Judgement Aggregation Over Conflicting Arguments: An Extended Abstract

Iyad Rahwan[1,2] and Fernando Tohmé[3]

[1] Masdar Institute of Science & Technology, Abu Dhabi, UAE
[2] Massachusetts Institute of Technology, Cambridge MA, USA
[3] LIDIA, Universidad Nacional del Sur, Bahía Blanca, CONICET, Argentina

Abstract. Consider a knowledge base consisting of a set of arguments and a binary relation characterising conflict among them. There may be multiple plausible ways to evaluate conflicting arguments. In this paper, we ask: *given a set of agents, each with a legitimate subjective evaluation of a set of arguments, how can they reach a collective evaluation of those arguments?* After formally defining this problem, we extensively analyse an argument-wise plurality voting rule, showing that it suffers a fundamental limitation. Then we demonstrate, through a general impossibility result, that this limitation is more fundamentally rooted. Finally, we show how this impossibility result can be circumvented by additional domain restrictions.[4]

1 Introduction

Argumentation has recently become one of the key approaches to automating and analysing reasoning in the presence of conflicting information. A key milestone in the development of argumentation in AI has been Dung's landmark framework [6]. Arguments are viewed as abstract entities, with a binary defeat relation among them (resulting in a so-called *argument graph*).

Often, there are multiple reasonable ways in which an agent may evaluate a given argument graph. Each possible evaluation corresponds to a so-called *extension* [6] or *labelling* [4]. We ask: *Given an argument structure and a set of agents, each with a legitimate subjective evaluation of the given arguments, how can the agents reach a collective compromise on the evaluation of those arguments?*

We formally define the problem of aggregating multiple evaluations of arguments, in the spirit of preference aggregation [1] and judgement aggregation [8, 9]. We define a specific aggregation operator (argument-wise plurality voting) and analyse some of its key properties. We then present an impossibility result on the existence of good aggregation operators (in particular, satisfying collective rationality). Then, we show one way in which the impossibility result can be avoided. In particular, we provide a full characterisation of the space of individual judgements that guarantees collective rationality using argument-wise plurality voting.

[4] A full version of this paper, with detailed proofs, appears in AAMAS 2010 [11]

The paper makes three key contributions to the state-of-the-art in computational models of argument. Firstly, the paper defines and analyses the argument-wise plurality voting mechanism for collective argument evaluation.

Our second contribution is a general impossibility result, showing that there is no aggregation operator that can satisfy a few simple requirements (common in social choice theory) for arbitrary argument graphs. This result not only helps us avoid the fruitless pursuit of such operator, but also because it motivates the need for specialised aggregation operators that work under more restrictive conditions.

This leads to the third contribution. By showing how the impossibility result can be avoided by restricting the space of possible individual judgements, we provide guidance on circumventing the practical implications of the problem.

2 Preliminaries

We briefly outline key elements of abstract argumentation frameworks [6], assuming finite sets of arguments.

Definition 1. *An* argumentation framework *is a pair* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\mathcal{A}$ *is a finite set of arguments and* $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ *is a defeat relation. We say that an argument* α defeats *an argument* β *if* $(\alpha, \beta) \in \rightharpoonup$ *(also written* $\alpha \rightharpoonup \beta$*).*

An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments. An example argument graph is shown in Figure 1. Argument α_1 has two defeaters (*i.e.* counter-arguments) α_2 and α_4, which are themselves defeated by arguments α_3 and α_5 respectively. Let $S^+ = \{\beta \in \mathcal{A} \mid \alpha \rightharpoonup \beta \text{ for some } \alpha \in S\}$. Also let

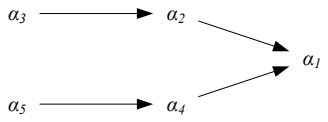

Fig. 1. A simple argument graph

$\alpha^- = \{\beta \in \mathcal{A} \mid \beta \rightharpoonup \alpha\}$. We first characterise the fundamental notions of conflict-free and defence.

