later, graph aggregation rules that also preserves being a defeat graph cannot satisfy all of the axioms we defined. We begin by showing that no quota rule can preserve being a defeat graph:

Proposition 1. *Being a defeat graph is not preserved by any quota rule.*

Proof. Consider a quota defeat aggregation rule F with an arbitrary quota $q \in [0,1]$. Then, take some natural number n such that $\frac{1}{n} < q$. Further, construct a $VAF = \langle A, \rightarrow, V, val \rangle$ such that $A = \{a_1, \ldots, a_n\}$, $\rightarrow = \{a_i \rightarrow a_{i+1} | i < n\} \cup \{a_n \rightarrow a_1\}$. Note that this attack relation forms a cycle. Also, let $val(a_i) = v_i$ for any argument a_i (now all arguments are assigned unique values). Then, consider a set of agents $N = \{1, \ldots, n\}$, submitting defeat graphs such that for any $i < n$, in agent i's perspective only $a_i \rightarrow_i a_{i+1}$, while for agent n only $a_n \rightarrow^n a_1$. It is easy to see that these are defeat graphs of VAF. For any agent i, the set of attacks $\{a_n \rightarrow a_m | a_n \not\rightarrow_i a_m\}$ is a chain of length $n-1$. Then, we can consider a preference ordering over values such that for any $a_n \rightarrow a_m$ such that $a_n \not\rightarrow_i a_m$, $val(a_m) \succ_i val(a_n)$. Clearly, this gives us a desired defeat graph.

Note now, that in the result of application of F to this profile, no attacks are preserved, as each of them has a support of fewer agents than $q * |N|$. But now suppose that we have an ordering P over V under which such a defeat graph would be obtained. Then, we would need to have than $v_n \succ_P v_{n-1} \succ_P \ldots \succ_P v_1 \succ_P v_n$. But then P is not transitive, so it is not a preference ordering.

Let us consider the following axiomatic characterisation of the class of quota rules:

Theorem 1 ([18])**.** *A graph aggregation F rule is anonymous, monotonic and independent iff F is a quota rule.*

This leads us immediately to an impossibility result:

Corollary 1. *Any graph aggregation rule preserving being a defeat graph violates anonymity, monotonicity, or independence.*

Proof. If a rule F is anonymous, monotonic and independent, by Theorem 1 F is a quota rule. But then, by Proposition 1, F does not preserve being a defeat graph.

By Corollary 1 it follows that all graph aggregation rule preserving being a defeat graph must violate intuitive axioms *in the general case*. However, we are especially interested in aggregating justifiable graphs. So, we would like to establish which rules restricted to justifiable inputs preserve being a defeat graph. As we will see, this is impossible for an important class of rules.

To establish this result we will employ Arrow's impossibility theorem, one of the cornerstones of social choice theory, stated here in its version for strict linear orders:

Theorem 2 ([2])**.** *Any unanimous and independent preference aggregation rule is dictatorial, when the set of alternatives has at least 3 elements.*

We can now show the following:

Theorem 3. *Any graph aggregation rule restricted to justifiable profiles and preserving being a defeat graph is not independent, not unanimous, or is dictatorial.*

Proof. Take a graph aggregation rule which preserves being a defeat graph. Also, suppose towards contradiction that it is unanimous, independent and is not dictatorial. Now, consider a $VAF = \langle A, \rightarrow, V, val \rangle$ such that $V = \{v_1, v_2, v_3\}$, $A = \{a_{v_i} | v_i \in V\}$, $val(a_{v_i}) = v_i$ and $\rightarrow = \{a \rightarrow b | a, b \in A \text{ and } a \neq b\}$. Now consider an argumentation framework AF which is a defeat graph of VAF. Note, that it corresponds to a

unique preference ordering over V. We can thus encode any profile $\mathbf{P}$ of preference orderings over V as a profile of defeat graphs $\mathbf{AF_P}$ of *VAF*. Define now a preference aggregation rule $F' : \mathscr{P}^m \to \mathscr{P}$, such that $v_i \succ v_j \in F'(\mathbf{P})$ iff $a_{v_i} \to_{F(\mathbf{AF_P})} a_{v_j}$. Notice that $F'(\mathbf{P})$ is a linear ordering: as $F(\mathbf{AF})$ is a defeat graph of *VAF*, we have that $F'(\mathbf{P})$ is connected since for every pair $\langle v_i, v_j \rangle \in V^2$ there is an attack $a_{v_i} \to a_{v_j}$. It is also anti-symmetric, as for every pair of arguments a_{v_i}, a_{v_j} we need to have that either $a_{v_i} \to_{F(\mathbf{AF})} a_{v_j}$ or $a_{v_j} \to_{F(\mathbf{AF})} a_{v_i}$ while it is not the case that $a_{v_i} \to_{F(\mathbf{AF})} a_{v_j}$ and $a_{v_j} \to_{F(\mathbf{AF})} a_{v_i}$. Otherwise it would be impossible to find a symmetric preference ordering **P'** such that $F(\mathbf{AF})$ is a defeat graph of *VAF* induced by **P'**. By a similar argument we can show that we cannot have that $a_{v_i} \to_{F(\mathbf{AF}} a_{v_j}$, $a_{v_j} \to_{F(\mathbf{AF})} a_{v_k}$ and $a_{v_k} \to_{F(\mathbf{AF})} a_{v_i}$. From this it follows that $F'(\mathbf{P})$ is transitive.

