Rationalisation of Profiles of Abstract Argumentation Frameworks: Extended Abstract*

Stéphane Airiau,[1] **Elise Bonzon,**[2] **Ulle Endriss,**[3] **Nicolas Maudet,**[4] **Julien Rossit** [2]
[1]LAMSADE, Université Paris-Dauphine
[2]LIPADE, Université Paris Descartes
[3]ILLC, University of Amsterdam
[4]LIP6, UPMC Université Paris 6, Sorbonne Universités
[1]stephane.airiau@dauphine.fr, [2]{elise.bonzon,julien.rossit}@parisdescartes.fr,
[3]ulle.endriss@uva.nl, [4]nicolas.maudet@lip6.fr

Abstract

We review a recently introduced model in which each of a number of agents is endowed with an abstract argumentation framework reflecting her individual views regarding a given set of arguments. A question arising in this context is whether the diversity of views observed in such a situation is consistent with the assumption that every individual argumentation framework is induced by a combination of, first, some basic factual information 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 occur in a multiagent system.

1 Introduction

The model of abstract argumentation introduced by Dung [1995] is at the root of a vast amount of work in AI. In a nutshell, this model abstracts away from the internal structure of an argument and simply represents 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 advanced together. The simplicity and generality of this model, 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.

Starting with the work of Coste-Marquis *et al.* [2007], 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 [Tohmé *et al.*, 2008; Bodanza and Auday, 2009; Dunne *et al.*, 2012; Endriss and Grandi, 2017; Bodanza *et al.*, 2017]. This is a fruitful line of research, bringing together concerns in abstract argumentation with the methodology of social choice theory, 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 may often seem more natural to encounter a profile with similar individual attack-relations rather than one with attack-relations that differ radically. So, 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 well depend on the individual circumstances of the agent considering the arguments in question, is central to the study of argumentation. Here we adopt one specific approach for modelling this phenomenon, the *value-based* approach proposed by Bench-Capon [2003]. 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, then 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? In other words: 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 $\succcurlyeq_i$. We may wish to impose any number of *constraints* on the solutions for this *rationalisability problem* we are interested in: e.g., constraints on the attack-relation of the master argumentation framework, constraints on the number of values used for rationalisation, or constraints on the preference orders admitted.

*This is a high-level summary of a paper originally presented at the 15th International Conference on Autonomous Agents and Multiagent Systems in 2016 [Airiau *et al*., 2016].

2 Examples

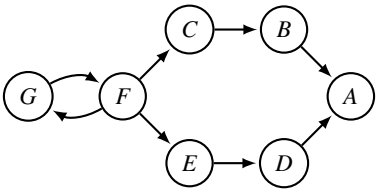
We refer to our original work for a formal definition of our model [Airiau *et al.*, 2016]. In this section, we instead introduce the basic ideas by means of examples.

An *argumentation framework* (AF), in the sense of Dung [1995], is a pair $AF = \langle Arg, \rightharpoonup \rangle$, where *Arg* is a set of arguments and $\rightharpoonup$ is a binary relation on *Arg*. If $A \rightharpoonup B$ holds between two arguments $A, B \in Arg$, we say that A *attacks* B. We shall assume that *Arg* is finite and that $\rightharpoonup$ is irreflexive.

Example 1. *A city council faces the issue of possibly banning polluting vehicles, and specifically diesel cars, from the city centre. The following arguments are under discussion:*

- (A) *Diesel cars should be banned from the inner city centre in order to decrease pollution.*
- (B) *Artisans, who deserve special protection, cannot change their vehicles, as that would be too expensive for them.*
- (C) *The city can offer financial assistance to artisans.*
- (D) *There are only very few alternatives to using diesel cars. Specifically, the 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, together with a natural attack-relation $\rightharpoonup$:



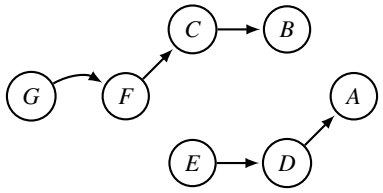
Observe that for this AF it is ambiguous whether or not we should accept argument A and ban diesel cars: Accepting either $\{A, C, E, G\}$ or $\{B, D, F\}$ is intuitively admissible.

