Rationalisation of Profiles of Abstract Argumentation Frameworks

Airiau, S.; Bonzon, E.; Endriss, U.; Maudet, N.; Rossit, J.

Publication date
2016

Document Version
Final published version

Published in
AAMAS'16

Rationalisation of Profiles of Abstract Argumentation Frameworks

Stéphane Airiau
LAMSADE, Univ. Paris-Dauphine
stephane.airiau@dauphine.fr

Elise Bonzon
LIPADE, Univ. Paris Descartes
elise.bonzon@parisdescartes.fr

Ulle Endriss
ILLC, University of Amsterdam
ulle.endriss@uva.nl

Nicolas Maudet
LIP6, Univ. Pierre et Marie Curie
nicolas.maudet@lip6.fr

Julien Rossit
LIPADE, Univ. Paris Descartes
julien.rossit@parisdescartes.fr

ABSTRACT

Different agents may have different points of view. This can be modelled using different abstract argumentation frameworks, each consisting of a set of arguments and a binary attack-relation between them. A question arising in this context is whether the diversity of views observed in such a profile of argumentation frameworks is consistent with the assumption that every individual argumentation framework is induced by a combination of, first, some basic factual attack-relation between the arguments and, second, the personal preferences of the agent concerned. We treat this question of *rationalisability* of a profile as an algorithmic problem and identify tractable and intractable cases. This is useful for understanding what types of profiles can reasonably be expected to come up in a multiagent system.

Keywords

Argumentation; Social Choice Theory

1. INTRODUCTION

The model of abstract argumentation introduced by Dung [12] is at the root of a vast amount of work in artificial intelligence and multiagent systems. In a nutshell, this model abstracts away from the content of an argument, and thus sees argumentation frameworks as directed graphs, where the nodes are arguments and the edges are attacks between arguments—in the sense that one argument undercuts or contradicts another argument. Different semantics provide principled approaches to selecting sets of arguments that can be viewed as coherent when taken together. The simplicity and generality of this framework, as well as its links with nonmonotonic reasoning, have stimulated a number of directions of research, e.g., at the level of the definition of the semantics, of their computation, of the expressivity of such frameworks, or regarding their application in a multiagent system.

In recent years, a number of authors have addressed the problem of aggregating several argumentation frameworks, each associated with the stance taken by a different individual agent, into a single collective argumentation framework that would appropriately represent the views of the group as a whole. Examples include the contributions of Coste-Marquis et al. [11], Tohmé et al. [27], Bodanza and Auday [8], and Dunne et al. [14]. Aggregating argumentation frameworks is a form of graph aggregation [15]: We are given a profile of attack-relations, one for each agent, and are asked to compute a suitable compromise attack-relation. This is an interesting and fruitful line of research, bringing together concerns in abstract argumentation with the methodology of social choice theory,[1] but it raises one important question: For a given profile of argumentation frameworks, is it in fact conceivable that that profile would manifest itself? Intuitively speaking, it will often seem more natural to encounter a profile with similar individual attack-relations rather than one with attack-relations that differ radically. How do we explain the differences in perspective of the individual agents for a given profile?

The point that the attack-relation should not be viewed as absolute and objective, but may very well depend on the individual circumstances of the agent considering the arguments in question, has been made before by multiple authors [2, 4, 9, 17, 16]. In fact, it is central to the study of argumentation, as also suggested by Modgil [22], who noted that abstract argumentation frameworks "should more properly be viewed as modelling human reasoning and debate, rather than as abstractions of underlying theories in some formal logic." A widespread explanation for such diversity of views is that agents have different preferences regarding the arguments at hand. For instance, arguments may come from different sources, which agents may trust more or less. Or arguments may be attached to different values, which agents may prioritise differently. This perspective still assumes an underlying ground truth, which however may be interpreted differently, depending on the agents. The same position is also taken by Searle [26]:

> "Assume universally valid and accepted standards of rationality, assume perfectly rational agents operating with perfect information, and

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

[1]The approach sketched here must be clearly distinguished from a second approach combining abstract argumentation and social choice theory found in the literature, which addresses the question of how to aggregate different extensions (or labellings) for a common argumentation framework. This is the approach of, amongst others, Caminada and Pigozzi [10] as well as Rahwan and Tohmé [24]. Bodanza and Auday [8] compare the two approaches explicitly.

> you will find that rational disagreement will still occur; because, for example, the rational agents are likely to have different and inconsistent values and interests, each of which may be rationally acceptable." (page *xv*)

In the literature on abstract argumentation, frameworks for modelling this phenomenon have been proposed by several authors, including Amgoud and Cayrol [1] and Bench-Capon [6]. Here we adopt a preference-based approach, in the *value-based* variant originally due to Bench-Capon [6]. In his model, whether argument A ultimately defeats argument B does not only depend on whether A attacks B in an objective sense, but also on how we rank the importance of the social or moral values attached to A and B: If we rank the value associated with B strictly above that associated with A, we may choose to ignore any attacks of A on B.

At the technical level, we thus ask the following question: Given a profile of argumentation frameworks $(AF_1, \ldots, AF_n)$, one for each agent, can this profile be explained in terms of a single master argumentation framework, an association of arguments with values, and a profile of preference orders over values $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$, one for each agent? Or, as we shall put it: Can the profile of argumentation frameworks observed be *rationalised*? To be able to answer this question in the affirmative, for every agent i, we require AF_i to be exactly the argumentation framework we obtain when the master argumentation framework with its associated values is reduced using the preference order $\succcurlyeq_i$.