We need to demonstrate that if F is unanimous, independent and non-dictatorial, then so is F'. If F is unanimous: suppose that F' is not and take a profile of preference orderings $\mathbf{P}$ such that for some pair of values v_i, v_j, for every $P_k \in \mathbf{P}$ $v_i \succ_k v_j$ but $v_j \succ_{F'(\mathbf{P})}$. Then, by construction of F we would have that for every member of the profile $\mathbf{AF}$ of graphs induced by $\mathbf{P}$ $a_{v_i} \to a_{v_j}$, but it would not be the case in $F(\mathbf{AF})$, which contradicts the assumptions. If F is independent, suppose that F' is not and take profiles of orderings $\mathbf{P}$, **P'** over V such that for some v_i, v_j we have that $v_i \succ v_j \in F'(\mathbf{P})$ while $v_j \succ v_i \in F'(\mathbf{P'})$ even though $v_i \succ v_j$ is supported by the same voters in both profiles. Then, by construction of F we would have that for the pair of profiles of graphs $\mathbf{AF}$, **AF'** induced by $\mathbf{P}$, **P'** $a_{v_i} \to_{F(\mathbf{AF})} a_{v_j}$ but $a_{v_j} \to_{F(\mathbf{AF'})} a_{v_i}$, so F wouldn't be independent. Finally, if F' would be dictatorial, by construction of F we would have that for some i, in any profile $\mathbf{AF}$, $F(\mathbf{AF}) = AF_i$, so by contraposition if F is non-dictorial, so is F'. But this means that by Theorem 2 we cannot have independent, unanimous and non-dictatorial rules restricted to justifiable profiles which preserve being a defeat graph.

This result shows that it is not possible to find any rule satisfying all the considered desiderata. However, we can still consider natural rules preserving being a defeat graph. Note that any *representative-voter* rule satisfies this property. These are the rules which always select some graph represented in the input profile (see e.g., [16]). Another intuitive move would be to select an argumentation framework which is a defeat graph of a *VAF* justifying the input profile and minimizing a distance from the submitted graphs, in line with techniques from belief merging cited in the related work section. This approach is however not guaranteed to work, and investigating this issue is an interesting direction for future work, as it might be the case that the output of such a rule is not a defeat graph of a second *VAF'* which also happen to justify the input profile.

3.2 Preference aggregation

Let us now proceed to the study of properties of preference aggregation based mechanisms. These mechanisms have a major advantage over graph aggregation: they always produce justifiable outputs. However, the outcome of the preference aggregation based mechanism is dependent on the context of *VAF*. By the context of a *VAF* we mean the assignment of values to arguments. Indeed, it is worth noting that a preference aggregation mechanism does not always ensure that if two profiles of preference orderings induce the same profile of defeat graphs, they will produce the same collective defeat graph. They might, in principle, provide different graphs depending on the context of *VAF*. As we will see further, this is the case for all reasonable preference aggregation based mechanisms.

Formally, we will be looking for preference aggregation rules corresponding to some graph aggregation rule.

Definition 14 (Corresponding graph aggregation)**.** *Let F be a preference aggregation rule. A graph aggregation function F' corresponds to F if for every profile of justifiable argumentation frameworks, VAF=$\langle A, \rightarrow, V, val\rangle$ and a profile of preference orderings $\boldsymbol{P}$ justifying $\boldsymbol{AF}$ with respect to VAF, the defeat graph $\langle A, \rightarrow_{F(\boldsymbol{P})}\rangle = F'(\boldsymbol{AF})$.*

A necessary condition for the existence of a graph aggregation rule corresponding to a preference aggregation rule F, is that given a *VAF* and a profile **AF** of its defeat graphs, application of F will result in the same collective defeat graph, no matter which profile of preference orderings justifying **AF** is chosen. Formally, we will refer to this property as to *interpretation independence*.

Definition 15 (Interpretation Independence)**.** *A preference aggregation rule F is* interpretation independent *if for every VAF and profile $\boldsymbol{AF}$ of defeat graphs of VAF, we have that for every pair of profiles $\boldsymbol{P}$, $\boldsymbol{P'}$ justifying $\boldsymbol{AF}$ with respect to VAF, $\langle A, \rightarrow_{F(\boldsymbol{P})}\rangle = \langle A, \rightarrow_{F(\boldsymbol{P'})}\rangle$.*

The aforementioned fact follows from the following lemma:

Lemma 1. *If there is a graph aggregation rule corresponding to a preference aggregation rule F, then F is interpretation independent.*

Proof. Proof by transposition. Take a preference aggregation rule F which is not interpretation independent. Then take a profile **AF** of defeat graphs of a *VAF* such that for two profiles of preference orderings **P**, **P'**, F produces different outcomes (*AF*, *AF'*). Then note that any graph aggregation function corresponding to F would need to have both *AF* and *AF'* as the output for **AF**, which is not possible.

However, interpretation independence only holds if F is independent.