Consider now an argumentation framework $\langle Arg, \rightharpoonup \rangle$, a finite set of social or moral *values Val*, a mapping $val : Arg \to Val$, and an agent i with *preference order* $\succcurlyeq_i$. That agent may reject any attack from an argument A to another argument B in case the former is associated with a value of less importance than the value the latter is associated with. Following Bench-Capon [2003], we say that argument $A \in Arg$ *defeats* argument $B \in Arg$, denoted $A \rightharpoonup_i B$, if and only if we have $A \rightharpoonup B$ but it is not the case that $val(B) \succ_i val(A)$.

Example 1 (continued)**.** *Let us introduce four different values. Arguments A and G concern environmental responsibility (value* `env`*), B and C are about social fairness (value* `soc`*), F promotes economic viability (value* `econ`*), and D and E pertain to infrastructure efficiency (value* `infra`*). Suppose a particular councillor i 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:*

$$\texttt{env} \sim_i \texttt{infra} \succ_i \texttt{soc} \sim_i \texttt{econ}$$

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

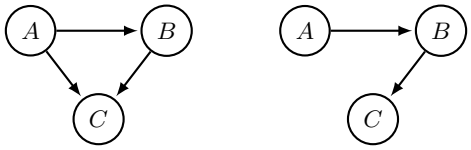


For instance, the attack from B to A got removed, because $val(A) = \texttt{env} \succ_i \texttt{soc} = val(B)$. For the new AF it is unambiguously clear that A should be accepted, and thus that diesel cars should be banned from the city centre.

Note that agent i's defeat-relation $\rightharpoonup_i$ is, just like an attack-relation $\rightharpoonup$, an irreflexive binary relation on *Arg*. Thus, we can (and will) think of $\langle Arg, \rightharpoonup_i \rangle$ as just another AF.

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *agents*. Suppose each agent supplies us with an AF, not necessarily over the same set of arguments. We call this a *profile* of AFs and denote it as $(\langle Arg_1, \rightharpoonup_1 \rangle, \ldots, \langle Arg_n, \rightharpoonup_n \rangle)$. Now, we call such a profile *rationalisable* if there exist a *master attack-relation* $\rightharpoonup$ on $Arg = Arg_1 \cup \cdots \cup Arg_n$, a set of values *Val* with a common mapping $val : Arg \to Val$, and a profile $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$ of individual preference orders on *Val*, such that, for all agents $i \in \mathcal{N}$ and all arguments $A, B \in Arg_i$, it is the case that $A \rightharpoonup_i B$ if and only if $A \rightharpoonup B$ but not $val(B) \succ_i val(A)$. Thus, we say that the profile is rationalisable in case we can explain its commonalities in terms of the master attack-relations and its differences in terms of the agent-specific preferences.

Example 2. *Suppose we observe two agents who are aware of the same set of arguments $\{A, B, C\}$ but who disagree on the status of the possible attacks between them:*



The only disagreement is whether A attacks C (agent 1 says it does; agent 2 says it does not). As preferences can only cancel attacks rather than create new ones, our best chance at rationalisation is to assume that the AF of agent 1 is also the master AF. Then, whatever values we end up associating with the arguments, if we assume that agent 1's preference order is such that she is indifferent between all values, we do not have to cancel any of the attacks for her, and rationalisation works vacuously as far as she is concerned. Thus, our original rationalisability problem reduces to the question of whether we can rationalise the second AF under the constraint that the first AF is the master AF. Now, if we associate each argument with a distinct value and if we permit incomplete preferences, then rationalisation is possible: If agent 2 prefers the value of C to the value of A and does not express a preference between any other pair, then the attack between A and C gets cancelled and all other attacks remain in place.