Of course, alternative justifications can be given for the fact that individual argumentation frameworks may differ, not just the preference-based explanation adopted here. In particular, agents may interpret arguments differently, especially when they are incomplete [7]. Also, while we adopt Bench-Capon [6]'s value-based approach as the technical foundation on the basis of which to construct our framework and for which to prove our results, there are alternative models of preference-based argumentation, for instance relying on meta-level argumentation [21]. We do not wish to commit to one specific view on the complex question of how to best model preferences in argumentation (see the work of Amgoud and Vesic [3] for an example of a contribution to this debate). Indeed, we believe that our general point is relevant beyond such specific modelling choices, and we see our contribution to be first and foremost as a methodological one. The same type of investigation could be undertaken for other models as well.[2] In a sense, this multiplicity of models is precisely what makes our contribution useful: by providing a collection of results that allow to check whether a profile can be rationalised on such grounds, we provide evidence for guiding the modelling process. The good news is that in many—albeit not all—cases verification of rationalisability can be performed efficiently, even when the assignment of values to arguments is not known beforehand.

The remainder of this paper is organised as follows. Section 2 presents the relevant background regarding value-based argumentation. Section 3 formally introduces the problem of rationalising a given profile of argumentation frameworks provided by a set of agents, and presents the different types of constraints on solutions we will consider. Section 4 analyses the single-agent case in detail, while Section 5 investigates the multiagent case. Finally, Section 6 discusses a number of application scenarios for our approach and Section 7 concludes with a review of open questions and possible directions for future work.

2. NOTATION AND TERMINOLOGY

Following Dung [12], below we define an *argumentation framework* (AF) as a binary attack-relation declared over a set of arguments. We will restrict ourselves to scenarios for which the set of available arguments is finite.

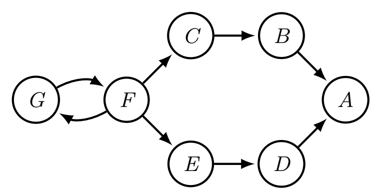
Definition 1 (AF). *An argumentation framework is a pair $AF = \langle Arg, \rightharpoonup \rangle$, where Arg is a finite set of arguments and $\rightharpoonup$, the attack-relation, is an irreflexive binary relation defined on Arg.*

If $A \rightharpoonup B$ holds for $A, B \in Arg$, then we say that A *attacks* B.

Example 1. *Pollution in the big cities is becoming a major health problem. City councils are facing the question of possibly banning polluting vehicles, and specifically diesel cars, from the inner centres of such cities. A city council might be entertaining the following arguments:*

- (A) *Diesel cars should be banned from in the inner city centre in order to decrease pollution.*
- (B) *Artisans, who deserve special protection by the city council, cannot change their vehicles, as that would be too expensive for them.*
- (C) *The city can offer financial assistance to artisans.*
- (D) *There are few alternatives: autonomy of electric cars is poor, as there are not enough charging stations around.*
- (E) *The city can set up more charging stations.*
- (F) *In times of financial crisis, the city should not commit to spending additional money.*
- (G) *Health and climate change issues are important, so the city has to spend what is needed to tackle pollution.*

The following graph shows the AF generated by these arguments and a natural attack-relation $\rightharpoonup$ between them:



Observe that for this AF it is ambiguous whether or not we should accept argument A and ban diesel cars: Accepting either $\{A, C, E, G\}$ or $\{B, D, F\}$ is intuitively admissible.

Recall that a *preorder* is a binary relation that is reflexive and transitive, and a *weak order* in addition is also complete [25]. We use preorders and weak orders to model *preferences*. Using a preorder means allowing for strict preferences, indifferences, and incomparabilities, while using a weak order excludes the possibility of two items being incomparable. We will use the terms 'preference order' and 'preorder' synonymously, i.e., a 'complete preference order' refers to a weak order. The strict part of a preference order $\succcurlyeq$ is denoted as $\succ$ and its indifference part as $\sim$.

[2]While some preference-based approaches are special cases of the one used here—e.g., in the work of Amgoud and Cayrol [1] each argument is mapped to a different value—others would require extensions, e.g., allowing several values per argument as in the work of Kaci and van der Torre [19].

Following Bench-Capon [6], we define an *audience-specific value-based argumentation framework* (AVAF) as an AF equipped with a function associating each argument with the social or moral value it advances, combined with a preference order declared over those values. While the mapping from arguments to values is fixed, the preferences over values are those of a particular agent (the "audience").

Definition 2 (AVAF). *An audience-specific value-based argumentation framework is defined as a 5-tuple $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, where $\langle Arg, \rightharpoonup \rangle$ is an argumentation framework, Val is a finite set of values, $val : Arg \to Val$ is a mapping from arguments to values, and $\succcurlyeq$ is the audience's preference order on Val.*

We call $\langle Val, val \rangle$ the AVAF's *value-labelling*. Let $=_{val}$ be the equivalence relation on arguments induced by val: $A =_{val} B$ if and only if $val(A) = val(B)$.

Now suppose an agent is presented with an AF and a value-labelling. In Bench-Capon's model [6], this agent will uphold a proposed attack $A \rightharpoonup B$ and therefore accept that A *defeats* B, unless she strictly prefers the value associated with B to the value associated with the attacker A.

Definition 3 (Defeated Arguments). *Given an AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, we say that argument $A \in Arg$ defeats argument $B \in Arg$, denoted $A \rightrightharpoons B$, if and only if $A \rightharpoonup B$ but not $val(B) \succ val(A)$.*

We call $\rightrightharpoons$ the defeat-relation *induced* by the AVAF. Note that saying '$val(B) \succ val(A)$ is not the case' is the same as saying '$val(A) \succcurlyeq val(B)$ is the case' only when the preference order $\succcurlyeq$ is complete.

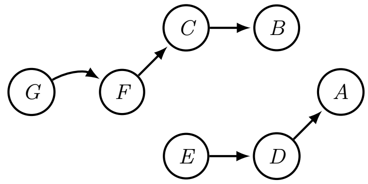
Note that for any given AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$ the induced defeat-relation $\rightrightharpoons$ is, just like an attack-relation $\rightharpoonup$, an irreflexive binary relation on Arg. That is, we can (and will) think of $\langle Arg, \rightrightharpoons \rangle$ as just another AF.