Proposition 2. *A preference aggregation rule F is interpretation independent iff F is independent.*

Proof. ($\Leftarrow$) Consider any $VAF = \langle A, \rightarrow, V, val\rangle$, as well as a profile of its defeat graphs **AF**. Also, let F be any independent preference aggregation rule. Further, take any pair of profiles $\mathbf{P}, \mathbf{P'}$ of preference orderings over V inducing **AF**. Now suppose that $\langle A, \rightarrow^{F(\mathbf{P})}\rangle \neq \langle A, \rightarrow^{F(\mathbf{P'})}\rangle$. Without loss of generality assume that there is an attack $a \rightarrow b$ such that $a \rightarrow b \in \langle A, \rightarrow^{F(\mathbf{P})}\rangle$ but $a \rightarrow b \notin \langle A, \rightarrow^{F(\mathbf{P'})}\rangle$. Then, by connectedness we know that $val(a) \succ val(b) \in F(\mathbf{P})$. Otherwise we would have that $val(b) \succ val(a) \in F(\mathbf{P})$, so the attack would be blocked. Then, take the set of voters $N_{\mathbf{P}}^{val(a)\succ val(b)}$. Note that they must correspond to defeat graphs in which $a \rightarrow b$ is included. Other defeat graphs can only be justified with orderings in which $val(b) \succ val(a)$ and, by connectedness requirement, preservation of $a \rightarrow b$ needs to be justified with an ordering in which $val(a) \succ val(b)$. So, $N_{\mathbf{P}}^{val(a)\succ val(b)}$ is also the set of supporters of $val(a) \succ val(b)$ in **P'**. So, by independence, $val(a) \succ val(b) \in F(\mathbf{P'})$. So, $a \rightarrow b \in \langle A, \rightarrow^{F(\mathbf{P'})}\rangle$. Contradiction.

($\Rightarrow$) Proof by transposition. Take a preference aggregation rule F which is not independent. Let us show that there is some *VAF* and a profile of its defeat graphs such that for two distinct profiles $\mathbf{P}_1$, $\mathbf{P}_2$ of preference orderings justifying it, the defeat graphs induced by $F(\mathbf{P}_1)$ and $F(\mathbf{P}_2)$ are not equal. We know that there is a pair of values v_1, v_2 and a pair of profiles of preference orderings $\mathbf{P}_1, \mathbf{P}_2$ such that $N_{\mathbf{P}_1}^{v_1\succ v_2} = N_{\mathbf{P}_2}^{v_1\succ v_2}$ but $v_1 \succ v_2 \in F(\mathbf{P}_1)$ while $v_1 \succ v_2 \notin F(\mathbf{P}_2)$. Take these profiles and construct a VAF=$\langle A, \rightarrow, V, val\rangle$ such that V is the set of values ordered by $\mathbf{P}_1$ and $\mathbf{P}_2$, $A = \{a_{v_k} | v_k \in V\}$, $\rightarrow = \{a_{v_1} \rightarrow a_{v_2}, a_{v_2} \rightarrow a_{v_1}\}$ and for any $val(a_{v_k}) = v_k$. Now, consider a profile **AF** of defeat graphs of *VAF* induced by $\mathbf{P}_1$. Note that both $\mathbf{P}_1$ and $\mathbf{P}_2$ justify **AF** since $N_{\mathbf{P}_1}^{v_1\succ v_2} = N_{\mathbf{P}_2}^{v_1\succ v_2}$. We know, however, that $v_1 \succ v_2 \in F(\mathbf{P}_1)$ while $v_1 \succ v_2 \notin F(\mathbf{P}_2)$. Thus, we have that in the defeat graph induced by $F(\mathbf{P}_2)$, $a_{v_2} \rightarrow a_{v_1}$, which is not the case in the graph induced by $F(\mathbf{P}_1)$.

It turns out that as a consequence of the Arrow's impossibility theorem, we can only guarantee that a preference aggregation rule is Interpretation Independent if it is not unanimous or is dictatorial.

Theorem 4. *The only unanimous preference aggregation rule corresponding to a graph aggregation function is a dictatorship.*

Proof. From Proposition 2 and Theorem 2 we have immediately that any unanimous, interpretation independent preference aggregation rule is dictatorial. Then, by Lemma 1 we get that the only unanimous preference aggregation rule corresponding to a graph aggregation function is a dictatorship.

On a more positive side, let us highlight two situations where preference aggregation rules are always interpretation independent. The first such situation is when a given *VAF* only admits defeat graphs justifiable by a *unique* preference ordering. The following proposition characterizes such *VAF*s.

Proposition 3. *For every VAF*$= \langle A, \rightarrow, V, val \rangle$ *we have that for every defeat graph AF of VAF, AF is justifiable by a unique preference ordering iff for every* $v_1, v_2 \in V$ *there is a pair of arguments* a_1, a_2 *such that* $val(a_1) = v_1$, $val(a_2) = v_2$ *and* $a_1 \rightarrow a_2$ *or* $a_2 \rightarrow a_1$.

