Preventing Manipulation in Aggregating Value-Based Argumentation Frameworks

MSc Thesis *(Afstudeerscriptie)*

written by

Grzegorz Lisowski
(born February 16th, 1993 in Warsaw, Poland)

under the supervision of **Umberto Grandi** and **Sonja Smets**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
July 5th, 2018

Members of the Thesis Committee:
Dr. Umberto Grandi
Prof. Dr. Sonja Smets
Prof. Dr. Yde Venema (Chair)
Prof. Dr. Robert van Rooij
Dr. Ronald de Haan
Dr. Davide Grossi

Institute for Logic, Language and Computation

Abstract

Recently, connections between abstract argumentation and decision making have gained increasing attention. In particular, value-based argumentation attempts to capture the specificity of deliberation concerning a choice of actions. This approach assumes that in such debates arguments appeal to certain values. Each agent ranks the values and evaluates arguments in accordance with her own preferences over values that they appeal to. The model assumes that an agent disregards attacks on strong arguments by weaker attackers. This move creates agent-specific argumentation frameworks. Another recent line of research in abstract argumentation involves situations in which agents aggregate agents' views on acceptability of arguments, or on the structure of argumentation.

In the thesis I study strategic behavior in argumentation based on values. The thesis consists of two major parts. The first one considers the single agent scenario. Here, I investigate the possibility for an agent to enforce that an argument supporting her desired decision is accepted when her preference ordering over values does not allow this acceptance. Then, methods of finding the closest preference ordering to the agent's original hierarchy of values sufficient to achieve this goal are considered.

In the second part I investigate the problem of manipulating the outcome discussion based on values in the multi-agent setting. Here, manipulation is understood as communicating an insincere preference ordering over values to ensure that a desired decision is made. A challenge tackled in this part is concerned with providing a procedure for aggregating opinions about the relative strength of arguments based on values that they appeal to. Two approaches to this problem are considered. In the first of them agents' preferences over values are aggregated directly with employment of preference aggregation functions. The other approach involves aggregation of argumentation frameworks corresponding to agents' views on the relative strength of arguments.

Acknowledgements

One of the first things I heard about the Master of Logic is that nobody can survive it alone. That is so true. But with all of the people I met along the way, I could not only see it through, but also enjoy this rough ride.

First of all, my deep gratitude goes to Sylvie Doutre and Umberto Grandi for their warm welcome in Toulouse and the most supportive supervision I could ever imagine. Thank you for your guidance and for your endless patience. Thanks for your ability to lift me up when necessary and to calm me down when my ideas were getting too far-fetched. Further, I would like to thank my ILLC supervisor, Sonja Smets. Thanks for the support in the early phase of working on this thesis, and in many other projects along my studies.

I could not be more thankful to Ulle Endriss for suggesting the opportunity of doing my thesis in Toulouse and for the most inspirational courses I ever attended.

Also, thanks to the thesis defense committee members for your time and useful comments.

Further, I want to thank the entire Toulouse research community for having me there for a couple of months. Especially, I would like to thank Arianna, Dennis and Audren for their warmth and understanding exactly when I needed it the most.

Thanks go to all the MoL students, especially my greatest flatmates, Morwenna, Miquel, Jonathan and Dean. You changed our Diemen house into a real home where things were anything but boring. Thanks for uijfertoffel burgers and tortilla de patatas, never to forget. Max, thanks for nice projects we did together and cool trips that followed them. Ian, thanks for the introduction to the specifity of French culture. Finally, thanks to all my foosball sparing partners, you know how much this game means to me.

Last, but for sure not least, my gratitude goes to Julia. Whatever I would say here would be too little.

Contents

Chapter 1

Introduction

Abstract argumentation theory, pioneered by Dung (1995), deals with selecting sets of arguments that one can rationally accept, given that there might be counterarguments undermining some of them. For example, let us consider a person telling her friend that she wants to go for a bike trip because she thinks that it is going to be sunny all day long. However, her friend claims that it is a bad idea, because a reliable forecast said that it is going to rain in the afternoon. Now, the decision-maker has a choice. Either she chooses to believe that the forecast is accurate and to drop her initial belief, or she decides to ignore the friend's advice and to go biking anyway. She cannot, however, accept both points of view. Then, she would need to accept that it is going to rain and that it is not going to rain at the same time. This simple argumentation would have a structure shown in the Figure 1.1.

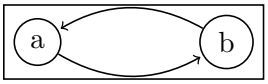


Figure 1.1: Structure of the example argumentation.

Naturally, in more complex cases it is not always clear which sets of arguments a decision-maker can sensibly select. Then, a formalization of rationality constraints, offered by the abstract argumentation theory is highly beneficial.

As we have seen in the example, argumentation theory can give us a tool for a decision-making support. This links argumentation to research done as well in artificial intelligence, as in philosophy, in which conditions

for rational decision-making have been intensely studied.

The questions classically asked with respect to this problem are connected with finding choices maximizing an agent's profit, or by making actions to fulfill own goals. However, another important aspect of decision-making is the possibility of explanation for a taken decision. Imagine that the previously described bike rider was planning to go for a trip together with her child. Then, if she decides to cancel the trip, it is not enough to determine that it is beneficial for her to do it. It is equally important to give the child an explanation *why* the trip has been cancelled.

Such an explanation can be provided in terms of argumentation theory (e.g Amgoud & Prade, 2009; Kakas & Moraitis, 2003). As Amgoud and Prade (2009) suggest, arguments can be associated with decisions which they either support, or undermine. For example, the argument that it is a good idea to take a trip because the weather is good supports the decision to go biking, while the argument that it is going to rain undermines it. Further, as we mentioned before, abstract argumentation theory provides natural criteria for acceptance of arguments. Then, if we identify which arguments are in favor of some decision, and which in favor of their rejection, we can find a justification of a decision based on accepted arguments.

In addition to the previously made points, we can notice that not all arguments are equally convincing to particular participants. It might be that some pieces of information included in the discussion are not reliable, or that some arguments were provided by a highly respected source. Further, arguments might appeal to particular values, which are of diversified importance to a selector of arguments. It is then plausible to assume that an attack on a strong argument from an argument of little importance should not be taken into account.

Suppose that in the previously described example the friend advising against going for a trip in fact refers to an information taken from a completely unreliable website. Meanwhile, the decision-maker bases her initial belief on a serious forecast. Then, it is not rational for the decision-maker to treat the available arguments as equal.

However, this point is not in line with the classical approach to *abstract argumentation* following Dung (1995). There, all arguments are atomic and their strength is uniform. Their acceptance relies purely on the structure of attacks between them. While this assumption allows for the high simplicity of the model, it is far from capturing argumentation between human agents plausibly.

Several approaches towards capturing the differences in the strength of

arguments have been introduced. Some of them involve assigning numerical values to arguments, constituting their strength (e.g. Dunne, Hunter, McBurney, Parsons, & Wooldridge, 2011). However, the plausibility of assigning arguments precise numbers which indicate their strength is difficult to justify when modeling argumentation between human agents is concernced. Overcoming this issue is one of the benefits of the qualitative approach to determination of the strength of arguments.

One of such approaches, *value-based argumentation*, was provided by Bench-Capon (2003). In this framework it is assumed that arguments appeal to specific values which are of a distinctive importance to a particular decision-maker. Then, an attack can be blocked from her perspective if she ranks the value of the attacked argument higher than the value of its attacker. This approach is suited to the problem of argumentation-based decision-making, in which factors different than credibility of information are important while assessing the acceptability of an argument. Also, it provides a clear justification for the determined strength of arguments. This is important when an argumentation serves as a support for decision-making; justification of the strength of arguments contributes to the justification of a decision. It is worth noting that this constitutes a strong advantage of this approach over assigning preferences over arguments directly, as it is the case in the *preference-based argumentation* (e.g Amgoud & Cayrol, 1998). Bench-Capon's approach makes sure that agents only consider some arguments as stronger than another, if they have a good reason to do so. Value-based argumentation will constitute the basic framework used in the thesis and will be described at length separately.

Furthermore, it is worth noting that argumentation is an inherently multi-agent phenomenon. It often occurs when agents exchange information, aiming at reaching a collective view with respect to some issue. In the previously discussed example, it was up to one agent to decide whether a trip should take place or not. However, the important information came from another agent, who could have been interested in successfully persuading her interlocutor not to take a trip. Further, described agents could have been planning to decide upon going together. Then, their collective decision would not only be dependent on the information that they exchange, but also on the collective view regarding the strength of arguments that they reach.

In the recent literature regarding abstract argumentation a growing interest in application of multi-agent systems techniques in modeling debates can be observed (e.g. Maudet, Parsons, & Rahwan, 2006; Bodanza, Tohmé,

& Auday, 2017). However, it is uncertain how to conceptualize the multi-agent character of argumentation. While multiple approaches towards solving this problem have been provided, I will focus on methods of *aggregation* associated with the social choice theory.

Within this approach two main types of aggregation can be distinguished. In the first of them, finding the collective view between agents' views on the outcome of deliberation is considered. In the second, a disagreement between perceived structure of arguments is taken into account. Then, aggregation of individual argumentation frameworks is studied.

With respect to aggregating outcomes of deliberation, application of judgment aggregation has been widely investigated (e.g. Caminada & Pigozzi, 2011; Awad, Booth, Tohmé, & Rahwan, 2015; Awad, Bonnefon, Caminada, Malone, & Rahwan, 2017; Awad, Caminada, Pigozzi, Podlaszewski, & Rahwan, 2017). Here, following the labeling based argumentation semantics (see, e.g. Caminada, 2008), agents are allowed to judge arguments as either accepted, rejected, or undecided. Then, judgments of this kind are aggregated to obtain a collective labeling. Another line of research is associated with merging the sets of arguments accepted by particular agents (Delobelle et al., 2016).

The second mentioned approach assumes that the differences in the outcomes of discussion from perspectives of particular agents are determined by differences in their perceived structure of argumentation. Multiple reasons for disagreements of this kind can be conceived of. They can be caused by the differences in interpretation of arguments themselves, which can be a major problem while reconstructing argumentation structure from natural language. Further, application of merging distinctive argumentation frameworks can be helpful while modeling argumentation in which agents do not have full access to existing arguments. Then, merging argumentation graphs can provide all participants of a debate with arguments that only some have access to. Another cause of differences in perceived structure of argumentation can be associated with differences of views on arguments' relative strength. Then, merging argumentation graphs can be associate with merging views on their strength.

As we have seen, the applications of techniques originating in multi-agent systems to argumentation theory help to capture the phenomenon of arguing plausibly. However, lifting the argumentation theory to the multi-agent level opens the possibility for agents to misrepresent the information available for them, in order to improve the outcome of discussion for themselves.

In the literature regarding multi-agent systems the possibility of agents'

strategic behavior has been widely studied (e.g. Gärdenfors, 1976; Brandt, Conitzer, Endriss, Lang, & Procaccia, 2016). Strategic behavior, or manipulation, is understood in this context as providing false information by an agent in order to receive a better outcome for herself. When collective decision-making is considered, attempts of manipulating the outcome of a decision procedure are not desirable. Collective decision systems aim at providing an outcome fair for all parties, under assumption that the information that they give as an input for the mechanism is accurate. Misrepresentation of a part of an input can distort the procedure and in the end induce an unfair result. This is the reason why engineering systems in which manipulation is never beneficial for any agent is of a high interest. I will also refer to such systems as *strategy-proof*.

Further, in case of systems or procedures which are not strategy-proof, but which enjoy other desired properties and are well suited to their applications, studying the computational complexity of manipulation is of great importance (e.g Caminada, Pigozzi, & Podlaszewski, 2011). The motivation for such a study is that a procedure in which computational complexity of manipulation is high enough to exclude the possibility of finding a beneficial way of misrepresenting own information in an efficient way can be treated as strategy-proof for practical purposes.

The problem of manipulation is highly relevant to the setting of multi-agent argumentation (e.g. Caminada et al., 2011). It is especially important when argumentation based decision systems are considered. Intuitively, the goal of collective solving argumentation problems is to select the best arguments taking into account all relevant information that agents have at disposal and to fairly combine views on the strength of arguments. Agents can have preferences over accepted arguments, for instance if acceptance of some distinguished arguments is determining the choice of some decision. Then, they might decide not to submit arguments that they know about, as considered by Rahwan and Larson (2008). Also, they can misrepresent their views on the strength of arguments to ameliorate the outcome of discussion for themselves. This is the reason why it is important to study the possibility of manipulation or the computational complexity of strategic behavior in such mechanisms.

1.1 Research Questions

The research of this thesis is situated along the lines of previously described points. We are interested in the behavior of agents who aim to ensure

that a certain decision is made by enforcing some view on the strength of available arguments. We will follow a particular model capturing the argument strength, namely the value-based approach.

Our investigations will be performed at two levels. Firstly, we will consider situations in which a single agent is responsible for making a decision. We will assume that she has an incentive to push a decision forward. Then, she is looking for a settlement of strength of arguments which would provide a justification for her desired decision.

The second direction covered in this thesis involves a situation in which a group of agents aims at selecting a decision collectively. However, they disagree upon the strength of particular available arguments. We will study how an agent, willing to push some desired decision forward, can manipulate the process of reaching a collective view with respect to the strength of arguments. The basic question in this approach is how to account for reaching such a collective view. We will study methods for reaching an agreement with respect to this problem using methods originating in social choice theory. Having established such methods, we will study the manipulation problem within them.

1.2 Structure of the thesis

In Chapter 2, I will provide basic definitions and results used in the remainder of the thesis. I will start with presenting the framework of abstract argumentation, with a special focus on value-based argumentation. I will define it and describe its philosophical motivation. Further, in Chapter 3, I establish results for the single agent case. Chapter 4 lifts the results obtained earlier to the multi-agent case. I consider two methods of aggregating views on strength of arguments within the value-based argumentation setting and study connections between them. Then, I investigate strategic behavior in considered settings. Finally, in Chapter 5, I provide conclusions and directions for further research.

Chapter 2

Preliminaries

In this part of the thesis I will present basic concepts and definitions used in the further chapters. I will begin with describing abstract argumentation theory, as defined by Dung (1995). Further, I will discuss the value-based argumentation, following Bench-Capon (2003). Finally, I will define several notions which will be used in subsequent parts of the thesis, such as distances between orderings.

2.1 Abstract argumentation

The setting employed in the current work is based on the model of argumentation provided by Dung (1995). In his view, argumentation is conceived as a set of arguments and a binary relation expressing which arguments attack which. Formally, this setting is defined as follows:

Definition 1. *An argumentation framework (AF) is a pair* $AF = \langle A, \rightarrow \rangle$, *where:*

- A *is the set of arguments*
- $\rightarrow \subseteq A^2$ *is the attack relation*

So, an argumentaion framework is a directed graph, where nodes are the arguments, and the edges are the attacks. Figure 2.1 displays an example of an argumentation framework, where $A = \{A, B, C, D, E, F, G\}$ and $\rightarrow = \{\langle B, A\rangle, \langle D, A\rangle, \langle E, D\rangle, \langle C, B\rangle, \langle F, C\rangle, \langle F, E\rangle, \langle F, G\rangle, \langle G, F\rangle\}$.

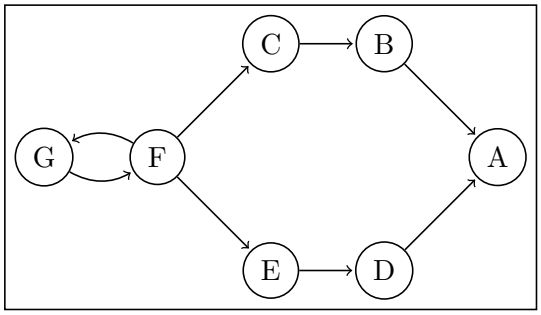


Figure 2.1: Example of an argumentation framework.

It is worth noting that the presented model does not capture the notion of support that arguments might provide for each other directly. However, such a notion is needed to provide plausible criteria of acceptance of sets arguments. In the discussed approach a support for an argument is understood as undermining credibility of its counterarguments. Consequently, it is the intuition that an argument a all attackers of which are attacked by some argument b, is supported by b. So, support for an argument is identified with its defense. Further, a set of arguments S is said to defend an argument a if for any attacker of a there is some member of S which attacks it.

Definition 2 (Defense)**.** *Given an argumentation framework $AF = \langle A, \rightarrow \rangle$, a set of arguments $S \subseteq A$ and some argument $a \in A$, S* defends *a iff for any $b \in A$ such that $b \rightarrow a$ there is an $a' \in S$ such that $a' \rightarrow b$. We say that S defends a set of arguments $S' \subseteq A$ iff S defends all $a \in S'$. A function $F : 2^A \rightarrow 2^A$ assigns every $S \subseteq A$ the set of all arguments that S defends. Also, S is said to be* self-defended *if S defends S.*

The notion of defense is then used to determine when a set of arguments can be rationally selected as an outcome of a discussion. Classically, the following criteria (i.e, semantics) for selecting sets of arguments (called extensions) have been considered (Dung, 1995):

Definition 3 (Argumentation Semantics)**.**

Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, and $S \subseteq A$. S is :

- **Conflict-free:** iff there are no $a, b \in S$ such that $a \rightarrow b$. We refer to the set of all conflict-free extensions of AF as CFR_{AF}.

- **Admissible:** iff S is conflict-free and self-defended. We refer to the set of all admissible extensions of AF as ADM_{AF}.
- **Complete:** iff S is admissible and $F(S) = S$. We refer to the set of all complete extensions of AF as CMP_{AF}.
- **Grounded:** iff S is the minimal complete extension of AF w.r.t. set inclusion. We refer to the grounded extension of AF as $GRND_{AF}$ [1].
- **Preferred:** iff S is a maximal complete extension of AF w.r.t. set inclusion. We refer to the set of all preferred extensions of AF as PRF_{AF}.
- **Stable:** iff S is conflict-free and for any $a \in A$ such that it is not the case that $S \rightarrow a$, $a \in S$. We refer to the set of all stable extensions of AF as STB_{AF}.

If S satisfies the condition of some argumentation semantics σ, we call it a σ-extension of AF.

Intuitively, the conditions of being conflict-free, self-defended and completenes can be considered as necessary for acceptability of an extension as an outcome of a discussion. If some set violates the first of them, then an agent selecting it accepts that two pieces of information are true even though they are in conflict with each other. Further, if a set of arguments fails to satisfy the second condition, then a decision-maker selecting it is forced to accept that there is a piece of information undermining a statement that she considers as true and she fails to justify why the attacker should not be considered. Finally, if the selected set of arguments is not complete, then an agent fails to accept all arguments whose attackers are undermined by the selected arguments.