Example 1 (continued). *Recall our earlier example about the arguments pondered by our city council. We can associate the arguments presented in this example with four types of values. Arguments A and G concern environmental responsibility (*`env`*), B and C are about social fairness (*`soc`*), F promotes economic viability (*`econ`*), and D and E pertain to infrastructure efficiency (*`infra`*). We thus have that $Val = \{\texttt{env}, \texttt{soc}, \texttt{econ}, \texttt{infra}\}$, as well as that $val(A) = val(G) = \texttt{env}$, $val(B) = val(C) = \texttt{soc}$, $val(F) = \texttt{econ}$, and $val(D) = val(E) = \texttt{infra}$.*

Let us now assume that a particular councillor wants to promote the values of environmental responsibility and infrastructure efficiency over the other two values. So her preferences might be given by the following weak order:

$$\texttt{env} \sim \texttt{infra} \succ \texttt{soc} \sim \texttt{econ}$$

This induces a defeat-relation $\rightrightharpoons$ for our councillor that corresponds to the following graph:



That is, three attacks have been removed. For this new AF it is unambiguously clear that argument A should be accepted (the only argument attacking A is itself attacked by an argument without any remaining attackers), and thus that diesel cars should be banned from the city centre.

In the sequel, we use standard set-theoretical operations (e.g., $\cap$, $\subseteq$) on binary relations (understood as sets of pairs). Furthermore, $R^{-1} = \{(x, y) \mid yRx\}$ is the inverse of a binary relation, R^+ its transitive closure, and R^* its reflexive-transitive closure. $R \circ R'$ is the composition of R and R'. We also define $R^+_{val} := (R \cup =_{val})^* \circ R \circ (R \cup =_{val})^*$, which is like the usual transitive closure, except that we can move to arguments with the same value, even if not connected by R.

3. THE RATIONALISABILITY PROBLEM

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *agents* (or *audiences*). Suppose each of these agents supplies us with an AF, not necessarily over the same set of arguments.[3] We call this a *profile* of AF's. Then we may ask whether the observed profile can be *rationalised* (explained) in terms of a common master AF and a common value-labelling, together with a profile of preference orders, one for each agent. As we think of each AF in the profile as the result of having imposed the relevant agent's preferences, we write individual AF's as $\langle Arg_i, \rightrightharpoons_i \rangle$ (rather than as $\langle Arg_i, \rightharpoonup_i \rangle$). Here, Arg_i is the set of arguments agent i is *aware* of and $\rightrightharpoons_i$ is the defeat-relation on Arg_i adopted by i. A profile of such AF's is denoted as $\boldsymbol{AF} = (\langle Arg_1, \rightrightharpoons_1 \rangle, \ldots, \langle Arg_n, \rightrightharpoons_n \rangle)$. Let $Arg := Arg_1 \cup \cdots \cup Arg_n$ denote the set of all arguments.

We now define the *rationalisability problem* as the problem of deciding whether a given profile can be rationalised in this sense.[4] In fact, we define an entire family of rationalisability problems, parameterised by a set of *constraints* imposed on the solutions admitted (concrete examples are given below).

Definition 4 (Rationalisability). *A profile of AF's $\boldsymbol{AF} = (\langle Arg_1, \rightrightharpoons_1 \rangle, \ldots, \langle Arg_n, \rightrightharpoons_n \rangle)$ is called rationalisable under a given set of constraints, if there exist an attack-relation $\rightharpoonup$ on $Arg = Arg_1 \cup \cdots \cup Arg_n$, a set of values Val with a mapping $val : Arg \to Val$, and a profile $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$ of preference orders on Val, all meeting said constraints, such that, for all agents $i \in \mathcal{N}$ and all arguments $A, B \in Arg_i$, it is the case that $A \rightrightharpoons_i B$ if and only if $A \rightharpoonup B$ but not $val(B) \succ_i val(A)$.*

We will refer to $\langle Arg, \rightharpoonup \rangle$ as the *master AF*, and consequently to $\rightharpoonup$ as the *master attack-relation*.

In this paper, we will consider the following types of constraints (but others may be of interest as well):

- the master attack-relation $\rightharpoonup$ may be fixed,
- the value-labelling $\langle Val, val \rangle$ may be fixed,
- the number of values may be bounded from above,
- the preference orders may be required to be complete.

[3]A common assumption in the literature on the aggregation of AF's is that every individual agent reports an AF over the exact *same set of arguments* [8, 14, 27]. Here, we instead follow Coste-Marquis et al. [11], who have argued that allowing for differences in the individual sets of arguments is more realistic. Note that the case of a single shared set of arguments is covered by our model as a special case.

[4]A different problem of rationalisability has recently been proposed by Dunne et al. [13]: If you observe a set of subsets of arguments, can it possibly correspond, for a given semantics, to different extensions of *some* single AF?

With these definitions in place, we may now ask: For a given set of constraints, can we characterise the class of all profiles of AF's that can be rationalised? And can we check efficiently whether a given profile is rationalisable?

4. THE SINGLE-AGENT CASE

We first consider the single-agent case of the rationalisability problem. This is not only useful for gaining an understanding of the multiagent case, but is also interesting in its own right. For example, it may be the case that there is some 'ground truth' available and we know what the correct attack-relation is (e.g., due to the logical structure of the arguments), but that a specific agent is still reporting a different AF. Can this subjective AF be explained in terms of the value-based model? That is, is this framework compatible with what we know to be the ground truth?