Proof. ($\Rightarrow$) Proof by transposition. Take a *VAF*$= \langle A, \rightarrow, V, val \rangle$ such that for a pair of values $v_1, v_2 \in V$, for any $a_1, a_2 \in A$ if $val(a_1) = v_1$ and $val(a_2) = v_2$, $a_1 \not\rightarrow a_2$ and $a_2 \not\rightarrow a_1$. We will show that there are $\succ, \succ'$ such that $\langle A, \rightarrow_{\succ} \rangle = \langle A, \rightarrow_{\succ'} \rangle$. Let for every $v_1, v_2 \in V$ such that $v_1, v_2 \notin \{v_i, v_j\}$, $v_1 \succ v_2$ iff $v_1 \succ' v_2$. Also, let for every $v_1 \notin \{v_i, v_j\}$, $v_1 \succ v_i$, $v_1 \succ v_j$, $v_1 \succ' v_i$ and $v_1 \succ' v_j$. Now take any attack $a \rightarrow b$. Note that by construction of $\succ, \succ'$, if $val(a) \notin \{v_i, v_j\}$ or $val(b) \notin \{v_i, v_j\}$, $a \rightarrow_{\succ} b$ iff $a \rightarrow_{\succ'} b$. But if it is not the case, $a \not\rightarrow b$. So, $\langle A, \rightarrow_{\succ} \rangle = \langle A, \rightarrow_{\succ'} \rangle$.

($\Leftarrow$) Take a *VAF*$= \langle A, \rightarrow, V, val \rangle$ such that for every $v_1, v_2 \in V$ there is a pair of arguments a_1, a_2 such that $val(a_1) = v_1$, $val(a_2) = v_2$ and $a_1 \rightarrow a_2$ or $a_2 \rightarrow a_1$. Then suppose that there are two preference orderings $\succ, \succ'$ such that $\langle A, \rightarrow_{\succ} \rangle \neq \langle A, \rightarrow_{\succ'} \rangle$. Consider a pair of values v_1, v_2 such that $v_1 \succ v_2$ but $v_2 \succ v_1$. W.l.o.g assume that there is a pair of arguments a_1, a_2 such that $val(a_1) = v_1$, $val(a_2) = v_2$ and $a_1 \rightarrow a_2$. But we know that $a_1 \rightarrow_{\succ} a_2$ iff $a_1 \rightarrow_{\succ'} a_2$, and hence $v_1 \succ v_2$ iff $v_1 \succ' v_2$, which contradicts the assumptions.

Another situation guaranteeing interpretation independence is when a *VAF* has not more than two values, as shown in the following proposition.

Proposition 4. *For every VAF* $= \{A, \rightarrow, V, val\}$ *such that* $|V| \leq 2$*, either all defeat graphs are justified by a unique preference ordering, or all defeat graphs of VAF are equal.*

Proof. Consider any such *VAF* $= \{A, \rightarrow, V, val\}$ and let $V = \{v_1, v_2\}$ (note that if V is a singleton, all defeat graphs of *VAF* are equal by definition of a defeat graph). Suppose that for some pair of arguments a_1, a_2 such that $val(a_1) = v_1$ and $val(a_2) = v_2$, $a_1 \rightarrow a_2$ or $a_2 \rightarrow a_1$. Then, by Proposition 3 we have that all defeat graphs of *VAF* are justified by a unique preference ordering. Also, if this is not the case, for all attacks $a \rightarrow b$ in $\rightarrow$, $val(a) = val(b)$. So, by definition of a defeat graph, for any preference ordering $\succ$, $\rightarrow = \rightarrow_{\succ}$.

4 Retrieved orderings

Note that both of the approaches considered so far have some drawbacks. On the one hand, we have shown that any graph aggregation function satisfying desirable properties cannot ensure that a *VAF* justifying a profile of graphs justifies also the collective graph. On the other hand, obtaining a collective

defeat graph using preference aggregation functions requires more information than the employment of graph aggregation and we cannot ensure its correspondence to any graph aggregation rule. In this section we explore an aggregation method which combines the two approaches. Our goal will be to provide a procedure which both makes sure that the collective argumentation framework is always a defeat graph of the same *VAF* as the input graphs, and that it corresponds to some graph aggregation rule.

To this purpose we consider the following mechanism. First, a profile of defeat graphs on a given *VAF* is submitted by the agents. Then, a profile of preference orderings justifying the profile is derived (multiple choices are possible). Further, these preferences are aggregated using a suitable preference aggregation function, which in turns induces a collective defeat graph on the initial *VAF*.

The considered mechanism is illustrated in Figure 6. The lower tier of the picture represents a profile of defeat graphs of a certain *VAF*. Each $\succ_i$ is a preference ordering inducing AF_i, and is aggregated into $\succ_{coll}$ by means of a preference aggregation function F. This step is depicted as an arrow in the upper tier. The collective defeat graph AF_{coll} is then induced by *VAF* using $\succ_{coll}$.

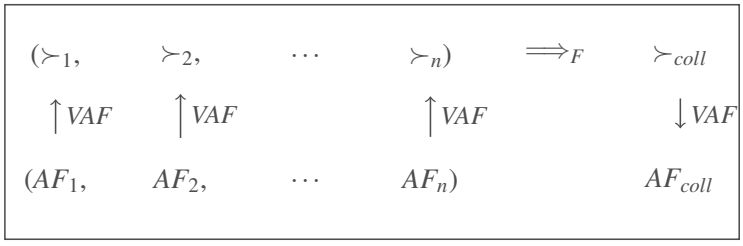

Figure 6: The individual attack graphs AF_i are provided a justification order over values $\succ_i$, which are then aggregated using a preference aggregation rule F, and the collective ordering $\succ_{coll}$ so obtained induces a collective attack graph AF_{coll} justified by it.