Definition 2. *Let* $\langle \mathcal{A}, \rightharpoonup \rangle$ *be an argumentation framework, let* $S \subseteq \mathcal{A}$ *and* $\alpha \in \mathcal{A}$*.*

- S *is* conflict-free *iff* $S \cap S^+ = \emptyset$*.*
- S defends *argument* α *iff* $\alpha^- \subseteq S^+$*. Equivalently, we say that argument* α *is* acceptable with respect to S*.*

Intuitively, a set of arguments is *conflict free* if no argument in that set defeats another. A set of arguments *defends* a given argument if it defeats all its defeaters. In Figure 1, for example, $\{\alpha_3, \alpha_5\}$ defends α_1. We now look at some ways to characterise the *collective acceptability* of a set of arguments.

Definition 3 (Characteristic function). *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. The* characteristic function *of AF is $\mathcal{F}_{AF}$: $2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ such that, given $S \subseteq \mathcal{A}$, we have $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$.*

When there is no ambiguity about the argumentation framework in question, we will use $\mathcal{F}$ instead of $\mathcal{F}_{AF}$.

Definition 4. *Let S be a conflict-free set of arguments in framework $\langle \mathcal{A}, \rightharpoonup \rangle$.*

- *S is* admissible *iff it is conflict-free and defends every element in S (i.e. iff $S \subseteq \mathcal{F}(S)$).*
- *S is a* complete extension *if $S = \mathcal{F}(S)$.*

Intuitively, a set of arguments is *admissible* if it is a conflict-free set that defends itself against any defeater – in other words, if it is a conflict free set in which each argument is acceptable with respect to the set itself.

An admissible set S is a *complete extension* if and only if *all* arguments defended by S are also in S (that is, if S is a fixed point of the operator $\mathcal{F}$). There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint.

Example 1. In Figure 1, the sets $\emptyset$, $\{\alpha_3\}$, $\{\alpha_5\}$, and $\{\alpha_3, \alpha_5\}$ are all admissible simply because they do not have any defeaters. The set $\{\alpha_1, \alpha_3, \alpha_5\}$ is also admissible since it defends itself against both defeaters α_2 and α_4. The admissible set $\{\alpha_1, \alpha_3, \alpha_5\}$ is the only complete extension, since $\mathcal{F}(\{\alpha_1, \alpha_3, \alpha_5\}) = \{\alpha_1, \alpha_3, \alpha_5\}$.

There are various approaches to differentiate between different complete extensions (*e.g.* by defining grounded, preferred, stable extensions and so on [6]). In this paper, we will take a liberal approach and consider any complete extension as a reasonable point of view for an agent, satisfying the minimal criteria of consistency and self-defence.

Crucial to our subsequent analysis is the notion of *argument labelling* [4]. It specifies which arguments are accepted (labelled `in`), which ones are rejected (labelled `out`), and which ones whose acceptance or rejection could not be decided (labelled `undec`). Labellings must satisfy two conditions: (i) an argument is `in` if and only if all of its defeaters are `out`; (ii) an argument is `out` if and only if at least one of its defeaters is `in`.

Definition 5 (Argument Labelling). *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. An* argument labelling *is a total function $L : \mathcal{A} \rightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ such that:*

- $\forall \alpha \in \mathcal{A} \ : \ (L(\alpha) = \texttt{out} \equiv \exists \beta \in \mathcal{A}$ *such that* $(\beta \rightharpoonup \alpha$ *and* $L(\beta) = \texttt{in}))$*; and*
- $\forall \alpha \in \mathcal{A} \ : \ (L(\alpha) = \texttt{in} \equiv \forall \beta \in \mathcal{A} \ : \ (\ $*if* $\beta \rightharpoonup \alpha$ *then* $L(\beta) = \texttt{out}))$

If none of the two conditions is satisfied, then $L(\alpha) = \texttt{undec}$ *(since* L *is a total function).*

Caminada [4] showed a one-to-one correspondence between possible labellings and the set of all complete extensions.

3 Motivation and Scope

In this section, we give a simple example and use it to motivate the paper and highlight the scope of its contributions. Consider the following simple example.