Other mentioned semantics can be treated as approaches towards selecting optimal complete extensions. The described semantics can be compared by the level of credulousness assumed when they are chosen as a criterion for selection of arguments. Clearly, a skeptical selector should be willing to choose the minimal complete extension, so she should prioritize the grounded semantics. This approach can be useful when the goal is to accept only the most reliable pieces of information.

On the other hand, a credulous selector might want to choose a maximal complete extension, following the preferred semantics. It is easy to show that stable semantics are also preferred.

[1] The grounded extension is always unique.

Further, it is easy to show that the mentioned semantics are included in each other. Here by inclusion of semantics σ in semantics σ' we mean that each σ-extension is also a σ'-extension. The hierarchy of argumentation semantics is depicted in Figure 2.2. Arrows correspond to the inclusion relation.

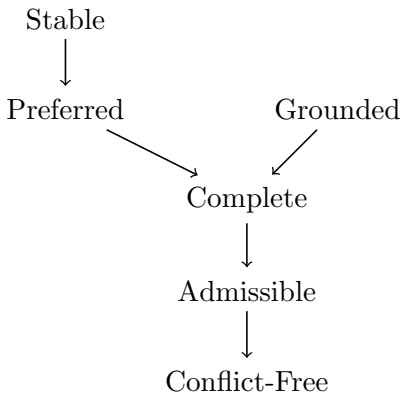


Figure 2.2: Inclusion of argumentation semantics.

Apart from the appropriateness of particular semantics to desired application, an important factor in deciding which of them should be selected as a rationality criterion is its computational complexity. Naturally, if some semantics is supposed to be used in practice, it is desired for it to be easily computable. While many decision problems are studied with respect to argumentation semantics, some of them will be particularly important in the remainder of the thesis.

It is worth noting that some of the mentioned semantics do not always provide the unique outcome of a discussion. Therefore, additional measures are needed in order to establish the set of selected arguments in case of semantics outputting multiple extensions. Two main ways of selecting accepted arguments, skeptical and credulous, are considered in the literature.

The skeptical acceptance condition requires that an argument is a member of all the extension under the chosen semantics.

Definition 4 (Skeptical Acceptance)**.** *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, $a \in A$ be an argument and σ be some argumentation semantics. We say that a is skeptically accepted with respect to σ iff for any σ-extension S of AF, $a \in S$.*

On the contrary, we might want to accept an argument if it is a member

of at least one extension.

Definition 5 (Credulous Acceptance)**.** *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework, $a \in A$ be an argument and σ be some argumentation semantics. We say that a is credulously accepted with respect to σ iff there is some σ-extension S of AF such that $a \in S$.*

The introduction of the skeptical and credulous acceptance conditions raises a question concerning the computational complexity of acceptance of arguments. In the thesis we will only consider the credulous acceptance.

Let us rephrase the definition of credulous acceptance as a decision problem.

CREDULOUS ACCEPTANCE(σ)

Instance: Argumentation framework $AF = \langle A, \rightarrow \rangle$, $a \in A$.
Question: Is a in at least one σ-extension of AF?

The complexity of the credulous acceptance is shown in the Table 2.1 [2]. This summary follows (Dunne & Wooldridge, 2009).

Semantics	Complexity of Credulous Acceptance
GRND	P
PRF	NP-complete
STB	NP-complete

Table 2.1: Complexity of credulous acceptance problem

It is worth noting that for both skeptical and credulous acceptance we can find examples of argumentation frameworks and argumentation semantics for which the set of accepted arguments is not an extension of the desired semantics. Consider for instance the framework displayed in Figure 2.3 and preferred semantics.

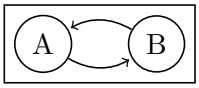


Figure 2.3: Preferred extensions: {A}, {B}

[2] See section 2.3 for a brief introduction to the computational complexity theory.

It is easy to check that in this case two preferred extensions exist: {A} and {B}. Then clearly the set of skeptically accepted arguments amounts to $\emptyset$, while the set of credulously accepted arguments is {A, B}. But neither the empty set, nor $\{A, B\}$ is a preferred extension of AF.

This observation raises problems with respect to the application of argumentation semantics to justifiable decision-making. If a semantics possibly outputs multiple extensions, a decision maker might be forced to either make a selection of arguments not complying with the chosen rationality criterion, or make an arbitrary choice among extensions of the chosen semantics. This is why application of semantics which always provide a single extension are of interest. The grounded semantics is an example of such semantics.

2.2 Value-based argumentation

In the recent literature regarding modeling specific applications of abstract argumentation theory, it has been argued that in order to capture the specificity of argumentation about decisions, it is needed to take into account the values to which arguments appeal (e.g Bench-Capon, 2003; Bench-Capon, Doutre, & Dunne, 2007; Modgil, 2009). This approach is referred to as *value-based argumentation*.

The motivation for the value-based argumentation approach, as proposed by Bench-Capon is to a large extent following a philosophical analysis of practical argumentation presented by Perelman (1971). This motivation is based on the insights from reasoning patterns in law or ethics. However, it can be plausibly applied in discussions about choice of actions broadly construed. It is claimed that in a discussion concerning certain practical decisions arguments are not primarily aimed at assessing the truthfulness of pieces of information. Instead, the discussion is concerned with the appropriate usage of available information towards reaching a collective view about some decision. It is also natural to observe that presenting arguments in such a discussion has a clearly specified goal. Namely, arguments are aimed at convincing a body responsible for making a decision that the decision should be made, or rejected, in accordance with the rhetorician's preferences.

Having made the discussed observation it can be noted that the decision-making body does not necessarily treat the presented arguments in an equal way. Some of the arguments might be more persuasive than another for particular assessors. Also, as Bench-Capon argues, these preferences are not always justifiable by rational reasoning. Instead, he submits that an

audience of argumentation assesses the strength of arguments relying on the importance of values that they appeal to. The term audience, used commonly in the literature on value-based argumentation, refers to a point of view on the hierarchy of values. It does not presuppose that there are multiple agents in an audience.

Following the described points, value-based argumentation assumes that an audience of a discussion can establish the relative strength of arguments on the basis of importance of values to which arguments appeal. Consequently, an attack on an argument appealing to a higher value than its attacker, can be disregarded by a relevant audience. As a result, some particular decision-makers can be persuaded by a given argumentation to a different extent than others.

Let us illustrate the presented line of reasoning on an example of a specific debate regarding making a practical decision.

Example 2.2.1. *(Airiau, Bonzon, Endriss, Maudet, & Rossit, 2016) 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 appealed to environmental responsibility (ER), B, C to social fairness (SF), F to economic viability (EV) and D, E - to infrastructure efficiency (IE).

These arguments can be represented as on the argumentation graph with a mapping of values depicted on the Figure 2.4. For each argument, the first

element of its description is its name, and the second is the name of the value it appeals to.

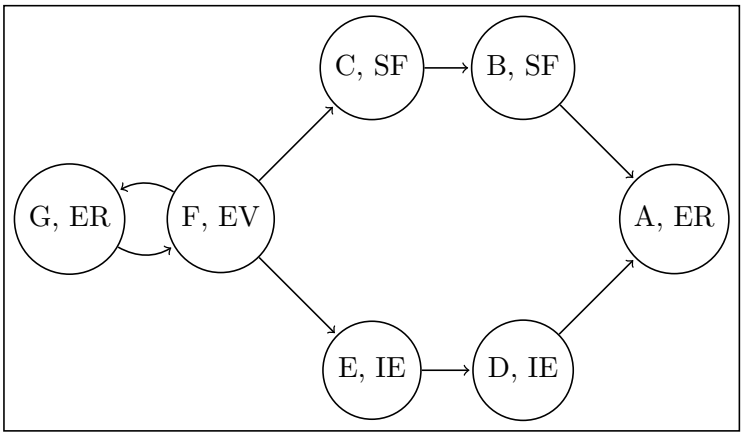


Figure 2.4: Argumentation structure for the example.

Let us now consider the structure of this discussion from the perspectives of two members of a decision-making jury. For the first of them, economic viability and infrastructure efficiency, which are equally strong for her, are more important than social fairness or environmental responsibility. She does not differentiate them. 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 following structure is obtained, after the elimination of disregarded attacks:

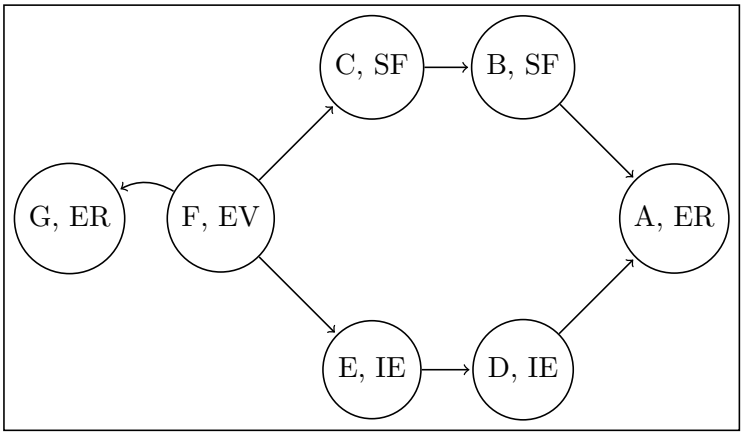


Figure 2.5: Argumentation from perspective of the liberal (EV $\sim$ IE $\succ$ SF $\sim ER$)

Clearly, some of the arguments appealing to economical viability or infrastructure efficiency are now in a better position than before. However, some arguments corresponding to different values cannot be accepted in the new structure.

Let us now consider another member of the jury, who believes that social fairness is the most important value. She ranks environmental responsibility second, and economic viability third. Finally, she considers infrastructure efficiency as the least important. From her perspective, the structure of successful attacks is much different.

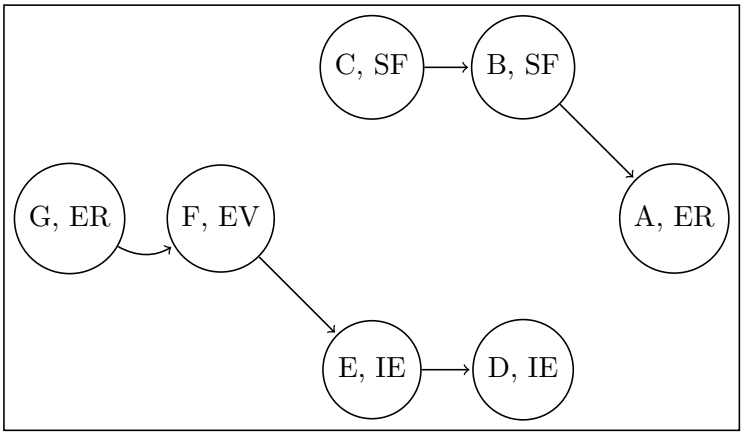


Figure 2.6: Argumentation from perspective of the left-winger (SF $\succ$ ER $\succ$ EV $\succ$ IE)

We can clearly observe that from her perspective arguments appealing to environmental responsibility or social fairness are in a much better position than from the perspective of the previously considered jury member. It is worth noting, however, that in the present setting some attacks hold regardless of the chosen preference ordering over values. If a pair of arguments attacking each other shares the same value, it is in conflict from any audience's perspective.

Let us now proceed to providing the formal account for the presented intuitions. The basic concept is that of the value-based argumentation framework. It is an extension of the abstract argumentation frameworks, presented in the Section 2.1. In addition to the set of arguments and a binary attack relation, a set of values and a function mapping them to arguments are taken into account.

Definition 6. *A* value-based argumentation framework *(VAF) is a tuple* $VAF = \langle A, \rightarrow, V, val \rangle$*, where:*

- A *is the set of arguments*
- $\rightarrow \subseteq A^2$ *is the attack relation*
- V *is the set of values*
- $val : A \rightarrow V$ *is the function assigning values to arguments*

Furthermore, in order to establish the relative strength of arguments, it is needed to provide a preference ordering over values. This move helps to determine what is the impact of arguments for a particular audience.

Definition 7. *Let $VAF = \langle A, \rightarrow, V, val\rangle$. An audience P is a reflexive and transitive relation (a preorder) over V. We denote that a value v_1 is at least as preferable as v_2 for P as $v_1 \succeq_P v_2$.*

Let us further introduce useful notations expressing types of relations between particular values.

Notation 1. *For a pair of values $v_i, v_j \in V$ and a given audience P, we say that $v_i \succ_P v_j$ iff $v_i \succeq_P v_j$, but it is not the case that $v_j \succeq_P v_i$. We say that $v_i \sim_P v_j$ iff $v_i \succeq_P v_j$ and $v_j \succeq_P v_i$. Also, we say that $v_i \bowtie_P v_j$ iff it is not the case that $v_i \succeq_P v_j$ or that $v_j \succeq_P v_i$. Finally, given a set of values V we call the preorder $P = \{v_i \succeq v_j | v_i = v_j\}$ the empty preorder over V.*

For the clarity of presentation, when examples of audiences are given, reflexivity is ommited.

Given these notions we can define what does it mean for an argument to defeat another from the perspective of a particular audience.

Definition 8. *Let $VAF = \langle A, \rightarrow, V, val\rangle$ be a VAF and P be an audience. We call a pair $aVAF = \langle VAF, P\rangle$ an* audience specific *VAF ($aVAF$). 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 $val(b) \succ_P val(a)$.*

Intuitively, it is stipulated that an audience might reject an attack on an argument, if it is stronger from their perspective than the attacking argument. This difference in strength is induced by the difference in the values that the arguments appeal to and the ordering over values.

Then, the argumentation framework on which a *VAF* is based can be transformed into a new framework, taking into account the values that arguments carry and preferences over them.

Definition 9. *Let $VAF = \langle A, \rightarrow, V, val\rangle$ and P be an audience. The defeat graph of VAF based on P is an argumentation framework $VAF_P = \langle A, \rightarrow^P\rangle$.*

It is worth noting that in this model attacks between arguments sharing the same value cannot be blocked. This can be seen as a factor contributing to the plausibility of value-based argumentation. It is only possible for an

agent to block an attack, if she has a reason to believe that some argument is stronger than another.

Additionally, following Bench-Capon (2003) we may consider that any plausible *VAF* does not include cycles with all arguments assigned the same value. This restriction is motivated by an observation that such a cycle would not be breakable under any preference ordering over values, and further allowing for a single outcome of the discussion. We will refer to cycles of this kind as monochromatic.

Definition 10 (Monochromatic cycle)**.** *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a VAF. We call a cycle of attacks* $C = a \rightarrow \cdots \rightarrow a$ *monochromatic iff for any* $b_1, b_2 \in C$, $val(b_1) = val(b_2)$.

Then, for any given audience a decision regarding the fair outcome of deliberation can be made by applying standard argumentation semantics to its respective defeat graph.

Clearly, specifying agents' preferences over values as arbitrary preorders contributes significantly to the cognitive plausibility of the current setting. It leaves room for agents' uncertainty about ordering of values. In this way we can allow for agents who are not sure about orderings of particular pairs of values, or treating them as equally important.

However, it is worth noting that ensuring that agents preferences are associated with linear orderings helps to secure beneficial computational properties of induced defeat graphs. It has been shown that, under the assumption that a *VAF* does not include any monochromatic cycles, then for any audience associated with a linear order over values, the defeat graph is acyclic.

Theorem 1. *(Bench-Capon, 2003) For any* $VAF = \langle A, \rightarrow, V, val \rangle$ *with no monochromatic cycles and an audience* P *associated with a linear ordering over* V*, the defeat graph* $VAF_P = \langle A, \rightarrow^P \rangle$ *is acyclic.*

This is an important feature of argumentation frameworks both from the perspective of computational complexity of computing argumentation semantics and from the perspective of appropriateness of use of particular semantics as a rationality constraint for selection of arguments. When an argumentation framework is acyclic, it only has a single, nonempty preferred extension.

On the other hand, allowing for specifying agents' preferences over values as arbitrary preorders allows for higher flexibility in terms of specifying relative strength of arguments. As a consequence, an agent allowed to only

specify her ranking of values as a linear ordering can block attacks in fewer ways than an agent specifying her preferences as arbitrary preorders.

As an example, consider the *VAF* displayed in Figure 2.7.

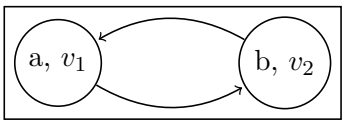


Figure 2.7: Non-monochromatic cycle

Notice that the argumentation framework displayed in Figure 2.8 is a defeat graph of the *VAF* for the empty preorder over $V = \{v_1, v_2\}$. However, any linear ordering over V would require one of the attacks to be eliminated.

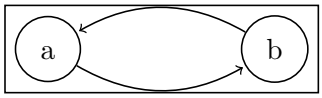


Figure 2.8: Not a defeat graph for linear preferences.

Due to this observation, particular properties shown in the thesis will be shown for particular types of orderings over values. While the main distinction will be made between arbitrary preorders and linear orderings, results concerning the case when preference orderings over values are connected preorders [3] will be provided.

2.3 Computational complexity

In the current work it is of major interest to study the computational complexity of agents' behavior. I will provide definitions of classes of complexity of problems which will be used further.

The computational complexity of a problem is a restriction of the amount of resources required to execute the best algorithm solving it, expressed as a function of the size of an input. These resources are understood as *time* and *space* needed to execute it. In this thesis results will be restricted to time. We define a problem as a set of inputs satisfying a certain property. Then, for a given input, we want to determine if it is a member of this set, or not.

[3] Connected preorders are preorders such that for any pair of items V_i, v_j, $v_i \succeq v_j$ or $v_j \succeq v_i$.

In order to define some important classes of complexity we need a notion of an *oracle*. An oracle for some problem PROBLEM provides a solution to PROBLEM in one unit of time.

Let us now list out the definitions of classes which will be used in the thesis.

- **P**: A decision problem PROBLEM is in the class P if it is computable by some algorithm with runtime bounded by a polynomial $f(n)$, where n is the size of the input for PROBLEM.
- **NP**:A decision problem PROBLEM is in the class NP if for a given input a we can guess the output for a and there is a polynomially computable algorithm checking if the guess was correct.
- $\mathbf{\Theta_p^2}$: A decision problem PROBLEM is in the class Θ_p^2 if it is computable in polynomial time with an access to the oracle for some problem in the class NP, where the number of times in which the oracle is accessed is bounded by some logarithmic function $f(n)$, where n is the size of an input for PROBLEM.

In order to determine if some problems are harder than others, we use a notion of *reduction* of problems. We say that a problem PROBLEM_1 is reducible to a problem PROBLEM_2 if there is a polynomially computable function f such that for any input a of PROBLEM_1, $f(a)$ is an input of PROBLEM_2 and a is in PROBLEM_1 if and only if $f(a)$ is in PROBLEM_2.