Let us define the proposed method formally. Note that to ensure the functionality of the proposed mechanism, it is required that the selected choice of preference orderings is predetermined: we do not distinguish between exact preferences over values that agents believe in, as long as they induce the appropriate profile of defeat graphs. To cope with this issue, we will always specify a *selection of justifications*. Intuitively, we infer preference orderings over values that participants of a debate have, based on their perceived strength of arguments.

Definition 16 (Justification selection)**.** *Let* $(Graphs^{VAF})^n$ *be the collection of all profiles of defeat graphs of VAF=*$\langle A, \rightarrow, V, val\rangle$ *of length n, and* $Prefs^n$ *be the set of all profiles of preference orderings over V of length n. Then a* justification selection *is a function* $J : (Graphs^{VAF})^n \rightarrow Prefs^n$ *such that* $J(\mathbf{AF})$ *induces* $\mathbf{AF}$ *for any profile* $\mathbf{AF}$ *in* $(Graphs^{VAF})^n$ *.*

Then, we can provide a definition of the proposed mechanism.

Definition 17 (C_F^J-mechanism)**.** *Take VAF=*$\langle A, \rightarrow, V, val\rangle$*, a profile* $\mathbf{AF}$ *of defeat graphs of VAF, a preference aggregation function F, and a justification selector J. Then,* $C_F^J(VAF, \mathbf{AF})$ *amounts to the defeat graph* $AF = \langle A, \rightarrow_{F(J(\mathbf{AF}))}\rangle$.

Henceforth, for the sake of clarity we will often assume that a justification selection is fixed and omit the superscript J. Let us illustrate the mechanism on the running example.

Example 4.1. *(Continuation of Example 2.3) Consider the profile of graphs submitted by the experts as in Example 2.3. Note that they are all defeat graphs of the VAF presented in Example 2.1. Thus, we can retrieve the preference orderings justifying them and apply the Borda preference aggregation rule. Let*

us assume that the tie-breaking protocol follows a fixed ordering $ER > SF > IE > EV$. Finally, once the collective preference ordering has been obtained, we can construct the collective defeat graph.

Suppose that the selector J associates with Expert 1 and Expert 3 the same orderings as in Example 2.2, while for Expert 2 the selectore chooses $EV \succ ER \succ IE \succ SF$. If we aggregate the three preferences using the Borda rule we obtain $EV \succ ER \succ SF \succ IE$, which is a different result than the one in Example 2.2. If we then construct the associated collective defeat graph, we observe that the attack from the argument B on A is blocked, while the remaining ones are as in the collective graph of Example 2.2.

The proposed mechanism C_F enjoys a number of interesting properties. First, this method does not require information about the exact preference orderings over values, so it does not require us to demand agents to provide additional information, as it was the case when the preference aggregation method was involved. However, we still have not ensured that there is a graph aggregation rule corresponding to any mechanism C_F. To satisfy this, we would need to have that the collective graph is not dependent on the chosen *VAF* justifying a profile of graphs.

Definition 18 (*VAF*-Independence)**.** *C_F is VAF-Independent if for any profile of justifiable graphs $\boldsymbol{AF}$ and a pair VAF, VAF' of VAFs justifying $\boldsymbol{AF}$, $C_F(\boldsymbol{AF}, VAF) = C_F(\boldsymbol{AF}, VAF')$.*

Unfortunately, as we will see, this can be the case only if F violates unanimity or is dictatorial.

Proposition 5. *If C_F is VAF-Independent, then F is dictatorial or non-unanimous for $|V| \geq 4$.*

Proof. Let us firstly show that any *VAF*-Independent C_F requires F to be independent for profiles based on at least 4 values. Suppose that F is not independent. Then, take two profiles **P**, **P'** over a set V with $|V| \geq 4$ such that for some pair of values v_1, v_2 we have that $N_{\mathbf{P}}^{v_1 \succ v_2} = N_{\mathbf{P'}}^{v_1 \succ v_2}$ but $v_1 \succ_{F(\mathbf{P})} v_2$ while $v_2 \succ_{F(\mathbf{P'})} v_1$. Then, consider a profile of graphs $\mathbf{AF} = \langle AF_1, \ldots AF_n \rangle$ such that $n = |\mathbf{P}|$ and for $AF_i \in \mathbf{AF}$, $AF_i = \{A, \rightarrow_i\}$ and $A = \{a_{v_i} | v_i \in V\}$, $\rightarrow_i = \{a_{v_1} \rightarrow_i a_{v_2}\}$ if $v_1 \succ_i v_2$, otherwise $\rightarrow_i = \{a_{v_2} \rightarrow_i a_{v_1}\}$. Now take two *VAF*s such that **AF** is a profile of defeat graphs of *VAF* justified by **P**, while **AF** is a profile of defeat graphs of *VAF'* justified by **P'**, such that for both *VAF*s, a_{v_1} and a_{v_2} attack each other, $val(a_{v_1}) = val'(a_{v_1}) = v_1$ and $val(a_{v_2}) = val'(a_{v_2}) = v_2$, while for some pair of arguments a_{v_i}, a_{v_j}, $val(a_{v_i}) = v_i, val(a_{v_j}) = v_j$ and $val'(a_{v_i}) = v_j, val'(a_{v_j}) = v_i$. Now note that the outcome of $C_F(\mathbf{AF}, VAF)$, $C_F(\mathbf{AF}, VAF')$ only depends on the collective preference ordering over v_1, v_2. So, we will have that $a_{v_1} \rightarrow_{C_F(\mathbf{AF}, VAF)} a_{v_2}$, while $a_{v_2} \rightarrow_{C_F(\mathbf{AF}, VAF')} a_{v_1}$. So, C_F is not *VAF*-Independent. It follows now from Theorem 2 that if C_F is *VAF*-Independent, then F is dictatorial or non-unanimous.