EXAMPLE 2. *Consider a scenario with three arguments, $Arg = \{A, B, C\}$, with a fixed master attack-relation $\rightharpoonup$ such that $A \rightharpoonup B$, $B \rightharpoonup C$, and $A \rightharpoonup C$. Suppose we observe a single agent who only declares $A \rightleftharpoons B$ and $B \rightleftharpoons C$. Can we rationalise this omission of the attack of A on C? Clearly, rationalisation requires A and C to be labelled with distinct values, say v_A and v_C, and our agent must prefer v_C to v_A for $A \rightharpoonup C$ to get cancelled. Are two values enough? The answer is no: If we reuse, say, value v_A to also label argument B, then $B \rightharpoonup C$ would get cancelled as well. Similarly, if we reuse v_C for B, then $A \rightharpoonup B$ would get cancelled. Thus, we need a third value v_B. Now there is a rationalisation, with the agent's preference order ranking v_C above v_A, and v_B being incomparable to the other two values. Observe that, even with three values, rationalisation is impossible if we require the preference order to be complete, i.e., if we require it to not leave any two values incomparable.*

In the single-agent case, we are given an AF $\langle Arg, \rightleftharpoons \rangle$. A solution consists of an AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$, over the same set of arguments Arg, that induces $\rightleftharpoons$. We consider this problem for several types of constraints on solutions.

FACT 1 (NO CONSTRAINTS). *In the absence of constraints, every single AF is rationalisable.*

PROOF. Given the AF $\langle Arg, \rightleftharpoons \rangle$ to be rationalised, let $(\rightharpoonup) := (\rightleftharpoons)$, choose the value-labelling $\langle Val, val \rangle$ arbitrarily, and let $(\succcurlyeq) := Val \times Val$ (meaning that our agent is indifferent between any two values). Then it is easy to check that $\rightleftharpoons$ is induced by the AVAF $\langle Arg, \rightharpoonup, Val, val, \succcurlyeq \rangle$. $\square$

Our proof shows that the same result also applies to rationalisation under any set of constraints referring only to Val and val. It also continues to apply if we require the preference order to be complete. The main insight here is that any natural instance of the single-agent problem that is nontrivial will involve a constraint on the master attack-relation. Therefore, for the remainder of this section, we only consider rationalisability problems with a given fixed master attack-relation.

PROPOSITION 2 (FIXED ATTACK-RELATION). *A single AF $\langle Arg, \rightleftharpoons \rangle$ is rationalisable by an AVAF with a given fixed master attack-relation $\rightharpoonup$ if and only if all of the following are the case:*

(i) $(\rightleftharpoons) \subseteq (\rightharpoonup)$;
(ii) $(\rightharpoonup \backslash \rightleftharpoons)$ *is acyclic;*
(iii) $(\rightleftharpoons) \cap (\rightharpoonup \backslash \rightleftharpoons)^+ = \varnothing$.

PROOF SKETCH. In this setting, there are no constraints on $\langle Val, val \rangle$. The first important insight then is that having more available values means more flexibility: we can rationalise if and only if we can rationalise by labelling every argument with a distinct value. Thus, we may think of the arguments *themselves* as representing values: w.l.o.g., assume that $Val = Arg$ and that val is the identity function. Hence, we can think of $\succcurlyeq$ as operating directly on arguments and need not consider values any longer.

Condition (i) is required, as our agent can never add (but only remove) edges. Let $R := (\rightharpoonup \backslash \rightleftharpoons)$ denote the set of edges to be removed. We must have $R^{-1} \subseteq (\succ)$ to ensure that the agent's preference order does indeed remove all of these edges. The second important insight now is that it is never beneficial to add more pairs to the preference order than we are absolutely forced to. That is, we should choose $\succ$ as small as possible, namely as the transitive closure of R^{-1}. We then still need to check two things. First, we need to check that $(R^{-1})^+$ is the strict part of some preorder, i.e., that it is transitive and irreflexive. This is equivalent to condition (ii), to R being acyclic. Second, we need to check that we are not removing any edges that should in fact stay, i.e., we need to make sure that $(\rightleftharpoons) \cap R^+ = \varnothing$, which is condition (iii). $\square$

All three conditions can be checked in polynomial time, so we obtain a tractability result:

COROLLARY 3 (FIXED ATTACK-RELATION). *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation can be decided in polynomial time.*

Note that our proof of Proposition 2 shows that requiring the preference order to be strict (i.e., not allowing any indifferences) does not affect rationalisability. On the other hand, our proof does not apply in case the preference order is required to be complete (this case will instead be covered by Proposition 6 below).

As discussed, a crucial ingredient of Proposition 2 and its proof was the fact that there were no constraints on the value-labelling. We now investigate what happens when we add such constraints, and first consider the most extreme case where the full value-labelling is fixed from the outset. This is a natural scenario to consider in those cases in which we are willing to assume that the question of which value a given argument relates to is a matter that can be settled in an objective manner.

PROPOSITION 4 (FIXED VALUE-LABELLING). *A single AF $\langle Arg, \rightleftharpoons \rangle$ is rationalisable by an AVAF with a given fixed master attack-relation $\rightharpoonup$ and a given fixed value-labelling $\langle Val, val \rangle$ if and only if all of the following are the case:*

(i) $(\rightleftharpoons) \subseteq (\rightharpoonup)$;
(ii) *the relation* $\bigcup_{A(\rightharpoonup \backslash \rightleftharpoons)B} \{(val(A), val(B))\}$ *is acyclic;*
(iii) $(\rightleftharpoons) \cap (\rightharpoonup \backslash \rightleftharpoons)^+_{val} = \varnothing$.