Further, we say that a problem PROBLEM is *C-hard* with respect to some complexity class C, if any member of C is reducible to PROBLEM. Then, we say that PROBLEM is *complete* with respect to C if it is C-hard and it is in C.

2.4 Distance between preference orderings

In the current setting preference orderings play a crucial role. They are the basis for determining the relative strength of arguments based on the values they appeal to, and they will be fundamental in establishing collective argumentation structures.

When we consider a number of agents with distinctive preferences over values, we might want to ask to what extent their positions are different from each other. This can be achieved by providing a distance metric between preference orderings. Deza and Deza (2009) provide an extensive overview of the literature on distances.

Establishing a distance between two preference orderings aims at providing a measure of how two views on importance of particular items of some set are different from each other. The main line of research on distances between orderings focuses on linear orderings, in the desired setting preference orderings are preorders. Thus, distance metrics need to be adapted to this application. In the reminder of the thesis we will not focus on particular metrics, but rather study classes of distances satisfying particular computational properties. In this section I will provide an example of applicable distances over preorders.

One of such metrics is the Hamming distance. Given two arbitrary preorders, it indicates the number of disagreements between them.

Definition 11 (Hamming distance)**.** *Let P_1, P_2 be preorders over a set V. The Hamming distance between P_1 and P_2 (denoted as $HD(P_1, P_2)$) is the number of disagreements between those pre-orders. It is the number of elements $a, b \in V$ such that $a \succeq_{P_1} b$ and $a \not\succeq_{P_2} b$, or $a \succeq_{P_1} b$ and $a \not\succeq_{P_2} b$*

Let us illustrate the concept of distances between preorders on the example.

Example 2.4.1. *(Continuation of Example 2.2.1.) Recall that we were considering two positions in the debate regarding the possible ban of Diesel cars: one of them was put forward by a left-winger, and another by a liberal. Let us also consider a neutral approach, in which all values used in the debate are not comparable. We know that those positions are different from each other. We would like to know, however, to what extent do they differ. Let us recall agents' preferences over values.*

1. $\{SF \succeq ER, ER \succeq EV, EV \succeq IE\}$
2. $\{EV \succeq IE, IE \succeq EV, EV \succeq SF, EV \succeq ER, IE \succeq SF, IE \succeq ER, SF \succeq ER, ER \succeq SF\}$
3. $\emptyset$

Let us now compare Hamming distances between these orderings. $HD(1,2) = 7$, $HD(1,3) = 3$, $HD(2,3) = 8$.

Chapter 3

Single agent setting

The goal of this chapter is to investigate the scenario in which a single agent is about to make a decision which needs to be justified. The type of justification which I will study is understood as a support of some ordering over values to which relevant arguments appeal. In this chapter it will be assumed that a decision-maker can have an incentive to make a decision which is not in line with her sincere preferences over values. Then, she might be willing to submit an insincere preference ordering as a justification of her choice of decision. This is not a desirable behavior and we would like the designed decision system to be immune to this kind of manipulation.

For the sake of simplicity of the setting, in this chapter I will focus on agents who wish to push the decision forward, not those who wish the decision not to be made.

I will begin by providing a motivation and formal account for this kind of justification of decisions. I will do it in Section 3.1. Further, in Section 3.2, I will study the complexity possibility of finding any preference ordering justifying agents' decision. Then, in Section 3.3, I will investigate the hardness of finding such an ordering which is minimally different than the agents' sincere hierarchy. In Section 3.4 describe some relevant connections of this problem with properties of particular value-based argumentation frameworks.

3.1 Decisive arguments

It is often the case that arguing agents aim at reaching a decision. We can say that there are some points in the discussion which clearly determine what should be decided, such as "We should go to war". However, some of them are not sufficient for resolving the issue at stake. For instance, an

argument "A lot of soldiers would die during war" can be used as a support for the pacifist view. Nevertheless, accepting it does not determine that the country would not go to war. It is worth noting that this point is different from only assuming that arguments are in favor of a decision or in favor of not taking it. If an argument stating that a decision should be taken is pushed forward, the decision should be accepted. This is independent of the balance of arguments which are in principle for or against the decision.

The described intuition can be captured by mapping information about a decisive support for certain decisions to a subset of considered arguments.

Definition 12. *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework. We call $DP = \langle AF, D \rangle$ a decision problem, where $D \subseteq A$ is a set of decisive arguments.*

A particular class of argumentation problems involves deliberation about performing a single action. In such an argumentation one argument is decisive for this action. We can associate it with a statement that the decision should be made. We refer to such argumentations as to binary argumentation problems. In the thesis I will focus on this type of decision problems. Unless specified otherwise, by *decision problem* I will refer to binary decision problems.

Definition 13 (Binary decision problem)**.** *Let $DF = \langle AF, D \rangle$ be a decision problem. DF is a binary argumentation problem iff $|D| = 1$.*

For convenience I will sometimes call a pair $\langle VAF, D \rangle$, were VAF is some value-based argumentation framework and D is an an argument a decision problem as well.

To illustrate this intuition, let us continue the example employed in the previous chapter.

Example 3.1.1. *(Airiau et al., 2016) Recall that in the Example 2.2.1 we were taking into account a debate regarding a possible ban of diesel cars, aimed at reduction of air pollution in big cities. Let us now assume additionally that the jury is not deciding upon any other actions during the meeting. The only possible outcomes of the decision process are that diesel cars are banned, or they are not.*

Then, despite the complexity of the structure of the argumentation, the result of the discussion is binary - either the city bans Diesel cars, or it does not. So, we consider only one decisive argument - A. If this argument is accepted, we should ban Diesel cars, and otherwise we should not.

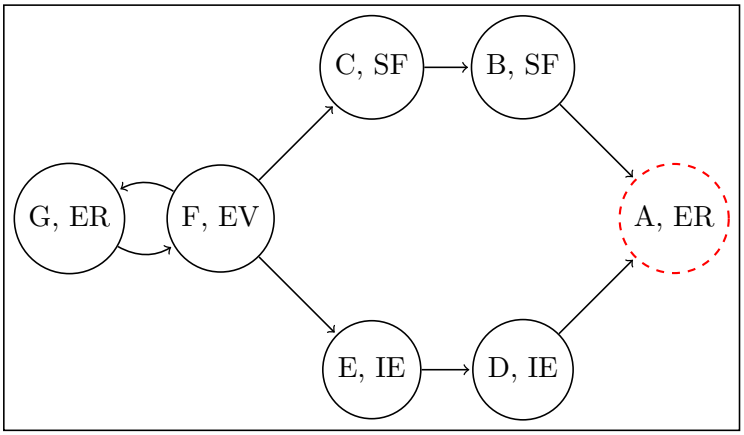


Figure 3.1: Argumentation structure with A marked as a decisive argument.

Notice now that as the arguments presented in the example clearly relate to particular values, preferences over them determine the relative strengths of arguments. Therefore, we can easily imagine that an agent who has an interest in pushing a decision forward would like to impose a ranking over values ensuring that the decision is made. To capture this behavior, let us introduce the notion of *preservability* of an argument.

Here, given a *VAF* and an argument we are looking for a preference ordering under which the argument is credulously accepted with respect to chosen semantics, as introduced in the Definition 5. This notion has been introduced as subjective acceptance of an argument by Bench-Capon (2003).

Definition 14 (Preservability)**.** *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a VAF,* $a \in A$*, and* σ *be an argumentation semantics. We say that* a *is* σ*-preservable iff there is an audience* P *such that* a *belongs to some* σ*-extension of the defeat graph* $AF = \langle A, \rightarrow^P \rangle$*.*

To illustrate this term, consider again the debate described in the previous example.

Example 3.1.2. *(Continuation of Example 3.1.1)*

Let us imagine that the decision is left to a single agent - a mayor. However, she also happens to be an owner of a factory of electric cars, so she would be highly interested in passing the ban. But she still needs to justify her decision. In the city it is customary to consider belonging the grounded extension as a fair justification of acceptance of an argument.

In reality, the mayor does not care about the environment at all. In fact, she does not have any preference over values, she only cares about

the decision being made, because this would give her a lot of money. So, in her ordering over values all values are equal. Thus, with respect to her preferences the induced defeat graph is of the form depicted on the Figure 3.2.

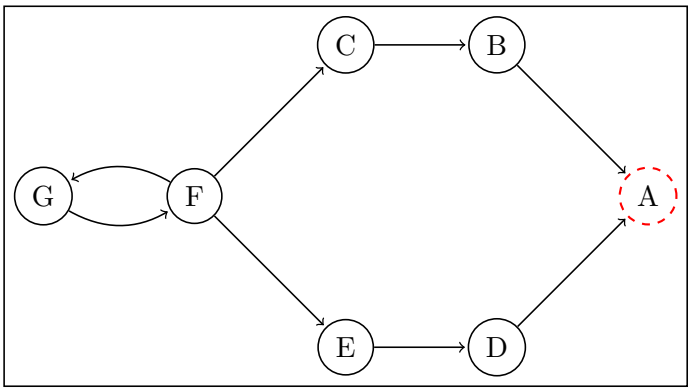


Figure 3.2: Defeat graph for mayor's sincere preference ordering (ER ⋈ IE ⋈ EV ⋈ SF).

It is easy to check that in this case the grounded extension is the empty set, so A is clearly not a member of it. So, the mayor cannot justify her decision with this ranking over values. However, it would be sufficient for her to pretend that the environment is more important for her than economical issues to transform the argumentation to the form in which her decision is justified. She decides to submit an insincere ordering over values is of the form:

$$ER \succ EV \succ SF \succ IE$$

Which induces the defeat graph presented in the Figure 3.3.

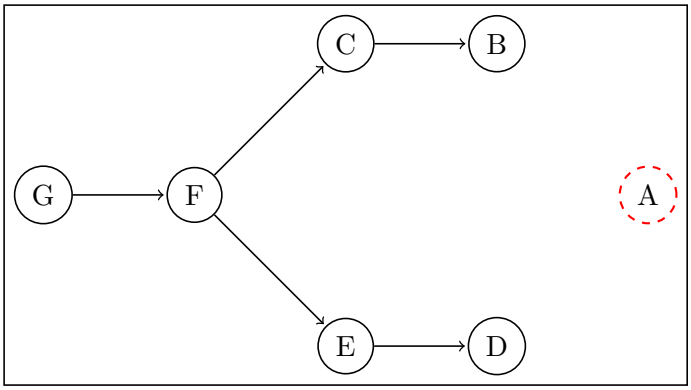


Figure 3.3: Defeat graph for mayor's insincere preference ordering.

In this graph, the grounded extension is $\{G, C, E, A\}$*. So now, clearly choosing* A *is justified. Thus,* A *is preservable.*

3.2 Single agent complexity results

Let us proceed to the determination of the computational complexity of finding preference orderings over values preserving decisive arguments. The preservability problem was studied in the literature under strong assumptions. Namely, it was assumed that agents specify their preferences over values as linear orderings and that there are no monochromatic cycles in *VAF*s (e.g. Dunne & Bench-Capon, 2004). We are interested in generalizing the results with respect to this problem to account for agents who are only willing to specify their preferences as arbitrary preorders. We will further study the complexity of finding a preference ordering preserving the decisive argument which is minimally different from the agent's sincere ordering.

For simplicity of this endeavor let us rephrase the definition of preservability as a decision problem.

PRESERVABILITY(σ)

Instance: $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$.
Question: Is there an audience P such that a is credulously accepted in the defeat graph of VAF based on P with respect to semantics σ?

The problem has been shown to be NP-complete under the assumption that audiences are associated with linear orderings over values and that

there are no monochromatic cycles in the considered *VAF*s (Dunne & Bench-Capon, 2004). It has also been shown that this theorem also holds when the structure of the considered *VAF*s is restricted to binary trees (Dunne, 2007).

It is worth noting that if we assume, following previous work on preservability, that preferences over values are strict and that there are no monochromatic cycles, any obtained defeat graph is acyclic. Such defeat graphs enjoy beneficial properties, for instance they are guaranteed to have a single, nonempty preferred extension (Bench-Capon, 2002). This is not the case, however, if we allow agents to have preferences over values expressed as arbitrary preorders. Therefore, we will study the complexity of the preservability problem separately for the grounded, the preferred and the stable semantics.

Let us first consider the preservability problem for the grounded semantics.

Proposition 1. *The preservability problem is NP-complete for the grounded semantics.*

Proof. **Membership**: Take a $VAF = \langle A, \rightarrow, V, val\rangle$ and $a \in A$. Then guess a preorder P over V. We will check if for the defeat graph AF of $aVAF{=}\ \langle VAF, P\rangle$, $a \in GRND_{AF}$. Checking if the guess was correct is polynomial, as the credulous problem is polynomial for the grounded semantics, as indicated in Table 2.1.

Hardness: We will follow the construction used by Dunne and Bench-Capon (2004) to prove NP-hardness for Theorem 2 and show, that it also holds for the currently considered case. We will reduce the 3-SAT problem to preservability with grounded semantics. Consider any 3-SAT formula over a set of variables $Z = \{z_1, \ldots, z_n\}$: $\varphi = \bigwedge_{i=1}^{m}(x_i^1 \vee x_i^2 \vee x_i^3)$. Then, let us construct a $VAF{=}\ \langle A, \rightarrow, V, val\rangle$ with $a \in A$ such that a is preservable iff φ is satisfiable.

Let us take the set of arguments $A = \{\varphi, C_1, \ldots, C_m\} \cup \bigcup_{i=1}^{n}\{p_i, q_i, r_i, s_i\}$. Now consider attacks: for any clause $x_i^1 \vee x_i^2 \vee x_i^3$, if $x_i^g = z_k$, $p_k \rightarrow C_i$. Also, if $x_i^g = \neg z_k$, let $q_k \rightarrow C_i$. Further, for any $i \leq n$, let $p_i \rightarrow q_i$, $q_i \rightarrow r_i$, $r_i \rightarrow s_i$ and $s_i \rightarrow p_i$. Finally, for any C_i, let $C_i \rightarrow \varphi$. Now consider the assignment of values: assign the value *con* to any argument in $\{\varphi, C_1, \ldots, C_n\}$. Also, assign a value pro_i to any argument p_i, r_i and con_i to q_i, s_i.

Suppose now that there is some model M for which φ is satisfied. Then, for any variable z_k assigned $\top$, set $pro_k \succ con_k$. Symmetrically, if z_k is assigned $\bot$, set $con_k \succ pro_k$. Now notice that in the defeat graph induced by this preference ordering, each argument C_i is attacked either by p_i or q_i

which is itself not attacked. So, φ is in the grounded extension. So, it is preservable.

Further, suppose that φ is in the grounded extension under some assignment of values. This means that for any C_i there is some argument p_k or q_k attacking it, which is itself not attacked. Thi is the case because for any i, $val(C_i) = val(\varphi)$, so attacks of the form $C_i \rightarrow$ cannot be blocked. But this means that for any clause C_i in φ and some z_k in C_i we have assigned $pro_k \succ con_k$ if z_k is positive in C_i, or $con_k \succ pro_k$ if z_k is negative in C_i. But this gives us a valuation under which φ is satisfied.

So, the 3-SAT problem is reduced to the preservability problem with respect to the grounded semantics. □

Therefore, we can solve the preservability problem with respect to the grounded semantics while considering arbitrary preorders still in NP time.

Let us study the preservability problem with respect to stable semantics.

Proposition 2. *The preservability problem is NP-complete with respect to the the stable semantics.*

Proof. **Membership**: Take a $VAF = \langle A, \rightarrow, V, val\rangle$, $a \in A$. Guess a preorder P over V. Since credulous acceptance is in NP for the stable semantics, we can verify if a is accepted in the defeat graph based on P under these conditions in polynomial time. So, the considered problem is in NP.

Hardness: Take the credulous acceptance problem with respect to stable semantics, which we know is NP-complete, as indicated in Table 2.1. Then, consider an instance of this problem: an argumentation framework $AF = \langle A, \rightarrow\rangle$ and an argument $a \in A$. Then, we can assign each argument the same value v. In this way we construct a $VAF = \langle A, \rightarrow, \{v\}, val\rangle$, where for any $x \in A$, $val(x) = v$. Now, take the only one ordering P over this set of values. Clearly, if a is in some stable extension of the defeat graph of $aVAF = \langle VAF, P\rangle$, it is credulously accepted with respect to stable semantics in AF. Otherwise, it is not. □

Let us further determine the complexity of the discussed problem with respect to the preferred semantics.

Proposition 3. *The preservability problem is NP-complete with respect to the preferred semantics.*

Proof. **Membership**: Take a $VAF = \langle A, \rightarrow, V, val\rangle$, $a \in A$. Then, guess a preorder P over V. We need to check if a is credulously accepted w.r.t preferred semantics in the defeat graph based on P. Then notice that since

credulous acceptance is in NP for the preferred semantics, we can verify if a is accepted in the defeat graph in question.

Hardness: Take the credulous acceptance problem with respect to preferred semantics. Now consider an instance of this problem: an argumentation framework $AF = \langle A, \rightarrow \rangle$ and an argument $a \in A$. Then, assign all arguments the same value and check if a is preservable. If it is, then clearly a is credulously accepted, and otherwise it is not. □

3.3 Preservability with minimal changes

In the previous section we have shown complexity results for determining if some preference ordering preserving a decisive argument can be found. However, it is often not sufficient to find any explanation for a decision, we can be concerned with finding the optimal one with respect to some criterion. Here, we consider being *minimally distant* from an agent's sincere ordering over values as such criterion. As this avenue of research is novel, I will restrict my investigations to the grounded extension.

In the current context we are considering agents who already have their initial preference orderings over values. Then, even if they are inclined to push a certain decision forward, they are hesitant to change their view on the hierarchy of values too much. This can be because they do not want to violate their principles. But it can also be the case that a decision-maker is concerned with her credibility. Then, she wants to avoid situations in which recipients of a justification for a decision are not willing to believe in its sincerity because it is substantially different from what they think is the sincere hierarchy of values that the decision maker has. Following this intuition we will study the problem of determining the preference ordering over values ensuring that a decision is made, which is minimally different from the agents' sincere hierarchy.

Let us begin with checking the complexity of finding if there is some audience enjoying the desired property within a distance of k from the original ordering.

k-DISTANCE PRESERVABILITY(σ, d)

Instance: $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$, semantics σ, distance $k \in \mathbb{N}$, preorder P over V.
Question: Is there an audience P' such that a is credulously σ accepted in the defeat graph of VAF based on P' and $d(P, P') \leq k$?