This leads us to conclude that when there are at least 4 values, if F is non-dictatorial and unanimous, a two arguably basic properties, then C_F does not correspond to any graph aggregation rule. Note, however, that this issue is solved once we fix the *VAF* justifying particular profiles of defeat graphs.

Let us first define a fixed selection of a *VAF* for a profile of graphs.

Definition 19 (VAF selection)**.** *Let $(Graphs)_J^n$ be the collection of all justifiable profiles of argumentation frameworks of length n, and VAFs be the set of all VAFs. Then a* VAF selection *is a function $S : (Graphs)_J^n \rightarrow VAFs$ such that for any profile $\boldsymbol{AF}$, $S(\boldsymbol{AF})$ is such that for every $AF_i \in \boldsymbol{AF}$, AF_i is a defeat graph of $S(\boldsymbol{AF})$.*

We can now define the modified mechanism.

Definition 20 ($C_{F,S}^{VAF}$ mechanisms with fixed *VAF*)**.** *Let F be a preference aggregation rule. Then, the combined mechanism is the function $C_{F,S}^{VAF} : Graphs_J \rightarrow Graphs$, where $Graphs_J$ is the set of all justifiable profiles of graphs, such that $C_{F,S}^{VAF}(\boldsymbol{AF}) = \langle A, \rightarrow_{F(\boldsymbol{P})} \rangle$ such that $\boldsymbol{P}$ is a profile of preference orderings justifying $\boldsymbol{AF}$ from the perspective of $S(\boldsymbol{AF})$.*

In what follows we will assume that S is fixed and drop the subscript when clear from context. Then, it is not difficult to show that C_F^{VAF} corresponds to a graph aggregation rule.

Observation 1. *Any C_F^{VAF} corresponds to a graph aggregation rule restricted to justifiable inputs.*

Proof. Take a profile of graphs $\mathbf{AF}$ justifable by a given *VAF* and a combined mechanism C_F^{VAF}. Then notice that the choice of *VAF* and of the profile of preference orderings justifying $\mathbf{AF}$ have been predetermined, and thus we can only obtain one graph. This is the case for all justifiable profiles of graphs. So, C_F^{VAF} indeed corresponds to a graph aggregation rule.

Not only this movement ensures the existence of a corresponding graph aggregation rule, but also that the correspondent inherits beneficial properties of the used preference aggregation rule.

Proposition 6. *Let F be a preference aggregation rule. Then it holds that: (1) If F is unanimous, then the graph aggregation rule F' corresponding to a C_F^{VAF}*

is unanimous. (2) If F is anonymous, then the graph aggregation rule F' corresponding to a C_F^{VAF} is anonymous.

is monotonic. (3) If F is independent, then the graph aggregation rule F' corresponding to a C_F^{VAF} is independent.

Proof. (1) Take any unanimous preference aggregation rule F. Then, suppose that the graph aggregation rule F' corresponding to C_F^{VAF} is not unanimous. Then, take a *VAF* and a profile of preference orderings $\mathbf{P}^*$ inducing a profile of defeat graphs $\mathbf{AF}$ such that there is an attack $a \to b$ such that for any $AF_i \in \mathbf{AF}$, $a \to b \in AF_i$, but $a \to b \notin F'(\mathbf{AF})$. Note that then $val(a) \neq val(b)$. Then, we must have that all agents submit that $val(a) \succ val(b)$. But then, by unanimity of F, $val(a) \succ val(b) \in F(\mathbf{P}^*)$, and thus $a \to b \in F'(\mathbf{AF})$. Contradiction.

(2) Take any anonymous preference aggregation rule F. Then, suppose that the graph aggregation rule F' corresponding to C_F^{VAF} is not anonymous. Then, take a *VAF* and a profile of preference orderings $\mathbf{P}$ and a sequence $\mathbf{AF}$ of defeat graphs induced by $\mathbf{P}$ such that there is a permutation π of $\mathbf{AF}$ such that $F'(\mathbf{AF}) \neq F'(\pi(\mathbf{AF}))$. But now note that this would imply that $F(\mathbf{P}) \neq F(\pi(\mathbf{P'}))$ which cannot be the case by anonymity of F.