PROOF SKETCH. As for Proposition 2, condition (i) reflects that our agent cannot add new edges. The crucial difference to the scenario of Proposition 2 is that now we cannot remove edges between arguments that are labelled with the same value. Let $R := (\rightharpoonup \backslash \rightleftharpoons)$ be the set of edges we need to remove. At the level of the values, this induces the relation $\bigcup_{(A,B)\in R}\{(val(A), val(B))\}$ mentioned in condition (*ii*). As before, the best we can do is to choose as small a preference order as possible, so we should use the transitive closure of the inverse of that relation on values. Condition (*ii*) then amounts to checking that this is indeed a well-formed preference order. Note that acyclicity implies irreflexivity, so we are correctly checking that we are not trying to remove an edge between two arguments labelled with the same value. Finally, we need to check that we are not removing any edges that should stay. This is taken care of by condition (*iii*). To see this, note that R^+_{val} is the set of edges getting removed. $\square$

Also this characterisation immediately provides us with a polynomial algorithm. Thus, we obtain the following result.

COROLLARY 5 (FIXED VALUE-LABELLING). *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation and a given fixed value-labelling can be decided in polynomial time.*

The final single-agent scenario we want to consider here is one where we are not given the full value-labelling but merely an upper bound on the number of values that may be used for rationalisation.[5] This scenario comes about when there is no unique objective mapping from arguments to values and we are looking for a "simple" explanation for an observed defeat-relation only involving a limited number of different values. From an algorithmic point of view, this is the most demanding problem considered so far. Still, at least for the case of complete preferences, also for this problem we are able to establish the existence of a polynomial algorithm, as the following result shows.

PROPOSITION 6 (BOUND ON VALUES). *Whether a single AF is rationalisable by an AVAF with a given fixed master attack-relation, a given upper bound on the number of values, and a complete preference order can be decided in polynomial time.*

PROOF. We are going to show how to translate our problem into an integer program with at most two variables per inequality. Deciding feasibility of such programs is known to be polynomial [18].

Let $\langle Arg, \rightleftharpoons\rangle$ be the AF, $\rightharpoonup$ the master attack-relation, and k (with $k \leqslant |Arg|$) the upper bound on the number of values. Observe that, if rationalisation is possible with fewer than k values, then it certainly is possible with exactly k values. As the rationalising preference order is required to be complete, w.l.o.g., we may assume that $Val = \{1, \ldots, k\}$ and that $\succcurlyeq$ is the usual relation $\geqslant$ defined over the natural numbers. Clearly, if $(\rightleftharpoons) \nsubseteq (\rightharpoonup)$, then rationalisation is impossible. So, from now on, assume that $(\rightleftharpoons) \subseteq (\rightharpoonup)$.

For every argument $A \in Arg$, introduce an integer variable x_A. We use inequalities of the form $1 \leqslant x_A$ and $x_A \leqslant k$ to ensure that each such variable must take a value from *Val*. Thus, these variables encode *val*. We have to be able to model two types of constraints. First, if $A \rightharpoonup B$ but not $A \rightleftharpoons B$, then we must ensure that the value of B is strictly preferred to the value of A: $x_A + 1 \leqslant x_B$. Second, if $A \rightleftharpoons B$ (and thus, by our assumption, also $A \rightharpoonup B$), then we must ensure that the value of B is *not* strictly preferred to the value of A: because of completeness, this can be written as $x_B \leqslant x_A$. The integer program thus constructed is feasible if and only if rationalisation is possible. $\square$

Let us reiterate that our proof makes use of the condition that the rationalising preference order should be complete. Without it, we would not be able to map requirements of the form $val(B) \not\succ val(A)$ into linear constraints. Assuming completeness of the preference order (i.e., excluding the possibility of an agent not being able to compare the importance of two given values) is sometimes reasonable, but certainly not always. Whether single-agent rationalisability for a bounded number of values remains polynomial for possibly incomplete preferences is an open question.

5. THE MULTIAGENT CASE

We now turn to the multiagent case. In presenting our results for each type of constraint considered, we will specifically focus on the extent to which the (positive) results obtained for the single-agent case carry over to this more general scenario. To get started, recall that we have seen that in the absence of constraints, *every* single AF can be rationalised (Fact 1). The following example shows that this result does not generalise to profiles with (at least) two AF's.

EXAMPLE 3. *Consider a profile of two AF's over a common set of three arguments. Suppose $A \rightleftharpoons_1 B$, $B \rightleftharpoons_1 C$, and $C \rightleftharpoons_1 A$, while $(\rightleftharpoons_2) = \varnothing$. Any value-labelled AF and preference profile that could possibly rationalise this profile would have to have an attack-relation $\rightharpoonup$ that includes, at least, the attacks $A \rightharpoonup B$, $B \rightharpoonup C$, and $C \rightharpoonup A$, as otherwise these edges could not have occurred in the first AF. But this means that the second preference order, to be able to cancel these attacks, must at least include the comparisons $val(B) \succ_2 val(A)$, $val(C) \succ_2 val(B)$, and $val(A) \succ_2 val(C)$. But then $\succ_2$ is not acyclic. Thus, this profile cannot be rationalised, even in the absence of any kind of constraint.*

Under what circumstances can we decompose a given multiagent rationalisability problem into a set of n single-agent rationalisability problems that can be solved independently of each other? For the scenarios covered by Propositions 2 and 4 this is easily seen to be possible:

- If the only constraint is that the master attack-relation is fixed, then every agent's rationalisability problem can be solved independently.
- If the only constraints are that master attack-relation and value-labelling are fixed, then every agent's rationalisability problem can also be solved independently.

But what if the master attack-relation is not given? Consider profile $\boldsymbol{AF} = (\langle Arg_1, \rightleftharpoons_1\rangle, \ldots, \langle Arg_n, \rightleftharpoons_n\rangle)$. Any rationalisation of $\boldsymbol{AF}$ must involve a master attack-relation $\rightharpoonup$ with $(\rightharpoonup) \supseteq (\rightleftharpoons_1) \cup \cdots \cup (\rightleftharpoons_n)$, because no agent can create an edge not already included in $\rightharpoonup$. Any additional edges in $\rightharpoonup$ will make rationalisation only harder, if they make a difference at all. Thus, rationalisation is possible at all if and only if rationalisation is possible with the fixed master attack-relation $(\rightharpoonup) := (\rightleftharpoons_1) \cup \cdots \cup (\rightleftharpoons_n)$.