Proposition 4. *The k-distance preservability problem is NP-complete for the Hamming distance and the grounded semantics.*

Proof. Take a $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$ and a distance k. For membership in NP, consider a procedure in which a preorder P^* based on V is guessed. Then it is polynomial to check if a is accepted under grounded semantics in the defeat graph of VAF based on P. Also, it is easy to see that it is polynomial to check if $HD(P, P^*) \leq k$ - it is sufficient to check for all possible pairs of values if they belong to one preorder but not to the other.

For NP-hardness consider a reduction of the preservability problem which we have shown to be NP-complete for grounded semantics. Take any $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$. Also, take the empty preorder P. Then consider $k = |V|^2$. Now check if there is a preorder preserving a within a distance k from P. Note that k is the maximal distance from P, so if a desired preorder P^* exists, $HD(P, P^*) \leq k$. But checking that is NP-hard, so k-distance preservability problem also is. □

This result is easily generalizable to any distance metric over preorders which is verifiable in polynomial time and for which the maximal distance between any pair of preorders within some set is polynomialy computable.

Proposition 5. *k-distance preservability problem is NP-complete for the grounded semantics and any distance metric over preorders d for which it is polynomial to check if $d(P_1, P_2) \leq k$, where P_1, P_2 are arbitrary preorders over some set V and $k \in \mathbb{N}$, and for fixed V, $d(P_1, P_2)$ is bounded by some constant M.*

Proof. It is sufficient to consider a construction symmetric to the one used in the Proposition 4. □

Let us then proceed to checking what is the complexity of finding what is the closest audience preserving an argument.

MIN-DISTANCE PRESERVABILITY(σ, d)

Instance: $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$, preorder P over V, preorder P' over V.
Question: Is P such that a is credulously σ accepted in the defeat graph of AF based on it and for any P', $d(P, P) \leq d(P, P')$?

By binary search we can show that for the grounded semantics this problem is in Θ_2^P (the definition of this class is given in the section 2.3).

Proposition 6. MIN-DISTANCE PRESERVABILITY *is in* Θ_2^P *for the grounded semantics and distance metrics satisfying conditions of Proposition 5.*

Proof. We will show that the problem is Θ_2^P by constructing a binary search algorithm. Take a $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$, a polynomially computable distance metric d with the given maximal distance between preorders over V, namely M. Also, consider some initial preorder over V, namely P. Now, first check if there is some preorder preserving a within distance M from P. We know that this step is computable in NP time. If there is one, check if there is one within distance $\frac{M}{2}$. If there is not, check within distance $M * \frac{3}{4}$. Perform this procedure until we find a preorder within the distance ϵ from P preserving a such that for any distance smaller than ϵ no preorder satisfying this property exists. As we know that the binary search is logarithmic from the size of input, we only need to solve the NP-complete problem of finding a preorder preserving a within some distance d. So, our problem is in Θ_2^p. □

3.4 Graph characterisation results

In this section it is of interest to find restrictions on the structure of *VAF*s which ensure that manipulation in the described sense is impossible. Another goal is to find structural restrictions of *VAF*s in which the preservability problem is polynomial. Although we know that this problem is difficult in general case, we might possibly find kinds of *VAF*s in which it is simplified. Those should be avoided when they are used for decision-making in the investigated manner.

We will begin with finding a restriction on the structure of *VAF*s ensuring that manipulation is not possible. To achieve this goal, let us introduce a notion of *guarded arguments*. Intuitively, these are arguments attacked by a chain of arguments preventing them from being accepted. This chain is required to be labeled with the same value as the argument in question, which

ensures that none of the attacks is possibly blocked under some audience. This is ensured by setting that the chain (including the decisive argument) is even and that even elements of the chain are only attacked by their direct predecessors.

Definition 15 (Guarded arguments)**.** *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a VAF,* $a \in A$*. We say that* a *is guarded iff there is an uneven-length sequence of distinct arguments (in terms of the number of arguments)* $a, b_1, \ldots, b_n$ *(a guard) such that* $n \geq 2$*,* $b_1 \rightarrow a$ *and for any* $i < n, b_i + 1 \rightarrow b_i$ *and for any* $i \leq n$*,* $val(b_i) = val(a)$*. Also, for any even* $i \leq n$*,* b_i *is attacked only by* b_{i+1} *or by no argument.*

To illustrate this notion let us consider a slightly modified version of the debate described before. It is shown in the Figure 3.4. In this *VAF* the argument A is guarded. A is attacked by A' which is attacked by A". A" is not attacked by any argument. So, we have an uneven, monochromatic chain such that for any uneven predecessor of A, it is not attacked by any argument than its own predecessor.

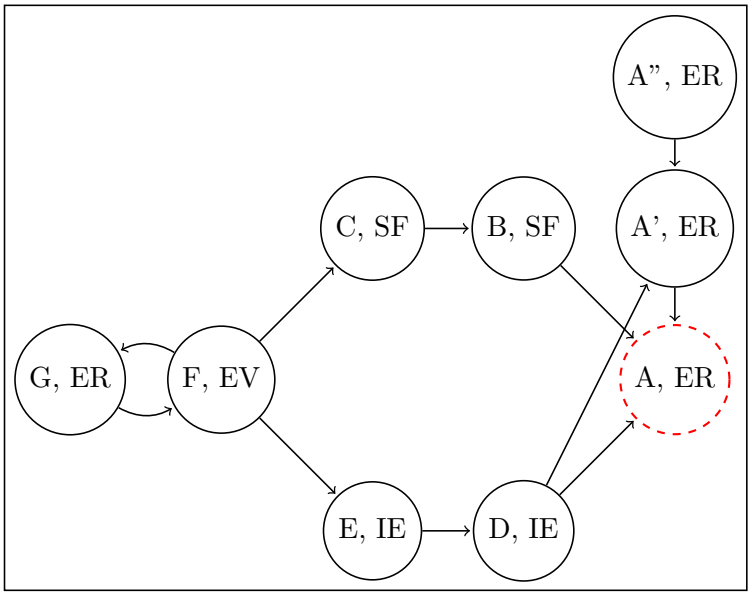


Figure 3.4: Argumentation structure with A being a guarded argument.

Guarded arguments are not preservable under any semantics which assume admissibility.

Proposition 7. *For any $VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$, if a is guarded, it is not credulously preservable under the admissible semantics.*

Proof. Let $VAF = \langle A, \rightarrow, V, val \rangle$ be a VAF, $a \in A$ such that a is guarded. Suppose that there is an audience P such that a is in some admissible set S of $AF = \langle A, \rightarrow^P \rangle$. We know, that by properties of the defeat relation, the attack sequence given by the guard of a is preserved. Now notice that as $b_1 \rightarrow a$, and b_2 is its only attacker, $b_2 \in S$. Consequently, for any even i, $b_i \in S$. Now consider the last uneven argument in the guarding chain. We know that it has an attacker, as the sequence is even, and that the attackers is not attacked. So, S is not admissible. □

Having established this restriction, which ensures that the setting is immune to manipulation, we might attempt to find cases in which it is manipulable. In fact, we can try to find cases in which it is easy. However, we know that the preservability problem is NP-hard even if we restrict the problem to binary trees. Therefore, relevant restrictions are deemed to be strict.

One of the simple restrictions in which finding a way to manipulate is easy, is when the attack relation is a chain. Let us introduce this notion formally.

Definition 16 (Chain of arguments)**.** *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework. A chain of arguments is a sequence of arguments $\langle a_1, \ldots, a_n \rangle$ such that for any $i < n$, $a_i \rightarrow a_{i+1}$. We say that a chain of arguments has a length n if it consists of n arguments.*

Proposition 8. *Let $VAF = \langle A, \rightarrow, V, val \rangle$ be a VAF and $a \in A$. If $\rightarrow$ is a chain, the preservability problem is polynomial for any semantics assuming admissibility.*

Proof. Consider the following procedure, assuming that the input attack relation is a chain:

Algorithm 1 CheckPreservable

procedure CheckPreservable($VAF = \langle A, \rightarrow, V, val \rangle$, $a \in A$)
  $Au \leftarrow$ empty preorder over V
  $NotBreak \leftarrow$ Empty list of pairs of values
  $En \leftarrow$ Enumeration of arguments in the chain, a is En_0
  if *length(En) is even* **then return** True
  for $En_i \in En$, where $i > 0$ **do**
    if i is even **then**
      append $\langle val(En_{i-1}), val(En_i) \rangle$ *to NotBreak* .
    if i is uneven&$val(En_{i-1}) \neq val(En_i)$&$\langle val(En_{i-1}), val(En_i) \rangle \notin$ *NotBreak* **then**
      Add $val(En_{i-1}) \succ val(En_i)$ to Au
      Return True
  Return False

Correctness: Suppose that the procedure returns True. Then, in the defeat graph based on Au the chain is broken at an even distance from a. Then it is easy to check that a is in some admissible extension. Now suppose that the procedure returns False. Then, the length of the chain is uneven. Let us show by induction on the legth of enumeration that then it is impossible to find an audience preserving a in some admissible extension. If $n = 1$, then the only argument attacking a is En_1. Then $val(En_1) = val(a)$, as otherwise the attack would be broken by construction of the procedure. Suppose that the claim holds for $m = n$. Now consider an enumeration of length $m = n + 1$. Then notice that the procedure would provide the m^{th} element of chain would be disregarded must be False, as otherwise, we would not reach the m^{th} step. If m is not even, as its attack is not blocked, it is either the case that its value is the same as of En_{m-1}, so it is impossible to block it for any audience, or the attack is in the $NotBreak$ list. Then, if we would break it, the chain would be broken at an even level, which would make a unacceptable. Also, if m is even, then by construction the procedure returns True.

Complexity: Clearly, the procedure is polynomial from the size of $\rightarrow$, as it only involves checking the chain once. □

This result can be easily generalized to frameworks in which a relevant chain is embedded in a bigger structure.

Proposition 9. *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a VAF and* $a \in A$. *If there is a chain of arguments* $Ch = a \leftarrow b_1, \leftarrow \ldots, \leftarrow b_n$, *such that for any element*

of the chain its only attacker is its successor in the chain, the preservability problem is polynomial.

Proof. Let $VAF = \langle A, \rightarrow, V, val \rangle$ be a VAF and $a \in A$. If there is a chain of arguments $Ch = a \leftarrow b_1, \leftarrow \ldots, \leftarrow b_n$, such that for any element of the chain its only attacker is its successor in the chain. Consider $VAF' = \langle A', \rightarrow' , V, val \rangle$, where $\langle A', \rightarrow' \rangle = Ch$. Then, it is sufficient to take the procedure used in the proof of Proposition 8 to prove the claim.

□

So, VAFs which satisfy the conditions of the last two propositions should be avoided, if there are reasons to believe that the decision-maker has an incentive to manipulate.

The next chapter deals with situations in which several agents interact on a given decision problem.

Chapter 4

Multi-agent setting

4.1 Introduction

In the recent literature on abstract argumentation the problem of establishing a collective view on particular aspects of deliberation has gained a growing interest (Bodanza et al., 2017). This problem becomes particularly interesting when argumentation is used as an assistance in collective decision-making. When a group wishes to reach a collective view with respect to some decision based on an argumentation, it is crucial that they share a view about its structure. Therefore, it is important to provide fair methods of reaching a collective view about it.

One of the intuitions behind the claim that argumentations can be viewed in distinctive ways is that agents might not agree on whether particular arguments are indeed in conflict with each other. Possible explanations for such a situation involve a scenario in which particular agents disagree on the relative strength of arguments. As it was argued earlier, it is plausible to assume that if an agent believes that some strong argument is attacked by a weak one, she might decide to disregard this attack. However, decisions about which arguments are stronger than another are at the discretion of individual assessors. Therefore, structures of *successful* attacks between arguments might vary among the considered group of agents.

This explanation undoubtedly applies to discussions based on values aimed at deciding upon a specific action. To illustrate this point let us continue the example of a debate discussed in the previous chapter.

Example 4.1.1. *We are faced with a debate concerning choosing a decision. In this debate arguments are associated with particular values. The VAF depicting the relevant argumentation is shown in the Figure 3.1.*

Recall that in the previous chapter we were considering justifications made by a single agent. Then, the mayor was responsible for passing the law and she was only looking for a way to explain her action. Let us now suppose that the decision is made by a committee of three members, who need to reach some collective view.

Then, there are many possible ways of determining the relative strength of arguments based on a ranking over values that they appeal to. Let us consider three committee members involved in the discussion, representing three different points of view on this subject. Each of them can be associated with a distinctive preference ordering over values and a distinctive defeat graph of the initial VAF. They are depicted in the Figure 4.1. One of them, in the graph (b), belongs to the mayor who tries to convince the others that passing the law is a good decision. The agent whose point of view is presented in the graph (c) supports her stance. Graph (a) shows the view of an agent who is against the bill[1]*. For the simplicity of the setting, unlike in the previous chapter we assume that agents' preferences over outcomes of discussion are induced by their sincere preferences over values. Then, it is natural to study behavior of both agents who wish the decisive argument to be accepted, and those who would prefer it to be rejected.*

[1] In the graph (a) the grounded extension is $\{F, B, D\}$, in graphs (b) and (c) it is $\{G, C, E, A\}$.

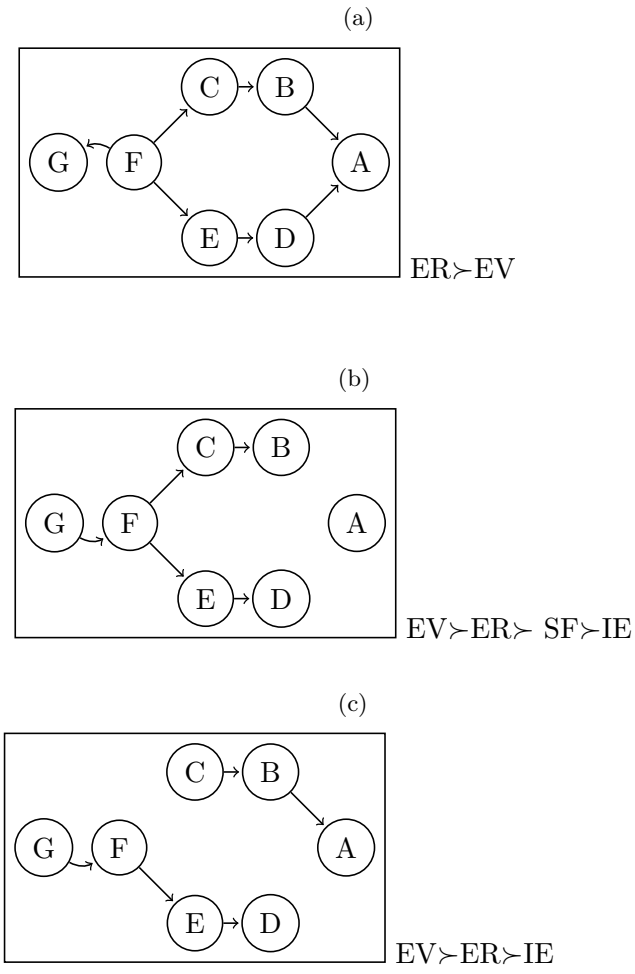


Figure 4.1: Defeat graphs of the example *VAF* for 3 agents with distinctive preferences over values.

When we are faced with a situation like this, we would like to find a method of providing a single defeat graph for the entire group of agents. Such a structure should take into account all perspectives of agents involved in the decision process. In this way we can provide a collective decision representing agents' views on arguments fairly, while making sure that the result of the decision process can be justified by some hierarchy of values.

Two natural approaches to this problem can be distinguished.

1. In the first of them, we focus on providing a collective preference order-

ing over values. Having established it, it is straightforward to compute a collective defeat graph based on it.

2. We can, however, take a different approach. We can consider providing a collective defeat graph basing just on the defeat graphs submitted by agents. In this way we can reach a compromise with respect to agents' views on the relative strength of arguments even without access to their precise preferences over values.

At the first glance it is difficult to determine which of the two ways of aggregating value-based argumentation frameworks is more appropriate. A goal of this work is to study them in detail to compare the benefits of their usage.

In Sections 4.2 and 4.3, a formal account for these approaches will be provided. I will present intuitions behind them and showcase several insightful properties of the proposed frameworks. Further, in Section 4.4, I will describe relations between the approaches. In particular, I will show a method of defining them in terms of each other. Finally, in Section 4.5, I will study the possibility of manipulating the outcome of aggregation within the proposed frameworks. I will focus on two aspects of strategic behavior in the considered settings. Firstly, it will be determined in which cases manipulation is impossible. Further, we will attempt to find methods of ensuring that strategic behavior is computationally difficult where it is possible.

4.2 Aggregating argumentation graphs

Let us firstly consider the approach in which the choice of a collective defeat graph is based on individual defeat graphs submitted by particular agents. It is worth noting that this approach has a significant advantage over aggregation of preference orderings. Namely, we do not need to have access to agents' preference orderings. What is sufficient here is to know what is the relative strength of arguments from the agents' perspectives, not the precise orderings which determined it.

In this thesis I focus on the aggregation of argumentation frameworks employing graph aggregation methods, originating in the social choice theory. It is worth noting that there exist other approaches to this problem, attempting to capture the specificity of multi-agent argumentation in distinctive manners. A comprehensive overview of such methods was provided by Bodanza et al. (2017).

4.2.1 The framework

The formalization of the approach in question is based on the model developed by Chen and Endriss (2017), which follows the results of Endriss and Grandi (2017). In this model the social choice mechanism is used in order to aggregate a number of submitted argumentation frameworks. It is further assumed that all agents agree upon the set of available arguments. What they are unsure about is the structure of an appropriate attack relation.

Let us demonstrate how such an aggregation rule might behave on an example.

Example 4.2.1. *Let us consider the three agents with distinctive views on the structure of attacks between a common set of arguments, giving rise to a profile of defeat graphs* ***AF*** *(see Figure 4.1).*

Then, we can consider a majority aggregation rule, in which an attack is accepted collectively if and only if more than a half of agents accepts it. Let us compute the result of application of this function to the depicted profile (see Figure 4.2.)

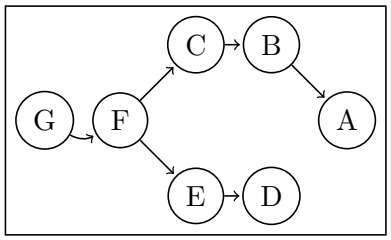


Figure 4.2: Application of the majority rule to the profile **AF**.

Let us now provide a way to aggregate defeat graphs. It is worth noting that the preferences over values that agents believe in are expressed in the structure of attacks that they believe are successful. The defeat graphs that agents submit primarily show the relative strength of arguments induced by preferences over values that they find appropriate. Therefore, given a *VAF* that serves as the basis for developing agents' individual defeat graphs we can provide a common argumentation graph expressing a collective view on the strength of arguments.