Example 2 (A Murder Case). A murder case is under investigation. To start with, there is an argument that the suspect should be presumed innocent (α_3). However, there is evidence that he may have been at the crime scene at the time (α_2), which would counter the initial presumption of innocence. There is also, however, evidence that the suspect was attending a party that day (α_1). Clearly, α_1 and α_2 are mutually defeating arguments since the suspect can only be in one place at any given time. This problem can be modelled as an argumentation framework $AF = \langle\{\alpha_1, \alpha_2, \alpha_3\}, \rightharpoonup\rangle$ with $\rightharpoonup = \{(\alpha_1, \alpha_2), (\alpha_2, \alpha_1), (\alpha_2, \alpha_3)\}$. Possible labellings are:

- $L(\alpha_1) = \texttt{in}$, $L(\alpha_2) = \texttt{out}$, $L(\alpha_3) = \texttt{in}$.
- $L'(\alpha_1) = \texttt{out}$, $L'(\alpha_2) = \texttt{in}$, $L'(\alpha_3) = \texttt{out}$.
- $L''(\alpha_1) = \texttt{undec}$, $L''(\alpha_2) = \texttt{undec}$, $L''(\alpha_3) = \texttt{undec}$.

The graph and possible labellings are depicted in Figure 2.

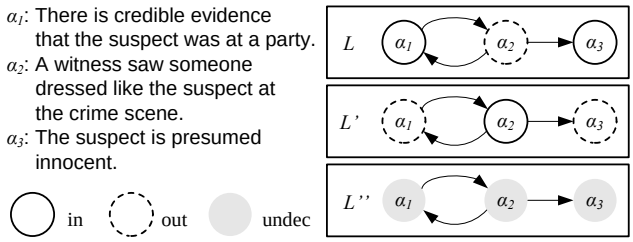

Fig. 2. Graph with three possible labellings

Example 2 highlights a situation in which multiple points of view can be taken, depending on whether one decides to accept the argument that the suspect was at the party or the crime scene. Consider the following example.

Example 3 (Three Detectives). A team of three detectives, named 1, 2, and 3, have been assigned to the murder case described in Example 2. Each detective's judgement can only correspond to a legal labelling (otherwise, his/her judgement

is not admissible and can be discarded). Suppose that each detective's judgement is such that $L_1 = L$, $L_2 = L'$ and $L_3 = L'$. That is, detectives 2 and 3 agree but differ with detective 1. These labellings are depicted in the labelled graph of Figure 3. The detectives must decide which (aggregated) argument labelling best reflects their collective judgement.

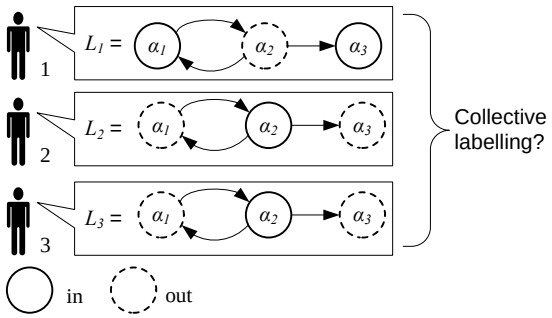

Fig. 3. Detectives with different judgements

Example 3 highlights an aggregation problem, similar to the problems of preference aggregation [1] and judgement aggregation [8]. It is perhaps obvious in this particular example that α_3 must be rejected (and thus the defendant be considered guilty), since most detectives seem to think so. For the same reason, α_1 must be rejected and α_2 must be accepted. Thus, labelling L' (see Example 2) wins by majority. As we shall see in our analysis below, things are not that simple, and counter-intuitive situations may arise. To summarise, the question is as follows: *Given a set of agents, each with a specific subjective labelling of a given set of conflicting arguments, how can agents reach a collective decision on how to evaluate those arguments?*

Below, we will explore the above question deeply. We introduce the argument-wise plurality voting rule and study its key properties. We show that while argument-wise plurality voting satisfies many desirable properties (*e.g.* anonimity, strategy-proofness *etc.*), it can produce counter-intuitive results. We then generalise this observation by presenting a general impossibility result on the existence of collectively rational aggregation operators for argument labelling. We then fully characterise restrictions on the space of individual judgements under which the argument-wise plurality voting avoids the impossibility result.