Given these insights, together with Corollaries 3 and 5, we obtain the following result:

[5]Thus, this scenario requires solving the decision problem corresponding to the optimisation problem of computing the minimal number of values needed for rationalisation.

PROPOSITION 7 (DECOMPOSABLE CASES). *Whether a profile of AF's is rationalisable can be decided in polynomial time by solving the problem independently for each agent, in at least the following cases:*

(a) *No constraints are given.*

(b) *Only the master attack-relation is fixed.*

(c) *Only the value-labelling is fixed.*

(d) *Master attack-relation and value-labelling are fixed.*

Thus, of all the constraints we have considered here, only the one specifying an upper bound on the number of values actually leads to a "genuine" multiagent rationalisation problem. Let us now consider this problem in some detail.

For the remainder of the paper, we will always assume that a fixed master attack-relation $\rightharpoonup$ is part of the constraints considered. By our reasoning above, any tractability result obtained under this assumption immediately extends to the case where no master attack-relation is specified.

Our first result on multiagent rationalisation with a bound on the number of values to be used is negative: In the most general case this problem is intractable.

PROPOSITION 8 (GENERAL CASE). *Deciding whether a profile of AF's is rationalisable by an AVAF with a given fixed master attack-relation and a given upper bound (of at least 3) on the number of values is an NP-complete problem.*

PROOF. NP-membership is immediate. To prove NP-hardness we provide a reduction from GRAPH COLOURING, which is known to be NP-hard [20]. Recall that in GRAPH COLOURING we are given an undirected graph $G = (V, E)$ and ask whether it is possible to colour the vertices V using at most $k \geqslant 3$ colours such that no two vertices with the same colour are linked by an edge in E.

So take any instance of GRAPH COLOURING with graph $G = (V, E)$ and bound k. Let $m := |V|$. We build an instance of our rationalisation problem for m arguments, $n := \binom{m}{2}$ agents, and a bound of k on the number of values as follows. First, let $Arg := V$ be the full set of arguments, and let the master attack-relation $\rightharpoonup$ be an arbitrary orientation of G. Second, for every pair $A \neq B \in Arg$ we create exactly one agent i, with $Arg_i = \{A, B\}$ and an empty defeat-relation $(\rightleftharpoons_i) = \varnothing$. (That is, there indeed are $\binom{m}{2}$ agents.) Now consider any edge (A, B) in G. As either $A \rightharpoonup B$ or $B \rightharpoonup A$, but neither $A \rightleftharpoons_i B$ nor $B \rightleftharpoons_i A$, the corresponding agent i must strictly rank $val(A)$ and $val(B)$, i.e., they must be different. As this is so for all edges in G and all agents, any two arguments linked in G must get labelled with distinct values. Hence, G is k-colourable if and only if the profile of AF's we constructed can be rationalised using at most k values. $\square$

This is bad news. But are there special cases where rationalisability is tractable after all? Observe that our proof heavily relied on the fact that different agents may be aware of different sets of arguments. This often is a reasonable assumption [11], but the special case where all agents consider the exact same set of arguments certainly is also of interest. Whether rationalisability for a given bound on the number of values remains intractable for this domain restriction is an open question. Furthermore, note that GRAPH COLOURING is *not* NP-hard for $k = 2$ colours, so our proof of intractability does not cover the case of exactly two values. Whether Proposition 8 can be strengthened to a bound of 2 is yet another interesting open question.

Recall that in case there is no bound on the number of values (or, equivalently, if $k = |Arg|$), we already know that rationalisation is tractable (as this follows from Proposition 7). Our final result shows that the problem remains tractable when the bound k is "large"—in the sense of only reducing the number of allowed values by a constant d (relative to the maximum $k = |Arg|$).

PROPOSITION 9 (LARGE BOUND ON VALUES). *Let $d \in \mathbb{N}$ be an arbitrary constant. Whether $\boldsymbol{AF} = (\langle Arg_1, \rightleftharpoons_1\rangle, \ldots, \langle Arg_n, \rightleftharpoons_n\rangle)$ is rationalisable by an AVAF with a given fixed master attack-relation and at most $k := |Arg_1 \cup \cdots \cup Arg_n| - d$ values can be decided in polynomial time.*

PROOF. Let $m := |Arg_1 \cup \cdots \cup Arg_n|$. There are $p := \binom{\binom{m}{2}}{d}$ ways of selecting d pairs from amongst all pairs of distinct arguments. This number is exponential only in d (not in m). Thus, as d is constant, p is polynomial. Note that p is a (generous) upper bound on the number of ways we can divide the m arguments into $k = m - d$ clusters: For any desired division into k clusters, there exists a choice of d pairs such that we obtain that clustering by merging exactly those pairs.

Note that it is not important *which* value is used to label a given argument: if rationalisation is possible at all, it remains possible after any given permutation of the values. The class of all clusterings with k clusters thus represents all relevant value-labellings with k values. Also note that, if rationalisation is possible with fewer than k values, then it certainly is possible with exactly k values. So we only need to check labellings with exactly k values.

To summarise, we have shown that our original rationalisation problem can be reduced to polynomially many (namely, p) new rationalisation problems, each for the same fixed master attack-relation and its own fixed value-labelling. But each of these individual problems is polynomial by Proposition 7 (item d), so we are done. $\square$

6. APPLICATION SCENARIOS

There are a number of different application scenarios where dealing with questions of rationalisability will be valuable. In this section, we list and illustrate some of them.