Let us first define a defeat aggregation problem. It consists of a *VAF* and a profile of defeat graphs corresponding to particular agents' views on the relative strength of arguments.

Definition 17 (Defeat Aggregation Problem)**.** *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be*

a VAF and $\mathcal{N} = \{1, \ldots, n\}$ *be a set of agents. We call a profile of defeat graphs of VAF* $\boldsymbol{AF} = \langle\langle A, \rightarrow^1\rangle, \ldots, \langle A, \rightarrow^n\rangle\rangle$ *a defeat aggregation problem.*

Then, we can take a function taking as an input a profile of defeat graphs of some *VAF* and providing a single argumentation framework.

Definition 18 (Defeat Aggregation Rule)**.** *Let* $VAF = \langle A, \rightarrow, V, val\rangle$ *be a VAF. Then, a defeat aggregation rule is a function* $F^{VAF} : \mathcal{AF}^n \rightarrow \mathcal{AF}$*, where* $\mathcal{AF}^n$ *is the set of all profiles of defeat graphs of VAF of length n and* $\mathcal{AF}$ *is the set of all argumentation frameworks with the set of arguments equal to A. Further, for any profile of defeat graphs* $\boldsymbol{AF}$ *and an attack* $a \rightarrow b$*,* $N^{a\rightarrow b}_{\boldsymbol{AF}}$ *denotes the largest set of agents such that for any* $i \in N^{a\rightarrow b}_{\boldsymbol{AF}}$*,* $a \rightarrow^i b$*. By a defeat aggregation rule F we denote a collection of defeat aggregation rules* F^{VAF}*, defined for all VAFs.*

4.2.2 Preservation of being an audience

In the intended application agents are determining the outcome of their discussion based on their beliefs about the relative strength of arguments, induced by a preference ordering over values that they think should hold. In such an approach it is crucial that they are capable of justifying their position by providing an appropriate ordering of values explaining their choices. Naturally, it is of interest for the collective view defeat graph to be justifiable as well. While it can be demanded from agents to submit explainable defeat structures, it is the matter of a proper design of an aggregation mechanism to ensure that the collective structure is explainable. This problem is highly related to the work by Airiau et al. (2016), who studied when it is possible to find a *VAF* for a set of argumentation frameworks such that all of them are its defeat graphs.

The problem of preservation of being an audience is highly related to the general issue of preserving properties of argumentation frameworks in their aggregation, studied by Chen and Endriss (2017). In abstract argumentation several structural properties of argumentation frameworks, such as being acyclic, are important to ensure that an outcome of a discussion is concluded efficiently and unambiguously. Therefore, it is important to ensure that if all agents submit a graph satisfying such a property, likewise does the collective framework.

We will study when can it be the case that the property of being a defeat graph of a certain *VAF* is preserved. In other words, we will investigate when it is the case that the collective framework is a defeat graph of the original *VAF*. To achieve this goal, let us state formally when is it the case that a

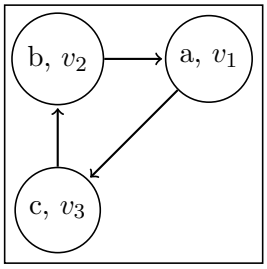


Figure 4.3: Triangle *VAF*.

graph is a defeat graph of *VAF*. We say that it holds when we can find a preference ordering over values resulting in it.

Definition 19. *Let* $VAF = \langle A, \rightarrow, V, val \rangle$ *be a VAF, and* $AF = \langle A, \rightarrow \rangle$. *We say that AF is a defeat graph of VAF iff there exists a preorder P over V such that AF is the defeat graph of a* $VAF = \langle VAF, P \rangle$. *We call such a P a* justification *of AF.*

Let us illustrate the problem of determining if an argumentation graph is a defeat graph of a given *VAF*. Consider the *VAF* shown in the Figure 4.3 and argumentation frameworks in the Figure 4.4.

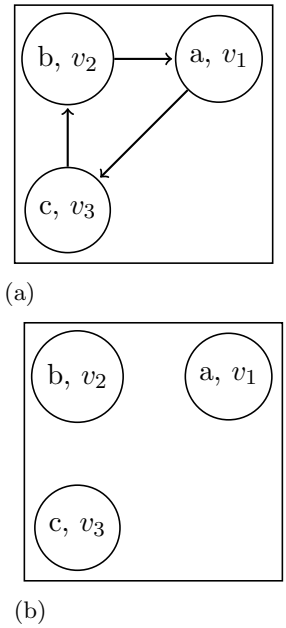


Figure 4.4: Potential defeat graphs.

The graph (a) is clearly a defeat graph of *VAF*. It is sufficient to consider the empty preorder over values. However, it is impossible to find a preorder to justify the framework (b). Observe that any preorder over values under which all three attacks would be eliminated would need to ensure that (1) $v_3 \succ v_1$, (2) $v_1 \succ v_2$, but also that (3) $v_2 \succ v_3$. So, such an ordering would violate the transitivity condition and thus would not be a preorder.

Along the lines of this observation, I will show that the property of being a defeat graph is not preserved by quota rules.

A quota rule is a defeat aggregation rule such that any attack is preserved if and only if it is supported by a number of voters larger than some fraction q of the total number of voters.

Definition 20 (Quota rule)**.** *Let $\boldsymbol{AF}$ be a defeat aggregation problem. 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 $N_{\boldsymbol{AF}}^{a \to b} \geq \lfloor q * N \rfloor$, where N is the total number of voters.*

A natural example of a quota rule is the weak majority rule, in which $q = \frac{1}{2}$.

We show that no quota rule can preserve being a defeat graph.

Proposition 10. *Being an audience is not preservable by any quota rule.*

Proof. Consider a quota defeat aggregation rule F with an arbitrary quota $q \in [0,1]$. Then, take some $n \in \mathbb{N}$ such that $\frac{1}{n} < q$. Further, construct a $VAF = \langle A, \rightarrow, V, val \rangle$ such that $A = \{a_1, \ldots, a_n\}$, $\rightarrow = \{a_i \rightarrow a_{i+1} | i < n\} \cup \{a_n \rightarrow a_1\}$. Notice 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.

Notice 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 $v_n \succ_P V_{n-1} \succ_P \cdots \succ_P v_1 \succ_P v_n$. But then P is not transitive, so it is not a preorder.

□

This is not a good news in the sense that we cannot use a large class of intuitive aggregation rules, if we wish to make sure that any collective defeat graph is justifiable in terms of preferences over values.

A class of rules which is capable of overcoming this issue is the class of *distance based* rules. Such functions, given a profile of defeat graphs of some VAF output a defeat graph which minimizes average distance between input graphs and the output. An example of such a rule is the rule selecting a defeat graph of VAF, minimizing the average Hamming distance between the input graphs. Here, the Hamming distance between two argumentation frameworks AF_1, AF_2 is the number of attacks $a \rightarrow b$ such that $a \rightarrow b \in AF_1$ and $a \rightarrow b \notin AF_2$, or $a \rightarrow b \in AF_2$ and $a \rightarrow b \notin AF_1$.

Nevertheless, the inability of a large class of defeat aggregation rules to preserve being an audience can be considered as an advantage of aggregating preference orderings directly. This approach, as we will see, gives an assurance that an obtained argumentation graph is a defeat graph of an initial VAF.

4.3 Aggregating orderings of values by preference aggregation

The other method of aggregating views on the relative strength of arguments is to reach a collective view about the ordering over values that they appeal to. This approach makes use of preference aggregation functions, pioneered by Arrow (1951), which constitute one of major parts of the social choice theory.

In this way we might establish a collective ordering over values and consequently compute a collective defeat graph of the initial *VAF*. Then, evaluation of acceptance of a decisive argument can be performed. This approach has been proposed earlier by Pu, Luo, Zhang, and Luo (2013). They suggested the application of preference aggregation techniques to aggregating views on preferences over values, however in their work connections between this approach and aggregating defeat graphs was not specified. Also, we generalize their work by allowing agents to express their preferences as arbitrary preorders.

In order to provide the described procedure formally, preference aggregation functions will be used. This mechanism, widely studied in social choice theory, considers a profile of orderings over a set of items. Further, it provides a single, collective ordering.

Definition 21 (Preference Aggregation Function)**.** *Let $V = \{v_1, \ldots, v_n\}$ be a set of options, $\mathcal{N} = \{1, \ldots, m\}$ be a set of agents, and $\mathcal{P}$ be the set of all preorders over V. Then, a preference aggregation function is a function $F : \mathcal{P}^m \to \mathcal{P}$. We denote the set of agents supporting $v_i \succeq v_j$ in a profile $\boldsymbol{P}$ as $N_{\boldsymbol{P}}^{v_i \succeq v_j}$.*

It is worth noting that in this mechanism it is required for the obtained preference ordering to be a preorder over V. Therefore, we are guaranteed that a collective argumentation framework based on it is a defeat graph of the initial *VAF*. This observation creates a strong restriction on the set of eligible rules in this context.

There are multiple examples of such rules. Many of them were originally defined under the restriction that all preference orderings in the input are linear. To introduce them, let us provide some handy notation.

Notation 2. *Let P be a linear order over some set V. We denote by $top(P)$ the option $v \in V$ such that for any $v' \neq v$, $v \succ_P v'$. Further, we denote as $rank_P(v)$ the position of the option v in the ordering P. Formally, $rank_P(v) = |\{v' \in V | v' \succ_P v\}| + 1$.*

An example of such a rule is the Borda rule. There, for any element P_i of a profile of linear preference orderings $\mathbf{P}$ of length n over a set of options V we assign to each option a number of points. A score an option v_j given by an agent i, called $BordaScore_i(v_j)$ amounts to $|V| - rank_{P_i}$. Then, an overall score of v_j, namely $BordaScore(v_j) = \Sigma_{i=1}^{n} BordaScore_i(v_j)$. Finally $Borda(\mathbf{P})$ is a preference ordering in which the rank of each option is determined by the number of gained points. To obtain a linear ordering as the output of this function additional tie-breaking rules are needed.

Another interesting rule of this kind is STV. Let $\mathbf{P}$ be a profile of linear preference orderings over a set of options V. It is computed in a number of steps. In each of them, an option which is ranked as $top(P_i)$ for the fewest members of $\mathbf{P}$ is eliminated. Then, $STV(\mathbf{P})$ is a preference ordering in which the rank of each option is $m - i + 1$, where m is the number of options and i is the number of round in which the option was eliminated.

When aggregation of preorders is considered, we can consider rules minimizing average *distance* between the collective preorder and preorders in the input. An example of such a rule would be a rule minimizing the average Hamming distance.

Then, it is straightforward to apply such a mechanism to aggregation of value-based argumentation frameworks. Given a *VAF*, it is sufficient to take a profile of preference orderings over the set of values and use a preference aggregation function to obtain a collective ordering over them. Further, we can compute a collective defeat graph of the *VAF*, under which the collective selection of arguments can be performed.

Let us illustrate this mechanism with employment of the running example.

Example 4.3.1. *Consider again the debate concerning banning Diesel cars depicted before. Recall that the debate we considered was captured as the VAF of the Figure 3.1.*

Let us consider again three committee members deciding upon the ban in question. Again, they represent different views on the hierarchy of values. This time, they are looking for a collective view with respect to the preferences over relevant values. The first of them believes that

$$ER \succ EV \succ IE \succ SF$$

Second thinks that

$$ER \succ IE \succ SF \succ EV$$

Finally, the third submits that

$$EV \succ ER \succ IE \succ SF$$

Let us use the Borda rule to compute the collective preference ordering over values. The score of ER is 8. Then, the score of EV is 5, while the score of IE is 4. Finally, the score of SF is 1. This gives us a collective ordering: $ER \succ EV \succ IE \succ SF$.

Given the obtained preferences over values, we can retrieve a defeat graph for the group. It is shown in the Figure 4.5.

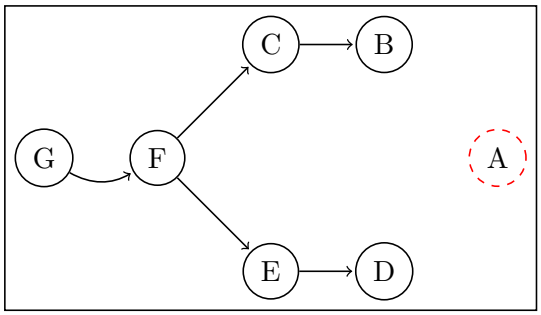


Figure 4.5: Application of the Borda rule for the example profile.

Then, the decisive argument A is accepted in the grounded extension of the collective defeat graph, which amounts to $\{A, G, C, E\}$. *So, we have a good reason to ban diesel cars.*

This approach, as we can see, gives us a handy way of computing compromise preference orderings over values which provide us with collective defeat graphs. However, it requires agents to submit their actual preference orderings, which is not the case in the defeat aggregation.

In the next section I will attempt to provide a method of combining defeat aggregation with preference aggregation which aims at overcoming the disadvantages of the described approaches.

4.4 Connections between aggregation approaches

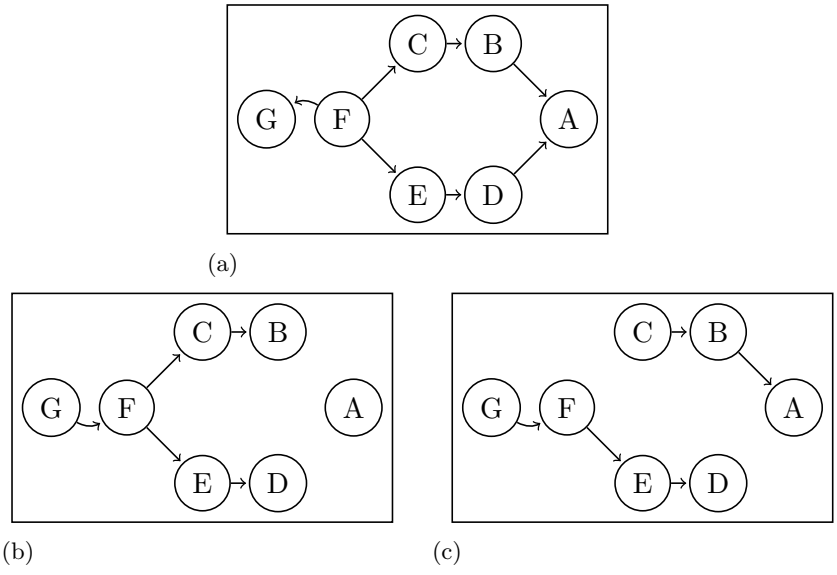
It is worth noting that the discussed approaches to aggregating *VAF*s are similar to a large extent. The goal of this section is to explore connections between them.

Let us first notice that the considered approaches can be simulated by each other. We can define a defeat aggregation rule in terms of a preference aggregation rule. Also, under certain conditions, we can find preference aggregation rule corresponding to defeat aggregation rules.

Let us begin with defining a preference aggregation rule corresponding to some defeat aggregation rule. Intuitively, given a *VAF* and a profile of its defeat graphs, we can consider a profile of preferences over values which induce the initial defeat graphs.

Let us illustrate this point on the example used earlier. Consider the previously described *VAF*. It is depicted in the Figure 3.1.

Also a profile of its defeat graphs. Let us use the defeat graphs depicted in the Figure 4.2. For the sake of clarity, let us repeat the graphs.



Further, take some graph aggregation function and apply it to the profile of defeat graphs. Let us consider the rule which provides the defeat graph of the *VAF* minimizing the average Hamming distance between input defeat graphs and the collective graph. Clearly, attacks $C \to B$ and $E \to D$ need to be in such a defeat graph. Further, attacks $G \to F$ and $F \to E$ should be included, while other should not. The result of the computation is identical with the one presented in the Figure 4.5.

This structure is indeed a defeat graph of *VAF*. To see that, consider a preference ordering (coll) = ER$\succ$EV$\succ$IE.

Then, we might also consider a procedure which instead of aggregating

defeat graphs, takes into account preference orderings which could induce defeat graphs considered before. For instance, in our case we could consider (a) EV$\succ$ER, (b)ER$\succ$EV$\succ$IE$\succ$SF, (c) SF$\succ$ER$\succ$EV$\succ$IE. Also, we can find a preference ordering which would induce the result of the graph aggregation function, such as (coll).

4.4.1 Defeat aggregation in terms of preference aggregation

Let us proceed to the formalization of the provided intuition. Firstly, consider a justification of a single defeat graph of some *VAF*, as defined in Definition 19.

Then, it is straightforward to define the justification of a profile of defeat graphs of a known *VAF*. By this term I mean a sequence of preference orderings over values such that any i-th member of this sequence justifies the i-th defeat graph.

Definition 22 (Justification of a profile of defeat graphs)**.** *Let $VAF = \langle A, \rightarrow V, val\rangle$ be a VAF and $\boldsymbol{AF}$ be a profile of its defeat graphs. We say that a profile of preference orderings $\boldsymbol{P}$ over V justifies $\boldsymbol{AF}$ iff*

1. *$\boldsymbol{AF}$ and $\boldsymbol{P}$ have the same length*
2. *For any $i \leq n$, where n is the length of $\boldsymbol{AF}$, the i-th element of $\boldsymbol{P}$ is a justification of the i-th element of $\boldsymbol{AF}$ with respect to VAF.*

It is worth noting that it is often the case that multiple preference orderings justifying a defeat graph exist. We will be interested in studying classes of orderings inducing identical defeat graphs with respect to particular *VAF*s.

Let us then introduce a notion of similar preference orderings with respect to a defeat graph of some *VAF*. Here, we mean that two preference orderings over values induce the same defeat graph.

Definition 23 (Similar preference orderings)**.** *Let $VAF = \langle A, \rightarrow, V, val\rangle$ be a VAF, $AF = \langle A, \rightarrow_D\rangle$ be one of its defeat graph and P, P' be preference orderings over V. We say that P and P' are similar with respect to AF iff both P and P' are justifications of AF with respect to VAF.*

With the defined notions in hand we can define a simulation of a defeat aggregation rule in terms of preference aggregation. Intuitively, for any profile of defeat graphs and a collective defeat graph given by some aggregation function, we can extract a profile of preference orderings justifying the

input of a defeat aggregation rule and a collective preference ordering justifying the aggregated defeat graph. If we consider such an operation for the entire defeat aggregation function, we can obtain a preference aggregation function.