(3) Take any independent preference aggregation rule F. Then, suppose that the graph aggregation rule F' corresponding to C_F^{VAF} is not independent. So, take a *VAF* and two profiles of defeat graphs $\mathbf{AF}$, $\mathbf{AF'}$ of *VAF* such that there is some attack $a \to b$ such that $N_{\mathbf{AF}}^{a \to b} = N_{\mathbf{AF'}}^{a \to b}$, but $a \to b \in F'(\mathbf{AF})$ while $a \to b \notin F'(\mathbf{AF'})$. Note that by connectedness of preference orderings, for any justification $\mathbf{P}, \mathbf{P'}$ of $\mathbf{AF}$, $\mathbf{AF'}$, $N_{\mathbf{P}}^{val(a) \succ val(b)} = N_{\mathbf{P'}}^{val(a) \succ val(b)}$. So, by independence, $a \succ b \in F(\mathbf{P})$ iff $a \succ b \in F(\mathbf{P})$. So, if $a \to b \in F'(\mathbf{AF})$, then $a \to b \in F'(\mathbf{AF'})$. Contradiction.

So, we have provided an alternative method of aggregating agents' views on the importance of values in the context of value-based argumentation. This has allowed us to overcome the issues we pointed out for the preference aggregation and graph aggregation approaches: we ensure that the outcome of the procedure is a defeat graph of the considered *VAF*, while making sure that the mechanism corresponds to a graph aggregation rule restricted to justifiable inputs.

5 Conclusion and Perspectives

In this paper we have explored three plausible methods of aggregating agents' views in the context of value-based argumentation, in order to obtain a collective argumentation framework: preference-based, graph-based, and combination-based aggregation methods.

For the preference aggregation approach, we have shown that non-dictatorial rules never correspond to graph aggregation rules and thus, depending on the context, different collective argumentation frameworks can be obtained even though agents perceive the defeat relation in the same manner. On the other hand, no graph aggregation approach satisfying intuitive properties can ensure that the outcome of the procedure is a defeat graph of the same *VAF* as the input graphs. These issues can be circumvented by an aggregation approach combining preference aggregation with the extraction of an ordering over values from the individual attack relations, which fixes a *VAF* and devises a suitable aggregation procedure for the specific context. This is one of the main conclusions of this work: obtaining a *justifiable* collective argumentation graph is not possible without taking into consideration the underlying *VAF*, which acts as the context of the problem.

This observation leaves a vast room for further research, investigating further aggregators that are specific to a given (class of) *VAF*. In the current setting we have restricted ourselves to studying preference orderings over values expressed as *linear* orderings. This is a strong requirement. It would be of high interest to study the aggregation problems stated in this paper when agents are allowed to consider particular values as equally important. Moreover, it would be interesting to study the scenario in which agents do not agree on the assignment of values to arguments. Further, it would be of natural interest to investigate, following e.g. Kaci and van der Torre [23], the scenario in which arguments appeal to multiple values. Finally, studying classes of argumentation problems which are not affected by the problems highlighted in the paper would be very beneficial towards showing the applicability of the proposed approaches.

Acknowledgments

We would like to thank Sonja Smets for useful comments in the early phases of this project. Most of the work was completed when Grzegorz Lisowski was hosted at the University of Toulouse in 2018.

References

[1] Stéphane Airiau, Elise Bonzon, Ulle Endriss, Nicolas Maudet & Julien Rossit (2017): *Rationalisation of profiles of abstract argumentation frameworks: Characterisation and complexity*. Journal of Artificial Intelligence Research 60, pp. 149–177, doi:10.1613/jair.5436.

[2] Kenneth J Arrow (1951): *Social Choice and Individual Values*. Wiley.

[3] Edmond Awad, Richard Booth, Fernando Tohmé & Iyad Rahwan (2015): *Judgement Aggregation in Multi-Agent Argumentation*. Journal of Logic and Computation 27(1), pp. 227–259, doi:10.1093/logcom/exv055.

[4] Trevor Bench-Capon (2003): *Persuasion in Practical Argument Using Value-Based Argumentation Frameworks*. Journal of Logic and Computation 13(3), pp. 429–448, doi:10.1093/logcom/13.3.429.

[5] Trevor Bench-Capon, Sylvie Doutre & Paul E Dunne (2007): *Audiences in Argumentatio Frameworks*. Artificial Intelligence 171(1), pp. 42–71, doi:10.1016/j.artint.2006.10.013.

[6] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang & Ariel D Procaccia (2016): *Handbook of Computational Social Choice*. Cambridge University Press, doi:10.1017/CBO9781107446984.

[7] Weiwei Chen & Ulle Endriss (2019): *Preservation of Semantic Properties in Collective Argumentation: The Case of Aggregating Abstract Argumentation Frameworks*. Artificial Intelligence 269, pp. 27–48, doi:10.1016/j.artint.2018.10.003.

[8] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex & Pierre Marquis (2007): *On The Merging of Dung's Argumentation Systems*. Artificial Intelligence 171(10-15), pp. 730–753, doi:10.1016/j.artint.2007.04.012.