4 Aggregation of Labellings

The problem we face is that of *judgement aggregation* [8] in the context of argumentation frameworks. In particular, taking as an input a set of *individual* judgements as to how each argument in AF must be labelled, we need to come

up with a *collective* judgement. If each agent $i = 1, \ldots, n$ has a labelling L_i, we need to find an *aggregation operator*, which we define as a partial function[5] $F : \mathbf{L}(AF)^n \rightarrow \{\texttt{in}, \texttt{out}, \texttt{undec}\}^{\mathcal{A}}$, where $\mathbf{L}(AF)$ is the class of labellings of AF. This means that for each $\alpha \in \mathcal{A}$, $F(L_1, \ldots, L_n)[\alpha]$ is the label assigned to α (if F is defined for α).

Aggregation involves comparing and assessing different points of view. There are, of course, many ways of doing this, as extensively discussed in the literature of Social Choice Theory [7]. In this literature, a consensus on some normative ideals has been reached, identifying what a 'fair' way of adding up preferences should be. So for instance, if everybody agrees, the outcome must reflect that agreement; no single agent can impose her view on the aggregate; the aggregation should be performed in the same way in each possible case, *etc.*These informal requirements can be formally stated as properties that F should satisfy [8, 5]:

Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, and suppose we have n agents.

1. *Universal Domain*: Every possible profile of labellings $(L_1, \ldots, L_n)$ is in the domain of F.
2. *Unanimity*: If $L_i = L$ for $i = 1, \ldots, n$, then $F(L_1, \ldots, L_n) = L$.
3. *Anonymity*: given any permutation $p : \{1, \ldots, n\} \rightarrow \{1, \ldots, n\}$, $F(L_1, \ldots, L_i, \ldots, L_n) = F(L_{p(1)}, \ldots, L_{p(i)}, \ldots, L_{p(n)})$.
4. *Independence*: for any $\alpha \in \mathcal{A}$, and any profiles $(L_1, \ldots, L_n)$ and $(L'_1, \ldots, L'_n)$, if $\forall i$ we have: $L_i(\alpha) = L'_i(\alpha)$, then $F(L_1, \ldots, L_n)[\alpha] F(L'_1, \ldots, L'_n)[\alpha]$.
5. *Neutrality*: for any pair $\alpha, \beta \in \mathcal{A}$, and any profile $(L_1, \ldots, L_n)$ if $\forall i$ $L_i(\alpha) = L_i(\beta)$ then $F(L_1, \ldots, L_n)[\alpha] = F(L_1, \ldots, L_n)[\beta]$
6. *Systematicity*: for any $\alpha, \beta \in \mathcal{A}$ and any profiles $(L_1, \ldots, L_n)$ and $(L'_1, \ldots, L'_n)$, if $\forall i$, $L_i(\alpha) = L'_i(\beta)$ then $F(L_1, \ldots, L_n)[\alpha] = F(L'_1, \ldots, L'_n)[\beta]$.
7. *Monotonicity*: For any $\alpha \in \mathcal{A}$, $l_\alpha \in \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ is such that given two profiles $(L_1, \ldots, L_i, \ldots, L_n)$ and $(L_1, \ldots, L'_i, \ldots, L_n)$ (differing only in i's labelling), if $L_i(\alpha) \neq l_\alpha$ while $L'_i(\alpha) = l_\alpha$, $F(L_1, \ldots, L_n)[\alpha] = l_\alpha$ implies that $F(L_1, \ldots, L'_i, \ldots, L_n)[\alpha] = l_\alpha$.
8. *Non-dictatorship*: there is no i such that for any profile $(L_1, \ldots, L_i, \ldots, L_n)$, $F(L_1, \ldots, L_i, \ldots, L_n) = L_i$.
9. *Collective Rationality*: $F(L_1, \ldots, L_n)$, is a labelling.

In words, *universal domain* requires that F admits any logically possible profile of agent judgements. *Unanimity* requires that if all agents submit the same labelling, this labelling must be the collective one. *Anonimity* means that all agents should have equal weight in the aggregation. *Independence* means that

[5] We state that the function is partial to allow for cases in which collective judgement may be undefined (e.g. when there is a tie in voting).

collective judgement on each argument should only depend on individual judgements about that particular argument. *Systematicity* combines independence with neutrality across arguments. *Monotonicity* states that if an agent switches its judgement on an argument in favour of the collective judgement, then the collective judgement remains the same. *Non-dictatorship* means that no single agent should always determine the collective judgement. *Collective rationality* means that the aggregation is always a legitimate labelling.