Rationalisation becomes more interesting—and also more difficult—when we impose *constraints* on the range of possible rationalisations we want to permit. In our work, we have considered the following types of constraints:

- the master attack-relation $\rightharpoonup$ may be fixed,
- the value-labelling $\langle \mathit{Val}, \mathit{val} \rangle$ may be fixed,
- the number of values $|\mathit{Val}|$ may be bounded by some k,
- preferences $\succcurlyeq_i$ may have to be complete (i.e., total).

In addition, we have paid special attention to two *restrictions* of the general problem, namely the case where all individual argument sets coincide (with $\mathit{Arg}_i = \mathit{Arg}_j$ for all $i, j \in \mathcal{N}$) and the single-agent case (with $n = 1$).

Example 3. *Let us again consider the rationalisability problem of Example 2, but now under the constraint that we may make use of at most two distinct values. Such constraints are of great practical interest, given that the number of distinct social values that we should expect an agent to reason about will usually be fairly low. Is rationalisation still possible under this cardinality constraint? No! We need two distinct values for A and C to cancel the attack between them. If B gets the same value as A, then we must also cancel the attack between B and C. And if B gets the same value as C, then we must also cancel the attack between A and B.*

Example 4. *Again for the scenario of Example 2, suppose we are interested in rationalisation with an arbitrary number of values but under the constraint that all preference orders involved must be complete. Is this possible? No! As we have seen before, agent 2 must strictly prefer the value of C to the value of A. Now, if we place the value of B anywhere above the value of A in her preference order, then we must also cancel the attack between A and B. On the other hand, if we place the value of B anywhere below the value of C, then we must also cancel the attack between B and C.*

3 Results

In this section, we give a high-level and informal overview of our main results. For the details, we refer to the original paper [Airiau *et al.*, 2016]. All results deal with the question of how to solve the rationalisability problem under certain constraints and, possibly, assuming certain restrictions. In the best possible scenario, we are able to provide a full characterisation of all problem instances thar are rationalisable, in such a way that the conditions featuring in the characterisation are computationally straightforward to check. The next best thing is a proof of the existence of a polynomial algorithm for solving the rationalisability problem. Most of our results are positive and either of the first or the second kind. But we also have identified scenarios where it is computationally intractable to decide whether a given profile can be rationalised.

First, let us consider the single-agent case. This is not only useful in view of 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?

It is easy to see that, in the absence of constraints, *every* single AF is rationalisable. We can simply assume that the master AF is equal to the AF observed an that our agent is indifferent between all values. Thus, nontrivial single-agent rationalisability problems involve some given master attack-relation. In case that is the only constraint, i.e., if we ask whether $\langle \mathit{Arg}, \rightleftharpoons \rangle$ can be rationalised by means of some given $\rightharpoonup$, we can fully characterise the positive problem instances. The following three conditions must be met:

(i) The observed attack-relation can only include attacks already present in the master attack-relation: $(\rightleftharpoons) \subseteq (\rightharpoonup)$.

(ii) The graph of attacks removed, $(\rightharpoonup \backslash \rightleftharpoons)$, must be acyclic, as it must correspond to the agent's preference order.

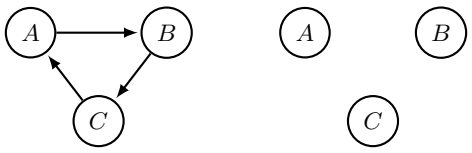
(iii) As preference orders are transitive, the transitive closure of that graph cannot be allowed to intersect with the attacks to be kept: $(\rightleftharpoons) \cap (\rightharpoonup \backslash \rightleftharpoons)^+ = \varnothing$.

These conditions are easy to check, which leads to a simple polynomial algorithm for deciding rationalisability. This approach can be generalised to also deal with problem instances where the value labelling $\langle \mathit{Val}, \mathit{val} \rangle$ is fixed.