Definition 24 (Simulation of a defeat aggregation rule)**.** *Let* $VAF = \langle A, \rightarrow, V, val\rangle$ *be a VAF and* F^{VAF} *be a defeat aggregation rule. Then,* $F^{Pref} : \mathcal{P}^m \rightarrow \mathcal{P}$ *is a preference aggregation simulating* F^{VAF} *iff for any profile of defeat graphs* $\boldsymbol{AF}$ *of VAF, a graph* $F^{VAF}(\boldsymbol{AF})$ *and a profile of preference orderings* $\boldsymbol{P'}$*, if* $\boldsymbol{P'}$ *justifies* $\boldsymbol{AF}$*, then* $F^{Pref}(\boldsymbol{P'})$ *justifies* $F^{VAF}(\boldsymbol{P})$*.*

It is worth noting that because some defeat graphs can be justified in multiple ways, simulation of a defeat aggregation rule is not unique. Then, we consider a class of preference aggregation rules which are simulating some defeat aggregation rule. On the other hand, justifications of defeat aggregation rules do not always exist. As we have seen, not all such rules preserve being an audience. Therefore, their simulation in some cases would not provide a justification of the output of the rule.

The notion of simulations of defeat aggregation rules is useful for a number of reasons. The first of them is that it enables us to obtain a clear condition for preservation of being a defeat graph by an argumentation aggregation rule. Intuitively, a defeat aggregation rule preserving being a defeat graph should always allow for finding a justification for its defeat graph.

Proposition 11. *A defeat aggregation rule preserves the property of being an audience iff it can be simulated by some preference aggregation rule.*

Proof. ($\Rightarrow$) Consider some defeat aggregation rule F which preserves being an audience. Also, suppose that it cannot be simulated by any preference aggregation rule. Then, there is a profile $\mathbf{AF}$ of defeat graphs of some *VAF* such that $F(\mathbf{AF})$ cannot by justified by any preference ordering. But then clearly it is not a defeat graph of *VAF*, so F does not preserve being an audience.

($\Leftarrow$) Assume that a defeat aggregation rule F is simulated by some preference aggregation rule F^*. Then suppose that it does not preserve being an audience. Then we know that there is a profile $\mathbf{AF}$ of defeat graphs of some *VAF* such that $F^{VAF}(\mathbf{AF})$ is not justified by any preference ordering over values. But then clearly F^* does not simulate F. □

Another beneficial factor of introducing justifications of profiles of defeat graphs is that they might allow for providing a justification of an outcome of

a discussion in terms of preferences over values without assuming knowledge of agents' actual hierarchies of values. If we have an access to the agents' views on the relative strength of arguments, expressed as defeat graphs, we can retrieve their justifications. Then, we can use a preference aggregation function to get a collective defeat graph.

It is worth noting, however, that because we can get multiple justifications of profiles of defeat graphs, it is possible that we would receive different outcomes of a preference aggregation function depending on a choice of justification. We can show, however, that for any independent preference aggregation rule, and any profile of defeat graphs the collective graph induced by aggregation of preference orderings justifying the profile of defeat graphs does not depend on the choice of orderings justifying them. This is true under assumption that only connected preorders [2] are used as justifications of the input and the employed preference aggregation function preserves being a connected preorder.

We will establish the mentioned results with respect to *independent* preference aggregation functions. Independence ensures that an attack should be treated equally in all profiles.

Definition 25 (Independence)**.** *F is independent if it holds that for any pair of profiles of preorders $\boldsymbol{P}$, $\boldsymbol{P'}$ and any pair of values $v_1, v_2 \in V$, if $N_{\boldsymbol{P}}^{v_1 \succeq v_2} = N_{\boldsymbol{P'}}^{v_1 \to v_2}$, then $v_1 \succeq v_2 \in F(\boldsymbol{P})$ iff $v_1 \succ v_2 \in F(\boldsymbol{P'})$.*

Proposition 12. *Let $VAF = \langle A, \to, V, val\rangle$ be a VAF and F be any independent preference aggregation rule and $\boldsymbol{AF}$ be a profile of its defeat graphs. Then, for any two profiles $\boldsymbol{P}, \boldsymbol{P'}$ of connected preorders over V justifying $\boldsymbol{AF}$, $\langle A, \to^{F(\boldsymbol{P})}\rangle = \langle A, \to^{F(\boldsymbol{P'})}\rangle$.*

Proof. Consider any $VAF = \langle A, \to, V, val\rangle$, as well as a profile of its defeat graphs $\mathbf{AF}$. Also, let F be any independent preference aggregation rule preserving being a connected preorder. Also, take any pair of profiles $\mathbf{P}, \mathbf{P'}$ of connected preorders over V similar with respect to $\mathbf{AF}$. Now suppose that $\langle A, \to^{F(\mathbf{P})}\rangle \neq \langle A, \to^{F(\mathbf{P'})}\rangle$. Without loss of generality assume that there is an attack $a \to b$ such that $a \to b \in \langle A, \to^{F(\mathbf{P})}\rangle$ but $a \to b \notin \langle A, \to^{F(\mathbf{P'})}\rangle$. Then, by connectedness we know that $val(a) \succeq val(b) \in F(\mathbf{P})$. 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 $N_{\mathbf{P}}^{val(a) \succ val(b)}$. Notice that they must correspond to defeat graphs in which $a \to b$ is included. Others defeat graphs can only be justified with orderings in which $val(b) \succ val(a)$ and, by connectedness

[2] A preorder $\succeq$ over a set V is connected iff for any $v_i, v_j \in V$, $v_i \succeq v_j$ or $v_j \succeq v_i$.

requirement, preservation of $a \rightarrow b$ needs to be justified with an ordering in which $val(a) \succeq val(b)$. So, $N_{\mathbf{P}}^{val(a) \succ val(b)}$ is also the set of supporters of $val(a) \succeq val(b)$ in **P'**. So, by independence, $val(a) \succeq val(b) \in F(\mathbf{P'})$. So, $a \rightarrow b \in \langle A, \rightarrow^{F(\mathbf{P'})} \rangle$. Contradiction. □

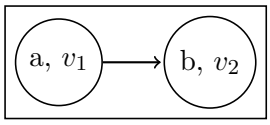
Corollary 1. *For any defeat aggregation rule F justifiable with connected preorders, $VAF = \langle A, \rightarrow, V, val \rangle$, a profile $\boldsymbol{AF}$ of defeat graphs of VAF, any independent preference aggregation rule F' restricted to connected preorders and a pair $\boldsymbol{P}, \boldsymbol{P'}$ of profiles preference orderings similar with respect to $\boldsymbol{AF}$, if $F'(P)$ justifies $F(\boldsymbol{AF})$, so does $F'(P')$.*

Proof. Take any independent defeat aggregation rule F justifiable with connected preorders, $VAF = \langle A, \rightarrow, V, val \rangle$, a profile **AF** of defeat graphs of *VAF*, any preference aggregation rule F' and any pair of profiles of preference orderings **P**, **P'** similar with respect to **AF**. Suppose that $F'(\mathbf{P})$ justifies $F(\mathbf{AF})$. Then, we know immediately that so does $F'(\mathbf{P'})$. □

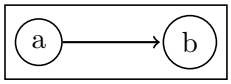
This result is very useful with respect to engineering aggregating defeat graphs. However, it ceases to hold when we consider preference aggregation based on arbitrary preorders.

Observation 1. *Corollary 1 does not hold when justification with arbitrary preorders is allowed.*

Example 4.4.1. *Consider the defeat aggregation rule F such that for any attack $a \rightarrow b$, it is selected iff all agents agree upon it. Also, take the preference aggregation function F' such that for any profile of preference orderings $\boldsymbol{P}$ and any pair of options v_i, v_j, $v_i \succeq v_j \in F'(\boldsymbol{P})$ if and only if all voters agree that it is the case. It is easy to check that this rule is independent. Now take the following VAF:*



and consider a profile $\boldsymbol{AF}$ of two defeat graphs of the VAF:



(d)

(e)

and consider two profiles of preorders over V justifying them: $\boldsymbol{P} = \langle\{v_1 \succeq v_2, v_2 \succeq v_1\}, \{v_2 \succeq v_1\}\rangle$*, and* $\boldsymbol{P'} = \langle\emptyset, \{v_2 \succeq v_1\}\rangle$*. Then,* $F(\mathbf{P})$ *amounts to* $v_2 \succ v_1$*, while* $F(\mathbf{P'})$ *is the empty preorder over* V*. Then clearly* $\langle A, \rightarrow F(\mathbf{P})\rangle \neq \langle A, \rightarrow^{F(\mathbf{P'})}\rangle$*, as the first one of them corresponds to the defeat graph (e), which justifies, while the latter to the defeat graph (d).*

4.4.2 Preference aggregation in terms of defeat aggregation

Let us now proceed to defining the simulation of preference aggregation rules in terms of defeat aggregation.

First, let us settle a useful notion of a profile of defeat graphs induced by a profile of preference orderings. This is simply a profile of defeat graphs of some *VAF* based on corresponding elements of the sequence of orderings.

Definition 26 (Induced profile of defeat graphs)**.** *Let* $VAF = \langle A, \rightarrow, V, val\rangle$ *be a VAF and* $\boldsymbol{P}$ *be a profile of preference orderings over* V*. Then, we say that the profile of defeat graphs* $\boldsymbol{AF} = \langle\langle A, \rightarrow^{P_1}\rangle, \ldots, \langle A, \rightarrow^{P_n}\rangle$*, where* P_i *is the* i*-th element of* P*, is the* profile induced *by* P*.*

Then, let us state when a defeat aggregation rule is a simulation of a considered preference aggregation rule.

Definition 27 (Simulation of preference aggregation rule)**.** *Let* F *be a preference aggregation rule. Also, let* $VAF = \langle A, \rightarrow, V, val\rangle$ *be a VAF. Then, a defeat aggregation rule* F' *is a simulation of* F *iff for any profile* $\boldsymbol{AF}$ *of defeat graphs of VAF, if* $\boldsymbol{AF}$ *is induced by a profile of preference orderings* $\boldsymbol{P}$*, then* $F'(\boldsymbol{AF})$ *is induced by* $F(\boldsymbol{P})$*.*

4.4.3 Preservation of axioms in simulating defeat aggregation

When we consider an application of a simulation of some argumentation aggregation we might wonder if it satisfies certain desired properties. In

particular, we are interested in finding if satisfaction of some axiom by an initial rule implies that its simulation is guaranteed to satisfy some condition. Further, as we will see, many axioms stated for the preference aggregation rules have their correspondents among defeat aggregation axioms. Thus, it is of special interest to check if satisfaction of these axioms implies satisfaction of their correspondents by the simulating rules.

Let us state some of such properties for the defeat aggregation. Their definitions are a modification of axioms provided by Chen and Endriss (2017).

The unanimity axiom states that if all agents agree that some attack should be included in the collective graph, then it is.

Definition 28 (Unanimity)**.** *F is anonymous if for any defeat aggregation problem $\boldsymbol{AF}_D$ consisting of some $VAF = \langle A, \rightarrow, V, val\rangle$ and a profile of defeat graphs $\boldsymbol{AF}$=$\langle AF_1, \ldots, AF_n\rangle$, if there is some pair of arguments $a, b \in A$ such that for any $AF_i \in \boldsymbol{AF}$ $a \rightarrow_i b$, $a \rightarrow b \in F(\boldsymbol{AF})$*

The anonymity condition expresses that a choice of attacks does not depend on the name of voters.

Definition 29 (Anonymity)**.** *F is anonymous if for any defeat aggregation problem $\boldsymbol{AF}_D$ consisting of some $VAF = \langle A, \rightarrow, V, val\rangle$ and a profile of defeat graphs $\boldsymbol{AF}$=$\langle AF_1, \ldots, AF_n\rangle$ it holds that for any permutation $\pi, F(\boldsymbol{AF}) = F(\pi\ (\boldsymbol{AF}))$.*

Independence states that all attacks are treated equally in any profile of defeat graphs.

Definition 30 (Independence)**.** *F is independent if for any defeat aggregation problem $\boldsymbol{AF}_D$ consisting of some $VAF = \langle A, \rightarrow, V, val\rangle$, an attacks $a \rightarrow b$ and for any pair of profiles of defeat graphs $\boldsymbol{AF}, \boldsymbol{AF'}$, if $N^{\boldsymbol{AF}}_{a\rightarrow b} = N^{\boldsymbol{AF'}}_{a\rightarrow b}$, then $a \rightarrow b \in F(\boldsymbol{AF})$ iff $a \rightarrow b \in F(\boldsymbol{AF'})$*

Monotonicity condition expresses that a selected attack should not be dropped if a support for it increases.

Definition 31 (Monotonicity)**.** *F is monotonic if it holds that for any pair of defeat aggregation problems $\boldsymbol{AF}$, $\boldsymbol{AF'}$ based on the same VAF=$\langle A, \rightarrow, V, val\rangle$, and any pair of arguments $a, b \in A$, if $N^{a\rightarrow b}_{\boldsymbol{AF}} \subseteq N^{a\rightarrow b}_{\boldsymbol{AF'}}$ and for any $a' \rightarrow b'$ such that $a' \rightarrow b' \neq a \rightarrow b$, $N^{a'\rightarrow b'}_{\boldsymbol{AF}} = N^{a'\rightarrow b'}_{\boldsymbol{AF'}}$, then if $a \rightarrow b \in F(\boldsymbol{AF})$, then $a \rightarrow b \in F(\boldsymbol{AF'})$.*

Furthermore, let us state some desired properties with respect to preference aggregation Notice, that the independence axiom was defined earlier.

We say that a preference aggregation function is unanimous if it never changes any ordering between options that all agents agree upon.

Definition 32 (Unanimity)**.** *A preference aggregation function F is unanimous iff in any profile of orderings $\boldsymbol{P}$ all voters submit that $v_i \succeq v_j$, then $v_i \succeq v_j$ in $F(\boldsymbol{P})$.*

We say that a preference aggregation function is anonymous if it provides the same output regardless of the ordering of items in its input.

Definition 33 (Anonymity)**.** *A preference aggregation function F is anonymous iff for any profile of orderings $\boldsymbol{P}$, any pair of items v_i, v_j and any permutation π of $\boldsymbol{P}$, $v_i \succ v_j \in F(\boldsymbol{P})$ iff $v_i \succ v_j \in F(\pi(\boldsymbol{P}))$.*

We call a preference aggregation function monotonic if increasing support for some already chosen ordering over items cannot cause dropping it.

Definition 34 (Monotonicity)**.** *A preference aggregation function F is monotonic iff for any two profiles of orderings $\boldsymbol{P}, P'$ over some set V and a pair of options $v_1, v_2 \in V$, if $v_1 \succeq v_2 \in F(P)$, then if $N_{\boldsymbol{P}}^{v_1 \succeq v_2} \subseteq_{\boldsymbol{P'}}^{v_1 \succeq v_2}$ and for pair of values v_1', v_2' such that $v_1', v_2' \neq v_1, v_2$, $N_{\boldsymbol{P}}^{v_1' \succeq v_2'} = N_{\boldsymbol{P'}}^{v_1' \succeq v_2'}$, then $v_1 \succeq v_2 \in F(\boldsymbol{P'})$*

Let us define formally the intuition behind the notion of the implication of some property. We wish to ensure that a preference aggregation rule simulating a defeat aggregation rule satisfying a certain property, satisfies its correspondent.

Definition 35 (Induction of a property)**.** *We say that a defeat aggregation rule property P_1 induces a property P_2 iff for any defeat aggregation rule F satisfying P_1, all its simulations satisfies P_2*

Unfortunately, it is usually not the case that defeat aggregation rules induce the correspondents of described axioms.

Observation 2. *It is not true that:*

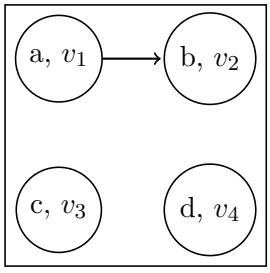
1. *Unanimity of defeat aggregation functions induces unanimity of their simulations.*

2. *Anonymity of defeat aggregation functions induces anonymity of their simulations.*

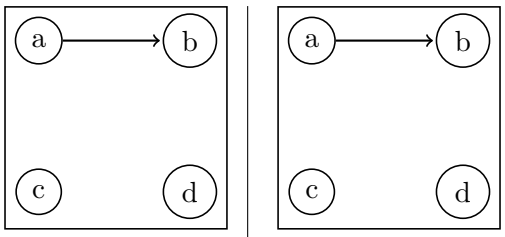
3. *Monotonicity of defeat aggregation functions induces monotonicity of their simulations.*

4. *Independence of defeat aggregation functions induces independence of their simulations.*

Example 4.4.2. *Consider the following VAF:*



Further, take a profile of its defeat graphs P:



Also, take a defeat aggregation rule F which outputs a graph identical to both elements of the profile

1. ***Unanimity**: Suppose that F is unanimous. Also, take a profile of preference orderings justifying the considered profile of defeat graphs:* $\boldsymbol{P'} = \langle v_4 \succ v_1 \succ v_2 \succ v_3, v_4 \succ v_1 \succ v_2 \succ v_3 \rangle$. *Then, we can consider a preference aggregation rule* F' *simulating F such that* $F'(\boldsymbol{P'}) = v_4 \succ v_3 \succ v_1 \succ v_2$. *Clearly,* $F'(\boldsymbol{P'})$ *justifies also the output. However,* F' *is obviously not unanimous.*

2. ***Anonimity**: Suppose that F is anonymous. Then, take a profile of preference orderings justifying the considered profile of defeat graphs:* $\boldsymbol{P'} = \langle v_4 \succ v_1 \succ v_2 \succ v_3, v_4 \succ v_3 \succ v_1 \succ v_2$. *Then, consider a preference aggregation rule* F' *simulating F such that* $F'(\boldsymbol{P'}) = v_4 \succ$

$v_1 \succ v_2 \succ v_3$ *but* $F'(\boldsymbol{P''}) = v_4 \succ v_3 \succ v_1 \succ v_2$*, where* $\boldsymbol{P''}$ *is the reversed* $\boldsymbol{P'}$*. Clearly, both outputs of* F' *justify the output of* F*, but* F' *is obviously not anonymous.*

3. ***Monotonicity****: Suppose that* F *is monotonic. Then, take two profiles of preference orderings justifying the considered profile of defeat graphs:* $\boldsymbol{P'} = \langle v_1 \succ v_2 \bowtie v_4 \succ v_3, v_1 \succ v_2 \bowtie v_4 \succ v_3 \rangle$ *and* $\boldsymbol{P''} = \langle v_1 \succ v_2 \succeq v_4 \succ v_3, v_1 \succ v_2 \succeq v_4 \succ v_3 \rangle$ *and a preference aggregation rule* F' *simulating* F *such that* $F'(\boldsymbol{P'}) = v_1 \succ v_2 \succeq v_4 \succ v_3$ *while* $F'(\boldsymbol{P''}) = v_1 \succ v_2 \bowtie v_4 \succ v_3$*. Clearly, both outputs of* F' *justify the output of* F*, but* F' *is obviously not monotonic.*

4. ***Independence****: Suppose that* F *is independent. Then, take two profiles of preference orderings justifying the considered profile of defeat graphs:* $\boldsymbol{P'} = \langle v_1 \succ v_2 \bowtie v_4 \succ v_3, v_1 \succ v_2 \bowtie v_4 \succ v_3 \rangle$ *and* $\boldsymbol{P''} = \langle v_1 \succ v_2 \succeq v_4 \succ v_3, v_1 \succ v_2 \succeq v_4 \succ v_3 \rangle$ *and a preference aggregation rule* F' *simulating* F *such that* $F'(\boldsymbol{P'}) = v_1 \succ v_2 \bowtie v_4 \succ v_3$ *and* $F'(\boldsymbol{P''}) = v_1 \succ v_2 \bowtie v_4 \bowtie v_3$*. Clearly, both outputs of* F' *justify the output of* F*, but* F' *is obviously not independent.*

4.4.4 Preservation of axioms in simulating preference aggregation

Having stated results concerning the implication of preference aggregation axioms by their simulations, it is interesting to investigate the reverse problem; namely to determine if defeat aggregation axioms imply their correspondents in preference aggregation.