Notice that these conditions are not independent since, for instance, *Systematicity* implies *Independence* (just by choosing $\alpha \equiv \beta$), but they reflect many properties that researchers consider a 'good' aggregation operator should have. In fact, it is trivial to show that Neutrality and Independence imply Systematicity. Suppose that for any pair $\alpha, \beta \in \mathcal{A}$ and any two profiles $(L_1, \ldots, L_n)$ and $(L_1^{'}, \ldots, L_n^{'})$ such that $\forall i\ L_i(\beta) = L_i(\beta) = L_i^{'}(\beta)$ we have, by Independence, that $F(L_1, \ldots, L_n)[\beta] = F(L_1^{'}, \ldots, L_n^{'})[\beta]$ and by Neutrality, $F(L_1, \ldots, L_n)[\alpha] = F(L_1, \ldots, L_n)[\beta]$. Therefore $F(L_1, \ldots, L_n)[\alpha] = F(L_1^{'}, \ldots, L_n^{'})[\beta]$.
Conversely, if we have Systematicity, Independence follows trivially by assuming $\alpha = \beta$ while Neutrality arises by considering that for every i $L_i = L_i^{'}$.

5 Argument-Wise Plurality Voting

An obvious candidate aggregation operator to check out is the *plurality* voting operator M. In this section, we analyse a number of key properties of this operator. Intuitively, for each argument, it selects the label that appears most frequently in the individual labellings.

Definition 6 (Argument-Wise Plurality). *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Given $\alpha \in \mathcal{A}$, $M(L_1, \ldots, L_n)[\alpha] = l_\alpha \in \{\texttt{in}, \texttt{out}, \texttt{undec}, \emptyset\}$ iff*

$$|\{i : L_i(\alpha) = l_\alpha\}| > \max_{l'_\alpha \neq l_\alpha} |\{i : L_i(\alpha) = l'_\alpha\}|$$

Otherwise, $M(L_1, \ldots, L_n)[\alpha] = \emptyset$.

Example 4 (Three Detectives (cont.)). Continuing on Example 3, applying argument-wise plurality:

- $M(L_1, L_2, L_3)[\alpha_1] = \texttt{out}$
- $M(L_1, L_2, L_3)[\alpha_2] = \texttt{in}$
- $M(L_1, L_2, L_3)[\alpha_3] = \texttt{out}$

Note that in the case of ties, M is well-defined since $\emptyset$ is a member of every set. However, when $M(L_1, \ldots, L_n)[\alpha] = \emptyset$ for some $\alpha \in \mathcal{A}$, then the output of M is obviously not a legal labelling (*i.e.* Collective Rationality will be violated).

5.1 Strategic Manipulation

First, we ask whether the plurality aggregation rule is *strategy-proof*. Before such analysis can be done, it is important to define what might *motivate* agents to behave strategically, i.e. agent's preferences over labellings.

We define agents' preferences with respect to restricted sets of arguments in order to model situations where agents have potentially different *domains of knowledge*. As a motivating example, consider a court case where a medical expert is called as an expert witness. This expert can put forward arguments related to medical forensics, but would be unable to comment on legal issues. Similarly, an agent's arguments can be limited by their *position to know*. For example, a friend may be in a position to comment on someone's character, while a stranger's comments would not be of interest.

Let $\theta_i \in \Theta_i$ denote the *type* of agent $i \in I$ which is drawn from some set of possible types Θ_i. The type represents the private information and preferences of the agent. More precisely, θ_i determines agent i's preferences are over *outcomes* $L \in \mathcal{L}$. By $L_1 \succeq_i L_2$ we denote that agent i *weakly prefers* (or simply *prefers*) outcome L_1 to L_2. We say that agent i *strictly prefers* outcome L_1 to L_2, written $L_1 \succ_i L_2$, if and only if $L_1 \succeq_i L_2$ but not $L_2 \succeq_i L_1$. Finally, we say that agent i is *indifferent* between outcomes L_1 and L_2, written $L_1 \sim_i L_2$, if and only if both $L_1 \succeq_i L_2$ and $L_2 \succeq_i L_1$.

Here, we consider *focal-set-oriented* agents. These agents have a core set of arguments which they care about, and their only interest is in their exact judgement on those arguments being adopted by the collective.