Our most interesting result regarding the single-agent case concerns scenarios where the master attack-relation is given, an upper bound k is placed on the number of values to be used for rationalisation, and all preference orders are required to be complete. We have been able to show that this type of rationalisability problem can be solved in polynomial time by reducing it to the feasibility problem for integer programs with at most two variables per inequality [Hochbaum and Naor, 1994]. Whether the problem remains polynomial in case we drop the requirement of preferences being complete is an open question.

Next, let us turn to the multiagent case. We have seen that, in the absence of constraints, *every* single AF can be rationalised. The following example shows that this does not generalise to profiles with (at least) two AFs.

Example 5. *Consider the following profile of two AFs over the common set of three arguments $\{A, B, C\}$:*



To achieve rationalisation, we would have to use a master attack-relation $\rightharpoonup$ that includes, at the very 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, so as to be able to cancel these attacks, must at least include the comparisons $\mathit{val}(B) \succ_2 \mathit{val}(A)$, $\mathit{val}(C) \succ_2 \mathit{val}(B)$, and $\mathit{val}(A) \succ_2 \mathit{val}(C)$. But then the relation $\succcurlyeq_2$ is not acyclic. Thus, this profile cannot be rationalised, even in the absence of any kind of constraint.

Our first result for the multiagent case identifies scenarios for which it is possible to reduce the multiagent rationalisability problem to a series of independent single-agent rationalisability problems. In a nutshell, this is the case when the constraints on rationalisation only concern the master attack-relation and the value-labelling. Any such problem can thus be solved efficiently, given that the single-agent problems involved can be solved efficiently.

Our remaining results concern the multiagent rationalisability problem with an upper bound k on the number of values to be used. The most general form of this problem is NP-complete, i.e., it is computationally intractable to determine rationalisability for such scenarios. The proof involves a reduction from the well-known problem of GRAPH COLOURING with k colours [Karp, 1972]. As GRAPH COLOURING is NP-complete only for $k \geqslant 3$, it remains an open question of whether multiagent rationalisability with $k = 2$ values also is intractable. Furthermore, our proof heavily exploits the fact that each individual agent may be aware of a different set of arguments. Whether intractability persists also under the restriction that all agents report the same set of arguments is yet another interesting open question. Finally, we have been able to show that in case the number of values that may be used is very large—in the sense that the difference between the overall number of arguments reported by the agents and the maximal number of values that can be used for rationalisation is bounded from above by a constant—it is possible to decide multiagent rationalisability in polynomial time.

Together, these results offer a good overview of the landscape of rationalisability. They enable the design of efficient algorithms for deciding whether a given profile of AFs can be rationalised for a given set of constraints for a range of natural scenarios, and they also pinpoint those scenarios were efficient algorithm design will be most challenging.

4 Application Scenarios

There are a number of different application scenarios where dealing with questions of rationalisability will be valuable.

First, given the growing interest in the abstract argumentation research community in questions of aggregation of AFs [Bodanza *et al.*, 2017], 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.

The second application concerns aggregation itself. In a scenario where multiple AFs need to be aggregated, we may use the notion of rationalisability to choose between alternative aggregation techniques. For example, if a profile is rationalisable for a given preference model, we may reasonably assume that this model is a good abstraction of reality and aggregate the AFs by aggregating the inferred preferences (which is a much better studied problem than that of aggregating AFs). But when rationalisation fails, this approach does not make sense, and we should look for a different method of aggregation. In such a case, there is a more substantial disagreement between the agents: the model of preferences may have to be changed, the agents may differ on the assignment of values to arguments, or the agents may interpret the arguments differently. Importantly, failure of rationalisation can also provide hints as to where disagreement occurs.