[9] Jérôme Delobelle, Adrian Haret, Sébastien Konieczny, Jean-Guy Mailly, Julien Rossit & Stefan Woltran (2016): *Merging of Abstract Argumentation Frameworks*. Proceedings of the 2016 Knowledge Representation (KR), pp. 33–42.

[10] Jérôme Delobelle, Sébastien Konieczny & Srdjan Vesic (2018): *On The Aggregation of Argumentation Frameworks: Operators and Postulates*. Journal of Logic and Computation 28(7), pp. 1671–1699, doi:10.1093/logcom/exy023.

[11] Phan Minh Dung (1995): *On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games*. Artificial Intelligence 77(2), pp. 321–357, doi:10.1016/0004-3702(94)00041-x.

[12] Paul E Dunne, Anthony Hunter, Peter McBurney, Simon Parsons & Michael Wooldridge (2011): *Weighted Argument Systems: Basic Definitions, Algorithms, and Complexity Results*. Artificial Intelligence 175(2), pp. 457–486, doi:10.1016/j.artint.2010.09.005.

[13] Paul E Dunne, Pierre Marquis & Michael Wooldridge (2012): *Argument Aggregation: Basic Axioms and Complexity Results*. Proceedings of the Coference on Computational Models of Argument (COMMA), pp. 129–140.

[14] Paul E. Dunne, Pierre Marquis & Michael Wooldridge (2012): *Argument Aggregation: Basic Axioms and Complexity Results*. In: Computational Models of Argument, pp. 129–140.

[15] Ulle Endriss (2016): *Judgment aggregation*. In F Brandt, V Conitzer, U Endriss, J Lang & A. D Procaccia, editors: Handbook of Computational Social Choice, Cambridge University Press, doi:10.1017/cbo9781107446984.018.

[16] Ulle Endriss & Umberto Grandi (2014): *Binary Aggregation by Selection of the Most Representative Voter*. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 668–674.

[17] Ulle Endriss & Umberto Grandi (2014): *Collective Rationality in Graph Aggregation*. In: 21st European Conference on Artificial Intelligence (ECAI), pp. 291–296.

[18] Ulle Endriss & Umberto Grandi (2017): *Graph Aggregation*. Artificial Intelligence 245, pp. 86–114, doi:10.1016/j.artint.2017.01.001.

[19] Patricia Everaere, Sébastien Konieczny & Pierre Marquis (2017): *An Introduction to Belief Merging and its Links with Judgment Aggregation*. In Ulle Endriss, editor: Trends in Computational Social Choice, chapter 7, AI Access, pp. 123–143.

[20] Xiuyi Fan & Francesca Toni (2015): *On Computing Explanations in Argumentation*. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 1496–1502.

[21] Allan Gibbard (1973): *Manipulation of Voting Schemes: a General Result*. Econometrica: Journal of the Econometric Society, pp. 587–601, doi:10.2307/1914083.

[22] Davide Grossi & Gabriella Pigozzi (2014): *Judgment Aggregation: a Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning 8(2), pp. 1–151, doi:10.2200/S00559ED1V01Y201312AIM027.

[23] Souhila Kaci & Leendert van der Torre (2008): *Preference-based argumentation: Arguments supporting multiple values*. International Journal of Approximate Reasoning 48(3), pp. 730–751, doi:10.1016/j.ijar.2007.07.005.

[24] Grzegorz Lisowski (2018): *Preventing Manipulation in Aggregating Value-Based Argumentation Frameworks*. Master's thesis, Universiteit Van Amsterdam.

[25] Grzegorz Lisowski, Sylvie Doutre & Umberto Grandi (2018): *Preventing Manipulation in Aggregating Audiences in Value-Based Argumentation Frameworks*. In: *Proceedings of International Workshop on Systems and Algorithms for Formal Argumentation (SAFA 2018)*, pp. 48–59.

[26] Paul-Amaury Matt & Francesca Toni (2008): *A Game-Theoretic Measure of Argument Strength for Abstract Argumentation*. In: *Logics in Artificial Intelligence*, pp. 285–297, doi:10.1007/BF01448847.

[27] Tim Miller (2019): *Explanation in Artificial Intelligence: Insights From the Social Sciences*. *Artificial Intelligence* 267, pp. 1 – 38, doi:10.1016/j.artint.2018.07.007.

[28] Sanjay Modgil (2009): *Reasoning About Preferences in Argumentation Frameworks*. *Artificial Intelligence* 173(9-10), pp. 901–934, doi:10.1016/j.artint.2009.02.001.

[29] Chaim Perelman (1971): *The New Rhetoric*. In: *Pragmatics of Natural Languages*, Springer, pp. 145–149, doi:10.1007/978-94-010-1713-8 8.

[30] Fuan Pu, Jian Luo, Yulai Zhang & Guiming Luo (2013): *Social Welfare Semantics for Value-Based Argumentation Framework*. In: *Proceedings of International Conference on Knowledge, Science, Engineering and Management*, Springer, pp. 76–88.

[31] Iyad Rahwan & Fernando Tohmé (2010): *Collective Argument Evaluation as Judgement Aggregation*. In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 417–424.

[32] Mark Allen Satterthwaite (1975): *Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions*. *Journal of Economic Theory* 10(2), pp. 187–217, doi:10.1016/0022-0531(75)90050-2.