Proposition 13. *Let* F *be a preference aggregation rule restricted to connected preorders. Then it holds that:*

1. *If* F *is unanimous, the induced defeat aggregation rule* F' *is unanimous.*

2. *If* F *is anonymous, the induced defeat aggregation rule* F' *is anonymous.*

3. *If* F *is monotonic, the induced defeat aggregation rule* F' *is monotonic.*

4. *For any independent preference aggregation rule* F*, restricted to connected preorders, the induced defeat aggregation rule* F' *is independent.*

Proof. 1. Take any unanimous preference aggregation rule F. Then, suppose that the induced defeat aggregation rule F' 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 $a \to b \notin F'(\mathbf{P})$. Then, by connectedness of all preference orderings, all voters submit that $val(a) \succeq val(b)$. But then, by unanimity of F, $val(a) \succeq val(b) \in F(\mathbf{P}^*)$, and thus $a \to b \in F'(\mathbf{P})$. Contradiction.

2. Take any anonymous preference aggregation rule F. Then, suppose that the induced defeat aggregation rule F' is not anonymous. Then, take a *VAF* and a profile of connected preorders $\mathbf{P}$ and a sequence $\mathbf{P'}$ of defeat graphs induced by $\mathbf{P}$ such that there is a permutation π of $\mathbf{P'}$ such that $F'(\mathbf{P'}) \neq F'(\pi(\mathbf{P'}))$. But now notice 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 monotonic preference aggregation rule F. Then, suppose that the induced defeat aggregation rule F' is not monotonic. Then take a *VAF* and two profiles of their defeat graphs$\mathbf{AF}$ and $\mathbf{AF'}$ such that there is some attack $a \to b$ such that $N^{a\to b}_{\mathbf{AF}} \subseteq N^{a\to b}_{\mathbf{AF'}}$, but $a \to b \in F(\mathbf{AF})$ while $a \to b \notin F(\mathbf{AF'})$. Then take a justification of these two profiles and of the outcomes of F. By connectedness of justifications we know that for any agent i submitting $a \to^i b$ it is the case that $val(a) \succeq_i val(b)$ for any justification. So, as we know that $N^{a\to b}_{\mathbf{AF}} \subseteq N^{a\to b}_{\mathbf{AF'}}$, we also know that $N^{val(a)\succeq val(b)}_{\mathbf{P}} \subseteq N^{val(a)\succeq val(b)}_{\mathbf{P'}}$. So, by monotonicity of F', $val(a) \succeq val(b) \in F'(P')$, so as F' simulates F, $a \to b \in F(\mathbf{AF'})$. Contradiction.

4. Take any independent preference aggregation rule F. Then, suppose that the induced defeat aggregation rule F' 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 $N^{a\to b}_{\mathbf{AF}} = N^{a\to b}_{\mathbf{AF'}}$, but $a \to b \in F'(\mathbf{AF})$ while $a \to b \notin F'(\mathbf{AF'})$. Notice that by connectedness of justifications, for any justification $\mathbf{P}, \mathbf{P'}$ of $\mathbf{AF}$, $\mathbf{AF'}$, $N^{val(a)\succeq val(b)}_{\mathbf{P}} = N^{val(a)\succeq val(b)}_{\mathbf{P'}}$. So, by independence, $a \succeq b \in F(\mathbf{P})$ iff $a \succeq b \in F(\mathbf{P})$. So, if $a \to b \in F'(\mathbf{AF})$, then $a \to b \in F'(\mathbf{AF'})$. Contradiction.

□

These results show the hazards of application of simulation of defeat aggregation by preference aggregation. While we can make sure that if we

simulate a preference aggregation rule with defeat aggregation its beneficial properties are preserved, we cannot be sure about it while simulating defeat aggregation.

4.5 Strategic Behavior

Following the study of properties of considered aggregation mechanisms, let us proceed to the study of strategic behavior within them.

Intuitively, the process of manipulation begins with determining if an agent is in favor of the decisive argument, or not. This is done by computing defeat graphs of the initial VAF based on her sincere preference orderings over values. Then, the aggregation is performed with employment of the agents' true preferences. Subsequently, agents might discover that they would be better off if they submitted an insincere view on either preferences over values or on the structure of defeat graphs.

In this section I will define the manipulation problem for both preference aggregation and defeat aggregation. Within these approaches, I will study under which conditions manipulation can be blocked. This will include both constraints on the construction of aggregation rules themselves, as properties of decision problems disabling strategic behavior. Furthermore, complexity of manipulation will be studied in both frameworks.

4.5.1 Manipulation in the preference aggregation approach

Let us begin with the study of manipulation in the setting based on the direct aggregation of preference orderings. As the investigation of strategic behavior in the proposed setting is a novel direction of research, the work in this section will follow several plausible restrictions.

Firstly, we will limit our investigations to the grounded extension. This is a significant simplification of the model. In this way we ensure that an extension of a desired type always exists and that it is unique. Also, we ensure that it is easily computable.

Secondly, we will assume that agents provide linear orderings over values. Also, it is assumed that aggregation rules output a linear ordering. While this choice is not fully plausible, it offers a strong simplification of the setting. The main reason behind taking this assumption is that it complies with the main line of research in preference aggregation.

As we will see in this section, unfortunately most intuitive rules are manipulable in general case. However, it is not difficult to ensure that the manipulation problem is computationally complex. Further, we can

also ensure strategy-proofness of our mechanism for *VAF*s enjoying some structural properties.

Let us now define what agents' preferences over outcomes of aggregation are. Recall that voters in our setting are interested in ensuring that the collective preference ordering induces a defeat graph in which the defeat graph is accepted if nd only if it is accepted in the graph induced by agents' own ordering.

Following this intuition we say that given a decision problem[3] and agents' ordering over values, an agent prefers some ordering to another if it treats the decisive argument consistently with agent's intentions, while the other does not. Notice that these preferences are dichotomous. Given a decision problem $DP = \langle VAF, D\rangle$ and a preference ordering P_i corresponding to same agent i, we say that i is in favor of D if it is in the grounded extension of the defeat graph induced by P_i. If it is not, we say that i is against D.

Definition 36 (Preferences Over Outcomes)**.** *Let* $DP = \langle VAF = \langle A, \rightarrow, V, val\rangle, D\rangle$ *be a decision problem and* i *be an agent with a preference ordering* P_i*. If* i *is* in favor of D*, then for any pair of preference orderings* P_1, P_2*,* $P_1 >_i P_2$ *iff* $D \in GRND_{\{A,\rightarrow^{P_1}\}}$ *while* $D \notin GRND_{\langle A,\rightarrow^{P_2}\rangle}$*. Also, if* i is against D*, then* $P_1 >_i P_2$ *iff* $D \notin GRND_{\langle A,\rightarrow^{P_1}\rangle}$ *while* $D \in GRND_{\langle A,\rightarrow^{P_2}\rangle}$*.*

With the definition of agents' preferences over outcomes of aggregation in hand, we can define when a preference aggregation function is strategy-proof with respect to the argumentation setting. As $F(P_i^*, \mathbf{P}_{-P_i})$ we denote the result of a preference aggregation function F for the preference ordering $\mathbf{P}$ with an ordering P_i replaced with P_i^*.

Definition 37 (Strategy-proofness with respect to argumentation)**.** *A preference aggregation rule* F *is strategy-proof with respect to argumentation iff for any profile of preference orderings* $\boldsymbol{P}$ *any agent* i *and any preference ordering* P_i^**, it is not the case that* $F(P_i^*, \boldsymbol{P}_{-P_i}) >_i F(\boldsymbol{P})$*.*

Let us also rephrase this definition as a computational problem.

MANIPULATION(F)

Instance: $DP = \langle VAF, D\rangle$, a profile of preference orderings $\mathbf{P}$, agent i.
Question: Is there a preference ordering P_i^* and such that $F(P_i^*, \mathbf{P}_{-P_i}) >_i F(\mathbf{P})$?

[3] As defined in Definition 13.

In the study of strategic behavior in aggregating preference orderings, a strong focus has been put on voting mechanisms. In this section I will relate the problem of manipulation in this setting to our approach.

Strategic Voting

Aggregating preference orderings is strictly connected with engineering voting rules. There, a group of voters elects an option out of a set of candidates. Mechanisms of this kind aim at ensuring that the winner of the elections represents agents' preferences accurately. Technically, a voting rule is a function $F : \mathcal{P}^m \to O$, where $\mathcal{P}$ is the set of all preference orderings over the set of options O and m is the number of voters. Notice that we have imposed that a rule always selects a single option. This property is known as the *resoluteness* condition.

Voting rules can be envisaged as preference aggregation rules. Then, the winner of elections is the top option of the collective preference ordering.

If this is the case, preferences of particular voters can be clearly defined. Each of them wants to make sure that the winner of the election is as good as possible from the perspective of their ranking.

Definition 38 (Strategic Voting Preferences)**.** *Let an agent i submit some ordering P_i over some set of options V. Then, for any pair of preference orderings P_1, P_2 over V, $P_1 >_i^V P_2$ iff $rank_{P_i}(top(P_1)) > rank_{P_i}(top(P_2))$.*

Then, we can ask if an agent can make herself better off with respect to strategic voting preferences by submitting an insincere preference ordering. If for some function F it is never the case, we say that F is strategy-proof with respect to voting preferences.

Definition 39 (Strategy-proofness in voting)**.** *A preference aggregation rule F is strategy-proof in voting iff for any profile of preference orderings $\boldsymbol{P}$ any agent i and any preference ordering P_i^*, it is not the case that $F(P_i^*, \boldsymbol{P}_{-P_i}) >_i^V F(\boldsymbol{P})$.*

This definition corresponds to a decision problem, in which we ask if there is some agent who would benefit from misrepresenting her views.

STRATEGIC VOTING(F)

Instance: Profile of preference orderings $\mathbf{P}$, agent i.
Question: Is there a preference ordering P_i^* such that $F(\mathbf{P}) >_i^V F(P_i^*, \mathbf{P}_{-P_i})$?

One of the crucial results in the social choice theory related to strategic behavior is related to a highly disadvantageous property of voting rules - being a dictatorship. This means that there is some individual whose most preferred option is always elected.

Definition 40 (Dictatorship with respect to strategic voting)**.** *Let F be a preference aggregation rule. We say that F is a dictatorship with respect to strategic voting iff there is some agent i such that for any profile of preferences orderings $\boldsymbol{P}$, $top(\boldsymbol{P}) = top(P_i)$.*

The Gibbard-Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975) states that any rule which is strategy-proof with respect to voting preferences is also dictatorial with respect to strategic voting. Its conditions involve non-imposition, which means that any option is elected by some preference ordering.

Theorem 2 (Gibbard - Satterthwaite [4])**.** *Any resolute, nonimposed, and strategy-proof voting rule for three or more alternatives must be a dictatorship.*

Application of voting mechanisms

As we will see it can be shown that any preference aggregation rule which is manipulable with respect to strategic voting, is also manipulable in the argumentation setting.

Proposition 14. *Any preference aggregation rule F which is manipulable with respect to voting preferences is also manipulable respect to argumentation.*

Proof. Consider any preference aggregation rule F which is manipulable with respect to strategic voting. This means that there is a set of voters $N = \{1, \ldots, n\}$, a set of options $V = \{v_1, \ldots, v_m\}$ and a profile of preference orderings submitted by voters $\mathbf{P} = \langle P_1, \ldots, P_n \rangle$ such that for some voter i, there is some preference ordering P_i^* over V such that $rank_i(top(F(\mathbf{P}))) < rank_i(top(F(\mathbf{P}^*)))$, where $\mathbf{P}^*$ is $\mathbf{P}$ with P_i replaced by P_i^*. Take such a profile. We will construct a decision problem $DP = \langle VAF, C \rangle$ which is manipulable by the successful manipulator with respect to strategic voter.

Let us take a set of values V and the set of arguments $A = \{a_1, \ldots, a_m\}$ (One per element of V). Further, take the valuation map val such that for any $a_i \in A$, $val(i) = v_i$. For simplicity let us say that $P_i = v_1 \succ_i v_2 \succ_i$

[4] Formulation from (Brandt et al., 2016, p.46)

$\cdots \succ_i v_m$. Now, let a_1 be the decisive argument. Then, let v_j correspond to top(F($\mathbf{P}$)). Construct the attack relation so that $a_j \rightarrow a_1$. Also, for any v_b such that $rank_i(v_b) > rank_i(j)$, let $a_b \rightarrow a_j$. No other attacks are considered. A simplified example of such a *VAF* is depicted below.

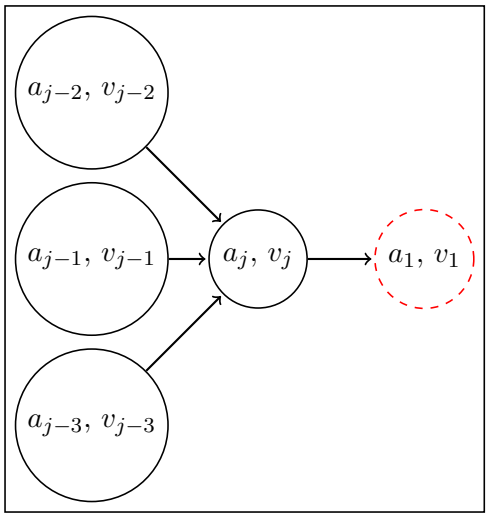


Figure 4.6: Illustration of the construction.

Then firstly notice that for the agent i the argument a_1 should be accepted, as it is in the grounded extension of the defeat graph based on i's preference ordering. However, it is not included in the grounded extension of the defeat graph based on $F(\mathbf{P})$, as all attacks on a_j are eliminated because v_j is the top value. However, we know that i can submit an ordering P^* such that some $v_b \succ_i v_j$ becomes the top option. Then, clearly one of the attackers of v_j, which is the only attacker of the decisive argument is the top option, so the attack is preserved. Therefore, v_i is accepted in the new defeat graph. So i manipulated successfully. □

This result is followed by an unfortunate conclusion. Namely, it turns out that for any preference aggregation rule F based on strict preferences, if F is not dictatorial with respect to strategic voting, it is manipulable in the current setting.

To justify this claim it is sufficient to take any preference aggregation rule F based on strict preferences and suppose that it is not dictatorial with respect to strategic voting. Then, by Gibbard-Satterwaite theorem we know that it is manipulable with respect to strategic voting. But then it follows

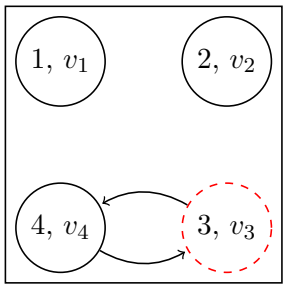
that it is also manipulable in the argumentation setting.

This means that aggregating *VAF*s by preference aggregation is at least as vulnerable to strategic behavior as voting mechanisms. Unfortunately, this is not the end of its unfortunate properties. In fact, we can find cases in which rules strategy-proof with respect to strategic voting are manipulable within the argumentation setting.

Observation 3. *It is not true that if a rule F is strategy-proof in voting, it is strategy-proof with respect to argumentation.*

Example 4.5.1. *Consider the following preference aggregation rule F: For any profile of preference orderings $\boldsymbol{P}$ distinguish a dictator d. Then, $top(F(\boldsymbol{P})) = top(P_d)$. To determine the rest of the collective preference ordering, eliminate the value $top(P_d)$ from the profile of orderings. Apply the Borda rule to the remainder of the profile. This rule is strategy-proof with respect to strategic voting, as the top option is known from the start.*

Now consider the following decision problem DP:



Now take a profile $\boldsymbol{P}$, where agent d is the dictator:

- $d : v_2 \succ v_1 \succ v_4 \succ v_3$
- $m : v_3 \succ v_4 \succ v_1 \succ v_2$
- $o : v_4 \succ v_3 \succ v_2 \succ v_1$

Let us now notice that the score of v_1 is 3, v_3 receives 5 points, and v_4 gets 6 . The score of v_2 does not matter as it is dictator's top option. So, we get $F(\boldsymbol{P}) = v_2 \succ v_4 \succ v_3 \succ v_1$. It is easy to see that under this ordering argument 3 is not in the grounded extension, as the attack $4 \to 3$ is preserved, while $3 \to 4$ is not. This leaves agent m dissatisfied, as in the defeat graph based on her preferences argument 3 is clearly accepted.

Consider, however, the profile:

- $d : v_2 \succ v_1 \succ v_4 \succ v_3$

- $m^* : v_3 \succ v_2 \succ v_1 \succ v_4$
- $o : v_4 \succ v_3 \succ v_2 \succ v_1$

After this change the score of v_1 is 3, v_3 receives 5 points, and v_4 gets 6. Therefore, the collective preference ordering is $v_2 \succ v_4 \succ v_3 \succ v_1$. Under this ordering 3 is accepted in the collective structure. So, m manipulated successfully.

This means that the current setting is strictly less immune to strategic behavior than strategic voting. However, we can show that for a large class of rules, the manipulation problem is difficult to compute. Thanks to this observation we can claim that the proposed mechanism can eliminate manipulation in practical applications. As we will show, if a rule is NP-hard with respect to the problem of strategic voting, so it is with respect to our setting. We know that a number of rules is NP-hard with respect to the strategic voting problem. For example, STV enjoys this property (Bartholdi & Orlin, 1991).