In the context of online debating platforms, value-based argumentation systems are used as a modelling tool [Pulfrey-Taylor *et al.*, 2011]. On these platforms, AFs are (typically) not obtained via a one-shot process, but rather retrieved interactively, by monitoring the utterances of the participants. Our approach could be used to detect inconsistencies as they occur, and thus to trigger clarification questions on the fly.

Our final point concerns the nature of what is observed. So far we have assumed that the agents express AFs, which we can observe directly. But in many situations, it may be more natural to assume that each agent only reports the set of arguments she accepts, or a (partial) labelling of arguments 'accepted' and 'rejected'. Dunne *et al.* [2014] have addressed the challenging problem of inferring an AF from such an extension (or a set of such extensions) that could serve as an explanation for the behaviour observed. Of course, there often will be *many* possible AFs that could explain a given set of accepted arguments. Our approach could be used to narrow down the range of possible explanations when performing this task for several agents in parallel, by imposing the constraint that the profile of AFs we infer, one for each extension observed, should be rationalisable. Similar ideas may also have useful applications in the context of analysing people's decisions *a posteriori*. For a set of arguments we observe to have been accepted in the course of a debate, we may first induce a number of possible AFs that could explain this extension, using the approach of Dunne *et al.* [2014], and then check whether any of these AFs is rationalisable, given the constraints regarding values extracted from the debate.

5 Future Work

Our work raises several interesting questions that may be addressed in future work. To start with, there are a number of open technical questions regarding the computational complexity of the rationalisability problem for certain combinations of constraints. Most prominent amongst them are (1) the single-agent case with a bound on the number of values and possibly incomplete preferences, (2) the multiagent case with exactly two values, and (3) the multiagent case with a bound on the number of values under the assumption that all agents are aware of the exact same set of arguments.

But future work should also investigate alternative instantiations of the general idea of rationalisability. For instance, the model of Bench-Capon 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. Finally, it is important to keep in mind that Dung's model of abstract argumentation is just that: an *abstract* model of argumentation. Other formalisms, which also model the internal structure of arguments, come closer to real forms of argumentation occurring between people. Therefore, our approach should also be applied to such richer models of argumentation.

Acknowledgements

We are grateful for the feedback received from several anonymous reviewers. Our work was partly supported by COST Action IC1205 on Computational Social Choice and by project AMANDE ANR-13-BS02-0004 of the French National Research Agency.

References

[Airiau *et al.*, 2016] Stéphane Airiau, Elise Bonzon, Ulle Endriss, Nicolas Maudet, and Julien Rossit. Rationalisation of profiles of abstract argumentation frameworks. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS-2016)*. IFAAMAS, 2016.

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

[Bodanza and Auday, 2009] Gustavo A. Bodanza and Marcelo 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.

[Bodanza *et al.*, 2017] Gustavo M. Bodanza, Fernando A. Tohmé, and Marcelo R. Auday. Collective argumentation: A survey of aggregation issues around argumentation frameworks. *Argument & Computation*, 2017. In press.

[Coste-Marquis *et al.*, 2007] Sylvie Coste-Marquis, Caroline Devred, Sébastien Konieczny, Marie-Christine Lagasquie-Schiex, and Pierre Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171(10–15):730–753, 2007.

[Dung, 1995] Phan Minh 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.

[Dunne *et al.*, 2012] Paul E. Dunne, Pierre Marquis, and Michael 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.

[Dunne *et al.*, 2014] Paul E. Dunne, Wolfgang Dvořák, Thomas Linsbichler, and Stefan 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.

[Endriss and Grandi, 2017] Ulle Endriss and Umberto Grandi. Graph aggregation. *Artificial Intelligence*, 245:86–114, 2017.

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

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

[Pulfrey-Taylor *et al.*, 2011] Sarah Pulfrey-Taylor, Emily Henthorn, Katie Atkinson, Adam Wyner, and Trevor 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.

[Tohmé *et al.*, 2008] Fernando A. Tohmé, Gustavo A. Bodanza, and Guillermo 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.