Definition 7 (focal-set-oriented). *An agent i with labelling L_i is* focal-set-oriented *if there is a set of arguments $\bar{A}^i \subseteq \mathcal{A}$, called i's* focal-set*, such that for any labelling L:*

1. $L_i \sim_i L$ *iff* $\forall \alpha \in \bar{A}^i$, $L_i(\alpha) = L(\alpha)$;
2. $L_i \succeq L$ *otherwise.*

Focal-set-orientation defines a very general class of agent preferences. An example of a focal-set-oriented agent is a *resolute agent*, that is only satisfied if the aggregated labelling exactly matches its own labelling. At the other extreme is an *agent with a focal argument*, which only cares about the final status of a single argument. In this case, the agent's focal-set includes a single argument only.

Strategy-proofness (also known as *dominant strategy incentive compatibility*) is an important property in analysing agents' strategic incentives [10, page 871].[6] In our context, it asks whether any agent has incentive to misreport its labelling, given any possible reported labellings by other agents. Let $L_{-i} = \{L_1, \ldots, L_n\} \backslash L_i$ denote the set of labellings of agents other than agent i.

[6] In the literature, strategy-proofness and incentive compatibility are sometimes used to mean the same thing, requiring us to state explicitly the type of equilibrium under which the mechanism is implemented (*e.g.* in dominant strategies).

Definition 8 (Strategy-Proof). *Let $i \in I$ be an arbitrary agent with a labelling L_i. F is a* strategy-proof *aggregation operator iff*
$\forall L_{-i}, \forall L_i^* \neq L_i$, $F(L_1, \ldots, L_i, \ldots, L_n) \succeq_i F(L_1, \ldots, L_i^*, \ldots, L_n)$

In the context of focal-set-oriented preferences, strategy-proofness means that if the outcome does not agree with an agent's labelling of its focal arguments, then the agent cannot alter this fact by mis-reporting its labelling. Formally, let $L = F(L_1, \ldots, L_i, \ldots, L_n)$ be the aggregated labelling when i reports its own truthfully, and let $L^* = F(L_1, \ldots, L_i^*, \ldots, L_n)$ be the aggregated result when i reports some arbitrary alternative L_i^*. Strategy-proofness means that $\forall \alpha \in \bar{A}^i$, $L_i(\alpha) \neq L(\alpha)$ implies $L_i(\alpha) \neq L^*(\alpha)$.

Theorem 1. *Let I be a set of focal-set-oriented agents. The argument-wise plurality rule $M(.)$ is strategy-proof.*

5.2 Other Social Choice Properties

Having analysed the strategic manipulability of argument-wise plurality voting, we now turn to analysing whether it satisfies the properties listed above.

Theorem 2. *The argument-wise plurality voting operator M satisfies properties* 1 *to* 8.

Despite these promising results, it turns out that plurality operator does not satisfy the collective rationality property.

Example 5. Consider arguments $\alpha_1, \alpha_2, \alpha_3$ and α_4, with the attack relation depicted in Figure 4. Suppose we have three agents with the labellings L_1, L_2 and L_3. We have:

- $M(L_1, L_2, L_3)[\alpha_1] = \texttt{out}$,
- $M(L_1, L_2, L_3)[\alpha_2] = \texttt{out}$,
- $M(L_1, L_2, L_3)[\alpha_3] = \texttt{out}$,
- $M(L_1, L_2, L_3)[\alpha_4] = \texttt{out}$.

But then, $M(L_1, L_2, L_3)$ is not a labelling (see 4).

The above counter-example is a variant of the discursive dilemma [8] in the context of argument evaluation, which itself is a variant of the well-known Condorcet paradox.

It is worth noting that, when the preferences are focal-set oriented, labellings are partitioned in two classes: top labellings, which satisfy the focal-set assignment of labels and bottom labellings, which do not. These kinds of preferences are called *dichotomous*. Brams and Fishburn [2] showed that *approval voting*, a method according to which each voter can vote for as many candidates as she likes, is strategy-proof on dichotomous preferences. So why not apply *labelling-wise approval voting* instead of argument-wise plurality?