First, given the growing interest in the abstract argumentation research community in questions of aggregation of AF's [8, 11, 14, 27], it is important to have a clear understanding for what types of scenarios the question of aggregation is in fact relevant. Our notion of rationalisability provides a suitable definition for this purpose. It allows for a systematic scan of the different examples used in the literature—not to dismiss those failing the test, but to point out that one must be careful with the interpretation used. For instance, let us see whether the example given by Coste-Marquis et al. [11, Example 7] passes the test. We are given $AF_1 = \langle\{A, B, E, F\}, \{(A, B), (B, A), (E, F)\}\rangle$, $AF_2 = \langle\{B, C, D, E, F\}, \{(B, C), (C, D), (F, E)\}\rangle$, and $AF_3 = \langle\{E, F\}, \{(E, F)\}\rangle$. It indeed does pass the test. This profile is rationalisable using as master attack-relation the union of the individual relations. But how many values are required to rationalise it? We see that it is sufficient to set $val(E) \neq val(F)$, while A, B, C, D can take the same value, either that of E or that of F. Thus, two values suffice.

Second, in applications where multiple AF's need to be aggregated, we may use the notion of rationalisability to choose between alternative aggregation techniques, depending on the result of the rationalisability test. For example, if a profile turns out to be rationalisable for a given preference model (e.g., for complete preference orders), we may reasonably assume that this model is a good abstraction of reality and aggregate the AF's by aggregating the inferred preferences (which is a much better studied problem than that of aggregating AF's). For instance, we may use the well-known Kemeny rule to aggregate the preferences, and then apply the collective preference order obtained to the master attack-relation inferred. But when rationalisation fails, this approach does not make sense, and we should look for a different method of aggregation. In that case, there is a more substantial disagreement: maybe the model of preferences has to be changed, maybe the agents differ on the assignment of values to arguments, or maybe the agents interpret the arguments differently. Importantly, failure of rationalisation can also provide hints as to where disagreement occurs.

Third, value-based argumentation systems are used in practice as a modelling tool for online debating platforms [23]. In this context, AF's are (typically) not obtained via a one-shot process, but rather retrieved interactively. Our approach could be used to detect inconsistencies as they occur, and thus to trigger clarification questions on the fly. Suppose, for instance, the following sequence occurs:

- Agent 1: A defeats B.
- Agent 2: B defeats A.
- Agent 3: There is no defeat between A and B.

At this stage it is clear that this collection of AF's cannot be rationalised. A clarification is required to identify the mismatch. For example, the system could ask agent 3 whether she really believes there is no attack between A and B.

Finally, it is interesting to note that our methodology can also be fruitfully combined with other approaches. Specifically, in many contexts, the input information provided is not directly an AF, but rather a set of acceptable arguments (i.e., an extension). This is the case, in particular, when the objective is to analyse *a posteriori* whether a given decision can be explained. A recent example of this kind is the study of a participatory decision setting involving an environmental project in Québec reported on by Tremblay and Abi-Zeid [28]. In their case analysis, they extracted seven values and attached them to arguments. They then enumerated all possible AVAF's, getting an overwhelming number of such frameworks, to test whether and how often the decision recommended by a given framework coincides with the decision actually observed in practice. Interestingly, by combining our technique and an approach for inferring AF's from target extensions [5, 13], a different methodology could be used instead: For a set of observed acceptable arguments, we may first apply such a technique to obtain candidate rationalisable AF's, and then apply our rationalisation method to check whether the AF is rationalisable for some values, against the ground truth built by extraction.

Suppose, for instance, that, regarding three arguments $\{A, B, C\}$ and using one of Dung's semantics, agent 1 reports extension $\{A\}$, agent 2 reports $\{A, C\}$, and agent 3 reports $\{A, B\}$. Realising these extensions—starting from an empty AF and considering a single defeat relation amongst arguments for simplicity—it must be the case that $A \rightharpoonup_1 B$ and $A \rightharpoonup_1 C$, while the relation between B and C is either (1-*i*) $B \rightharpoonup_1 C$, (1-*ii*) $C \rightharpoonup_1 B$, or (1-*iii*) no defeat between B and C. Now, for agent 2, there must be no defeat between A and C, leaving five possible cases: (2-*i*) $A \rightharpoonup_2 B$ and $C \rightharpoonup_2 B$, (2-*ii*) $A \rightharpoonup_2 B$ and $B \rightharpoonup_2 C$, (2-*iii*) $C \rightharpoonup_2 B$ and $B \rightharpoonup_2 A$, (2-*iv*) $A \rightharpoonup_2 B$, or (2-*v*) $C \rightharpoonup_2 B$. Finally, for agent 3, we have no defeat relation between A and B, and thus either (3-*i*) $A \rightharpoonup_3 C$ and $B \rightharpoonup_3 C$, (3-*ii*) $B \rightharpoonup_3 C$ and $C \rightharpoonup_3 A$, (3-*iii*) $A \rightharpoonup_3 C$ and $C \rightharpoonup_3 B$, (3-*iv*) $A \rightharpoonup_3 C$, or (3-*v*) $B \rightharpoonup_3 C$. But now, in terms of rationalisation, we see that some combinations are impossible, such as for instance (1-*iii*, 2-*ii*, 3-*iii*). To see this, note that the master attack-relation would have to contain both $B \rightharpoonup C$ (for agent 2) and $C \rightharpoonup B$ (for agent 3). But then agent 1 would have to have one of these attack relations in her system, as she cannot both strictly prefer the value of B to that of C and *vice versa*. While this does not allow us to uniquely define a collection of AF's, this method can nevertheless guide the search for AF's compatible with the extensions observed.