Proposition 15. *For any preference aggregation rule F for which the strategic voting problem is NP-hard, so is the VAF-manipulation problem with respect to F.*

Proof. Take any preference aggregation rule F for which the strategic voting problem is NP-hard. Let us show the way to reduce this problem to manipulation in the current setting. Take a profile of preference orderings $\mathbf{P}$ and an agent i with a preference ordering P_i. Let us construct a decision problem in which i can manipulate if and only if she can manipulate with respect to strategic voting. The construction is parallel to the one depicted in the Figure 4.6. As before, take a *VAF* in which we have an argument corresponding to any ranked option. Also, map each of the options as values of corresponding arguments. Further, let i's favourite option correspond to the decisive argument - a_i. Now, let the argument a_j, corresponding to $top(F(\mathbf{P}))$ attack a_i iff $a_i \neq a_j$. Also, let any argument a_b such that $val(a_b) \succ_{P_i} val(a_i)$ attack a_j. Clearly, i is in favor of a_i.

We need to show that i can manipulate with respect to argumentation setting iff she can manipulate with respect to strategic voting. If i can manipulate with respect to argumentation setting, then there is a preference ordering P_i^* such that $rank_i(top(F(P_i^*, \mathbf{P}_{-P_i}))) > rank_i(top(F(\mathbf{P})))$. Otherwise, a_i would not be in the grounded extension of the defeat graph induced by $F(P_i^*, \mathbf{P}_{-P_i})$. But then, i can manipulate with respect to strategic voting. But also, if there is a preference ordering P_i^* such that $rank_i(top(F(P_i^*, \mathbf{P}_{-P_i}))) >$

$rank_i(top(F\mathbf{P}))$, then D becomes in the grounded extension of the defeat graph induced by $F(P_i^*, \mathbf{P}_{-P_i})$. So, i can manipulate with respect to manipulation. □

In addition to this result concerning the computational complexity of the manipulation problem, we can select a large class of decision problems for which all rules are strategy-proof. This can be the case if the decisive argument is accepted, or rejected, from the perspective of any audience. This point holds for both considered approches.

As we have seen before, we encounter such a case when the decisive argument is guarded.

Proposition 16. *For any preference aggregation rule F, F is strategy-proof for decision problems in which the decisive argument is guarded.*

Proof. Take any decision problem $DP = \langle\langle A, \rightarrow, V, val\rangle, D\rangle$ in which D is guarded. Then, as we have shown in Proposition 7, D cannot be accepted in the grounded extension under any preference ordering over values. So, agents can only submit preference orderings rejecting it, so they cannot manipulate. □

4.5.2 Manipulation in defeat aggregation

Let us now proceed to the study of manipulation in the aggregation of defeat graphs. It is worth noting that the nature of strategic behavior in this setting is largely similar to manipulation in preference aggregation. As before, agents are concerned with making sure that the decisive argument is accepted in the collective structure if and only if it is accepted in their own. What differs in this setting from the previously discussed is that the manipulation does not take into account the exact preferences over values that agents have. Instead, it focuses only on the relative strength of arguments in which agents believe.

In this study we will focus on connections between the manipulability of defeat aggregation and manipulability of preference aggregation. This approach requires making strong assumptions.

Firstly, we will study how strategic behavior in the setting of defeat aggregation relates to the strategic behavior rules in their simulations. This will force us to restrict the domain of the study to rules which are preserving being an audience. However, this restriction is justified. We are looking for rules which always provide us with the outcome explainable by some preference ordering over values.

Secondly, a stricter restriction comes into play as we wish to use the similar assumptions to those used in results concerning manipulation in preference aggregation. We would will use defeat aggregation functions which only take as input defeat graphs justifiable with linear orderings, and always outputting graphs satisfying this condition. This restriction applies to rules studied henceforth. Overcoming this restriction would be an interesting avenue for further research. In this section we will only refer to defeat aggregation rules satisfying the mentioned restrictions.

Let us first define agents' preferences over the outcomes of the aggregation of defeat graphs. We say that an agent i is in favor of D if D is in $GRND_{AF_i}$, where AF_i is i's sincere defeat graph. If i is not in favor of D, she is against it.

Definition 41 (Preferences Over Collective Defeat Graphs)**.** *Let $DP = \langle VAF = \langle A, \rightarrow, V, val\rangle, D\rangle$. Also, take an agent i and her sincere defeat graph AF_i of VAF. Then, for any pair AF_1, AF_2 of defeat graphs of some VAF, if i is in favor of D, $AF_1 >_i^{def} AF_2$ when $D \in GRND_{AF_1}$ while $D \notin GRND_{AF_2}$. Otherwise, $AF_1 >_i^{def} AF_2$ if $D \notin GRND_{AF_1}$ while $D \in GRND_{AF_2}$.*

With the definition of agents' preferences over outcomes of the aggregation process in hand, we can define the manipulation problem with respect to a given defeat aggregation function in the current context. We say that an agent i can manipulate a rule F if she can replace her sincere defeat graph in order to ameliorate the outcome of defeat aggregation for herself. A defeat aggregation rule is said to be strategy-proof, if it is never the case.

As $F(AF_i^*, \mathbf{AF}_{-AF_i})$ we denote the result of a preference aggregation function F for the profile $\mathbf{AF}$ of defeat graphs of some with a defeat graph AF_i replaced with AF_i^*.

Definition 42. *A defeat aggregation rule F is strategy-proof iff for any profile of defeat graphs $\boldsymbol{AF}$ of some VAF, any agent i and any defeat graph AF_i^*, it is not the case that $F(AF_i^*, \boldsymbol{AF}_{-AF_i}) >_i^{def} F(\boldsymbol{AF})$.*

Let us rephrase this definition as a computational problem.

MANIPULATION IN DEFEAT AGGREGATION (F)

Instance: Decision problem $DP = \langle VAF = \langle A, \rightarrow, V, val\rangle, D\rangle$, a profile $\mathbf{AF}$ of defeat graphs of VAF, agent i.
Question: Is there a defeat graph AF_i^* of VAF such that $F(\mathbf{AF}^*) >_i^{def} F(\mathbf{AF})$?

We will study this approach with employment of the simulation of defeat aggregation with preference aggregation.

Proposition 17. *A defeat aggregation rule F is manipulable iff some preference aggregation rule simulating it is manipulable.*

Proof. ($\Rightarrow$) Take any manipulable defeat aggregation rule F. Now suppose that some preference aggregation rule simulating it are strategy-proof. Also, consider a rule F' simulating F. Then, take a decision problem $DP = \langle VAF, D\rangle$ and a profile of defeat graphs $\mathbf{AF}$ of VAF. Also, consider an agent i who can manipulate F with respect to $\mathbf{AF}$, by submitting some defeat graph AF_i^*. Without loss of generality assume that i is in favor of D. This means that $D \notin F(\mathbf{AF})$, but $D \in F(AF_i^*, \mathbf{AF}_{-AF_i})$. Also, consider any justification $\mathbf{P}$ of $\mathbf{AF}$. Notice that as F' justifies F, $D \notin GRND_{\langle A, \to^{F'(\mathbf{P})}\rangle}$.

Now we can consider a preference ordering P_i^* in justifying AF_i^*. Notice that as F' justifies F, it must be the case that $D \in F'(P_i^*, \mathbf{P}_{-P_i^*})$. So, i could manipulate in F'.

($\Leftarrow$) Symmetric. □

This result shows us that if we require defeat aggregation rules to satisfy the taken restrictions, we need to accept that it is vulnerable to strategic behavior.

Corollary 2. *For any strategy-proof defeat aggregation F, all its simulations are dictatorial with respect to strategic voting.*

Proof. Take any strategy-proof defeat aggregation rule F. We know, that all of its simulations are strategy-proof. But then we know, that they are also strategy-proof with respect to strategic voting. So, they need to be dictatorial. □

However, we can cope with this problem by choosing functions which are computationally difficult to manipulate. It is easy to show that defeat aggregation rules whose simulations are difficult to manipulate are hard to manipulate themselves.

Proposition 18. *For any defeat aggregation function F preserving being an audience and its simulation F', if manipulating F' is NP-hard, manipulating F is also NP-hard.*

Proof. Take any defeat aggregation function F which is simulated by the preference aggregation function F' restricted to strict preference orderings, which is NP-hard to manipulate. Then take an agent i, a VAF and a profile

of its defeat graphs **AF**. Now compute their justifications **AF**. Now an agent can compute a way to manipulate F'. But then by Proposition 17 we know that she can also get the way to manipulate F. So we reduced the problem of manipulating defeat aggregation rule to manipulating a preference aggregating function, which is NP-hard. □

4.6 Conclusions

In this chapter we have studied two ways of aggregation of agents' views on preferences over values. In the first of them we considered the application of preference aggregation to the determination of a collective view on the importance of values. In the second, we studied an application of graph aggregation techniques to aggregating defeat graphs. We have shown a limitation of this approach, namely, we have shown that no quota rule can ensure that a collective argumentation framework is a defeat graph of the initial *VAF*.

Further, we have studied the ways of simulating preference and defeat aggregation in terms of each other. In particular, we have shown that simulating preference aggregation with defeat aggregation preserves a number of desirable properties. This is not the case, however, when simulation of defeat aggregation with preference aggregation is considered.

Finally, we have studied the strategic behavior within both considered settings. For the preference aggregation approach, we showed that strictly more rules are manipulable with respect to the studied decision-making setting than with respect to strategic voting. This means, that any rule strategy-proof with respect to aggregating *VAF*s is dictatorial with respect to strategic voting. It is worth noting that this is not necessarily a very problematic result. A strategy-proof preference aggregation function which is only dictatorial with respect to one value can still be fair with respect to a large part of the preference ordering.

We have also studied strategic behavior with respect to defeat aggregation. We have shown that defeat aggregation rules preserving being justified with linear orderings over values are manipulable if and only if their simulations are.

Chapter 5

Conclusions and further research

5.1 Conclusions

The starting point of the thesis was to establish the connections between value-based argumentation and decision-making. We focused on argumentation frameworks in which a single argument was decisive. Then, we stipulated that the decision is made if the decisive argument is selected as an outcome of discussion.

Following this assumption, we have studied strategic behavior of agents involved in the process of reaching a decision. We studied two distinctive types of behavior. We investigated situations in which a single agent is responsible for choosing an action. Further, we moved on to the study of collective decision-making.

In the single agent case we were interested in individuals who are willing to deviate from their sincere preferences over values in order to make sure that their desired decision is taken. We studied the computational complexity of such a behaviour, extending results due to Dunne (2007) and Dunne and Bench-Capon (2004). We generalized their findings to relaxed assumptions regarding agents' preferences over values. We also investigated the complexity of finding an ordering over values preserving the decisive argument which is minimally different from agents' sincere hierarchy. Then, we provided some restrictions on the structure of VAFs under which strategic behavior is not possible.

Further, in the multi-agent scenario, we studied applications of social choice mechanisms to aggregating views on preferences over values. We

focused on two approaches. In the first we considered aggregating defeat graphs submitted by particular agents, using graph aggregation methods. In the second, we used preference aggregation functions to determine a collective preference ordering over values. We studied connections between these two approaches, including translations between them.

Further, we have studied strategic behavior within the proposed models for collective decision-making. We used results concerning strategic voting to establish conditions for manipulability in preference aggregation. Following this connection we also obtained results concerning the complexity of manipulation problem in the preference aggregation approach. Further, we used the possibility of translating preference aggregation into defeat aggregation to provide results concerning manipulation within this framework.

5.2 Future work

The results provided in the thesis leave a vast room for further research. In the single agent setting we have focused on agents who wish to push the decision forward. However, we can imagine that particular agents might have an incentive to make sure that the decision is not taken. It would be interesting to extend our results to account for such agents. Another issue which was not resolved in this section is the complexity of finding preferences orderings saving decisive arguments minimally different than the agent's sincere hierarchy for semantics other than the grounded semantics.

Furthermore, it is worth noting that we have interpreted searching for preference orderings preserving decisive arguments as a negative behavior. But the negative approach is not the only way to see it. We can imagine that a person who wants to justify the action she desires genuinely wishes to change her beliefs in order to stay consistent with her incentives. Possibly, she might want to change her beliefs only to a minimal needed extent. Then, we might be interested to study the ways in which a desired ordering can be found easily. For instance, it would be interesting to develop heuristic algorithms capable of finding preference orderings preserving arguments quickly despite the high complexity of the problem in general case.

In the multi-agent setting the study of particular defeat aggregation rules was left out. It would be interesting to develop specific functions designed to perform well in the studied context. Further, the study of strategic behavior can be extended further. It would be interesting to study the problem of manipulating in the minimal way, similarly to the manner in which it was studied in the single agent setting. Also, it would be beneficial to investigate

the complexity of manipulation for semantics other than grounded. Finally, in the multi-agent setting we have assumed that agents' incentives are always consistent with the acceptance or rejection of the decisive argument in the defeat graph based on agents' sincere preferences over values. It would be highly interesting to investigate the possibility of manipulating the studied mechanisms when agents' incentives are not dependent on their preferences over values.

Moreover, in the current work we have focused on a restricted class of decision problems. It would be of high interest to study the problems considered in this thesis with respect to argumentations which might lead to non-binary decisions.

References

Airiau, S., Bonzon, E., Endriss, U., Maudet, N., & Rossit, J. (2016). Rationalisation of Profiles of Abstract Argumentation Frameworks. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multi-Agent Systems (AAMAS)* (pp. 350–357).

Amgoud, L., & Cayrol, C. (1998). On the Acceptability of Arguments in Preference-Based Argumentation. In *Proceedings of the 1998 Conference on Uncertainty in Artificial Intelligence (UAI)* (pp. 1–7).

Amgoud, L., & Prade, H. (2009). Using Arguments for Making and Explaining Decisions. *Artificial Intelligence*, *173*(3), 413 – 436.

Arrow, K. J. (1951). Social Choice and Individual Values.

Awad, E., Bonnefon, J.-F., Caminada, M., Malone, T. W., & Rahwan, I. (2017). Experimental Assessment of Aggregation Principles in Argumentation-Enabled Collective Intelligence. *ACM Transactions on Internet Technology (TOIT)*, *17*(3), 29.

Awad, E., Booth, R., Tohmé, F., & Rahwan, I. (2015). Judgement Aggregation in Multi-Agent Argumentation. *Journal of Logic and Computation*, *27*(1), 227–259.

Awad, E., Caminada, M. W., Pigozzi, G., Podlaszewski, M., & Rahwan, I. (2017). Pareto Optimality and Strategy-Proofness in Group Argument Evaluation. *Journal of Logic and Computation*, *27*(8), 2581–2609.

Bartholdi, J. J., & Orlin, J. B. (1991). Single Transferable Vote Resists Strategic Voting. *Social Choice and Welfare*, *8*(4), 341–354.

Bench-Capon, T. (2002). Value Based Argumentation Frameworks. *arXiv preprint cs/0207059*.

Bench-Capon, T. (2003). Persuasion in Practical Argument Using Value-Based Argumentation Frameworks. *Journal of Logic and Computation*, *13*(3), 429–448.

Bench-Capon, T., Doutre, S., & Dunne, P. E. (2007). Audiences in Argumentatio Frameworks. *Artificial Intelligence*, *171*(1), 42–71.

Bodanza, G., Tohmé, F., & Auday, M. (2017). Collective Argumentation: A Survey of Aggregation Issues Around Argumentation Frameworks. *Argument & Computation*, *8*(1), 1–34.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., & Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

Caminada, M. (2008). A Gentle Introduction to Argumentation Semantics. *Lecture material, Summer*.

Caminada, M., & Pigozzi, G. (2011). On Judgment Aggregation in Abstract Argumentation. *Proceedings of the 2011 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, *22*(1), 64–102.

Caminada, M., Pigozzi, G., & Podlaszewski, M. (2011). Manipulation in Group Argument Evaluation. In *Proceedings of the 2011 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1127–1128).

Chen, W., & Endriss, U. (2017). Preservation of Semantic Properties During the Aggregation of Abstract Argumentation Frameworks. In *Proceedings of the 2017 Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*.

Delobelle, J., Haret, A., Konieczny, S., Mailly, J.-G., Rossit, J., & Woltran, S. (2016). Merging of Abstract Argumentation Frameworks. *Proceedings of the 2016 Knowledge Representation (KR)*, 33–42.

Deza, M. M., & Deza, E. (2009). Encyclopedia of Distances. Springer.

Dung, P. M. (1995). On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, *77*(2), 321–357.

Dunne, P. E. (2007). Computational Properties of Argument Systems Satisfying Graph-Theoretic Constraints. *Artificial Intelligence*, *171*(10-15), 701–729.

Dunne, P. E., & Bench-Capon, T. (2004). Complexity in Value-Based Argument Systems. In *European Workshop on Logics in Artificial Intelligence (JELIA)* (pp. 360–371).

Dunne, P. E., Hunter, A., McBurney, P., Parsons, S., & Wooldridge, M. (2011). Weighted Argument Systems: Basic Definitions, Algorithms, and Complexity Results. *Artificial Intelligence*, *175*(2), 457–486.

Dunne, P. E., & Wooldridge, M. (2009). Complexity of Abstract Argumentation. In *Argumentation in Artificial Intelligence* (pp. 85–104). Springer.

Endriss, U., & Grandi, U. (2017). Graph Aggregation. *Artificial Intelligence*, *245*, 86–114.

Gärdenfors, P. (1976). Manipulation of Social Choice Functions. *Journal of Economic Theory*, *13*(2), 217–228.

Gibbard, A. (1973). Manipulation of Voting Schemes: a General Result. *Econometrica: Journal of the Econometric Society*, 587–601.

Kakas, A., & Moraitis, P. (2003). Argumentation Based Decision Making for Autonomous Agents. In *Proceedings of the 2003 International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 883–890).

Maudet, N., Parsons, S., & Rahwan, I. (2006). Argumentation in Multi-Agent Systems: Context and Recent Developments. In *Proceedings of International Workshop on Argumentation in Multi-Agent Systems* (pp. 1–16).

Modgil, S. (2009). Reasoning About Preferences in Argumentation Frameworks. *Artificial Intelligence*, *173*(9-10), 901–934.

Perelman, C. (1971). The New Rhetoric. In *Pragmatics of Natural Languages* (pp. 145–149). Springer.

Pu, F., Luo, J., Zhang, Y., & Luo, G. (2013). Social Welfare Semantics for Value-Based Argumentation Framework. In *Proceedings of International Conference on Knowledge, Science, Engineering and Management* (pp. 76–88).

Rahwan, I., & Larson, K. (2008). Mechanism Design for Abstract Argumentation. In *Proceedings of the 2008 International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1031–1038).

Satterthwaite, M. A. (1975). Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *Journal of Economic Theory*, *10*(2), 187–217.