7. CONCLUSION

We have introduced the concept of *rationalisability* of a profile of abstract argumentation frameworks, proposed a specific instantiation of the general idea in terms of social values associated with the arguments and preferences over those values held by the agents, and studied the resulting decision problem from an algorithmic point of view, for several types of constraints on admissible solutions. We have been able to show that the single-agent rationalisability problem is tractable for all the constraints considered. These positive results extend to the multiagent case for several types of constraints. However, in the presence of a constraint limiting the number of values we may use, the most general variant of the multiagent problem is NP-complete.[6]

While our technical results offer a good initial overview of the landscape of rationalisability, our work also pinpoints a number of interesting open questions. These include the complexity of single-agent rationalisability with a limited number of values for incomplete preferences, as well as the identification of further tractable cases of the multiagent rationalisability problem with a limited number of values.

Besides addressing these questions, future work should also investigate alternative instantiations of the general idea of rationalisability expounded here. For instance, as mentioned already in the introduction, the model of Bench-Capon [6] is but one approach to modelling the emergence of different individual argumentation frameworks. Defining the rationalisability problem for competing approaches is likely to be fruitful as well. Another idea, still within the value-based framework, is to treat the fact that agents may be aware of different sets of arguments somewhat differently. In this paper, we have projected the master attack-relation onto each individual argument set before rationalisation. Alternatively, one could ask whether there exists a *possible completion* of an agent's individual defeat-relation for the full set of arguments induced by her preferences.

Acknowledgments. This work has been partly supported by COST Action IC1205 on Computational Social Choice as well as project AMANDE ANR-13-BS02-0004 of the French National Research Agency.

[6]Recall that we have assumed attack relations to be irreflexive. The complexity of the rationalisability problem is not affected by this assumption. As no assignment of values and choice of preference orders can ever cancel out a self-attack, all you need to do on top of checking our existing conditions is checking that all agents agree on all self-attacks.

REFERENCES

[1] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34:197–216, 2002.

[2] L. Amgoud, Y. Dimopoulos, and P. Moraitis. Making decisions through preference-based argumentation. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR-2008)*, 2008.

[3] L. Amgoud and S. Vesic. A new approach for preference-based argumentation frameworks. *Annals of Mathematics and Artifcial Intelligence*, 63(2):149–183, 2011.

[4] R. Baumann. What does it take to enforce an argument? Minimal change in abstract argumentation. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012)*. IOS Press, 2012.

[5] R. Baumann and G. Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. In *Proceedings of the 3rd International Conference on Computational Models of Argument (COMMA-2010)*. IOS Press, 2010.

[6] T. J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[7] E. Black and A. Hunter. A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55–78, 2012.

[8] G. A. Bodanza and M. R. Auday. Social argument justification: Some mechanisms and conditions for their coincidence. In *Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2009)*. Springer-Verlag, 2009.

[9] R. Booth, S. Kaci, and T. Rienstra. Property-based preferences in abstract argumentation. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT-2013)*. Springer-Verlag, 2013.

[10] M. Caminada and G. Pigozzi. On judgment aggregation in abstract argumentation. *Journal of Autonomous Agents and Multiagent Systems*, 22(1):64–102, 2011.

[11] S. Coste-Marquis, C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex, and P. Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171(10–15):730–753, 2007.

[12] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[13] P. E. Dunne, W. Dvorák, T. Linsbichler, and S. Woltran. Characteristics of multiple viewpoints in abstract argumentation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR-2014)*, 2014.

[14] P. E. Dunne, P. Marquis, and M. Wooldridge. Argument aggregation: Basic axioms and complexity results. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA-2012)*. IOS Press, 2012.

[15] U. Endriss and U. Grandi. Collective rationality in graph aggregation. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI-2014)*. IOS Press, 2014.

[16] S. Gabbriellini and P. Torroni. Arguments in social networks. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2013)*. IFAAMAS, 2013.

[17] D. Grossi and W. van der Hoek. Audience-based uncertainty in abstract argument games. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI-2013)*, 2013.

[18] D. S. Hochbaum and J. Naor. Simple and fast algorithms for linear and integer programs with two variables per inequality. *SIAM Journal on Computing*, 23(6):1179–1192, 1994.

[19] S. Kaci and L. van der Torre. Preference-based argumentation: Arguments supporting multiple values. *International Journal of Approximate Reasoning*, 48(3):730–751, 2008.

[20] R. M. Karp. Reducibility among combinatorial problems. In *Proceedings of a Symposium on the Complexity of Computer Computations*. Plenum Press, 1972.

[21] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9):901–934, 2009.

[22] S. Modgil. Revisiting abstract argumentation frameworks. In *Proceedings of the 2nd International Workshop on Theory and Applications of Formal Argumentation (TAFA-2013)*. Springer-Verlag, 2014.

[23] S. Pulfrey-Taylor, E. Henthorn, K. Atkinson, A. Wyner, and T. J. M. Bench-Capon. Populating an online consultation tool. In *Proceedings of the 24th Annual Conference on Legal Knowledge and Information Systems (JURIX-2011)*. IOS Press, 2011.

[24] I. Rahwan and F. A. Tohmé. Collective argument evaluation as judgement aggregation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*. IFAAMAS, 2010.

[25] F. S. Roberts. *Measurement Theory*. Addison-Wesley, 1979.

[26] J. R. Searle. *Rationality in Action*. MIT Press, 2001.

[27] F. A. Tohmé, G. A. Bodanza, and G. R. Simari. Aggregation of attack relations: A social-choice theoretical analysis of defeasibility criteria. In *Proceedings of the 5th International Symposium on Foundations of Information and Knowledge Systems (FoIKS-2008)*. Springer-Verlag, 2008.

[28] J. Tremblay and I. Abi-Zeid. Value-based argumentation for policy decision analysis: Methodology and an exploratory case study of a hydroelectric project in Québec. *Annals of Operations Research*, 236(1):233–253, 2016.