As it turns out, approval voting on labellings also fails to satisfy collective rationality. Just consider a system with only two arguments, α and β in a cycle of mutual defeat. Three rational labellings are possible for (α, β): $(\texttt{in}, \texttt{out})$,

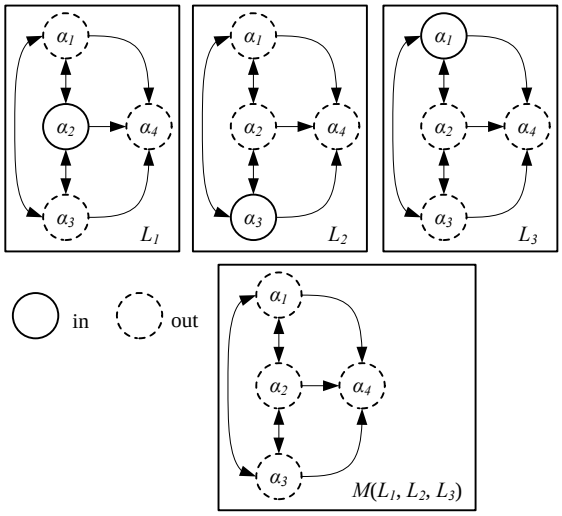

Fig. 4. Counter example to collective rationality

(`out`, `in`) and (`undec`, `undec`). Suppose there are only two agents, 1 and 2, with focal sets, $\{\alpha\}$ and $\{\beta\}$, respectively. Then each one will have a top preferred labelling, (`in`, `out`) for 1 and (`out`, `in`) for 2. Each one will vote only for her top labelling. Then, instead of having a single labelling as an outcome, a tie obtains, i.e. a set of two labellings, which certainly is not a rational labelling.

6 Is 'Good' Aggregation Possible?

In the previous section, we analysed a particular judgement aggregation operator (namely, argument-wise plurality voting). We showed that while it satisfies most key properties, it fails to always generate collectively rational judgements. This is a significant limitation, and gives rise to a more important question of whether any such operator exists. We now give a negative answer to this question, then show how this impossibility result can be avoided by restricting the domain of the voting rule.

6.1 An Impossibility Result

Social Choice Theory has been built around an impossibility result on the aggregation of preferences (Arrow's Theorem). A similar result has been found on the aggregation of judgements in propositional settings [8] and extended to more general logics [5]. The theorem below provides a counter-part for abstract argumentation framework.

We now show that there exists no possible aggregation operator F that satisfies collective rationality along with only four other minimal conditions, namely: universal domain, anonymity, systematicity, and unanimity.

Theorem 3. *There exists no F satisfying* Universal Domain*,* Anonymity*,* Systematicity*,* Unanimity*, and* Collective Rationality*.*

The idea of the proof (shown in [11]) is as follows. Given any profile and its permutation, any aggregation operator must yield the same result. By systematicity the result should be obtained on the same profile in the same way for any pair of arguments. But this means that the number of agents that vote for the 'winning' labellings on both arguments must be the same. Then a profile is constructed, for which the aggregation operator is unable to yield an outcome (i.e. legal profile) without violating this last requirement.

The above impossibility result highlights a major barrier to reaching good collective judgement about argument evaluation in general. As mentioned earlier, this is similar in flavour to Arrow's celebrated impossibility theorem on preference aggregation [1] and List and Pettit's impossibility theorem on judgement aggregation in propositional logic [8] (with the addition of unanimity). In our case, unanimity was required because, unlike in propositional logic, we have three possible labels. Unanimity prevents agents from switching to the third label (*e.g.* `undec`) in order to settle disagreement (*e.g.* between `in` and `out`).

In our context, the result means that rational aggregation on argument evaluation can only be achieved at a cost to universal domain, unanimity, anonymity or systematicity. Unfortunately, there is no escape from violating these conditions or accepting irrational aggregate argument labellings without somewhat lowering our standards in terms of desirable criteria.

6.2 Circumventing the Impossibility

Faced with the impossibility result, how can agents guarantee, in some way, reaching collective argument evaluation that is collectively rational? Following the tradition of social choice theory, we explore what restrictions on the domain of the argument-wise plurality voting rule guarantee collective rationality. In particular, we provide a full characterisation of the space of labelling profiles that guarantees collective rationality using the argument-wise plurality voting rule.

We first need a few conditions. The first is the *no-tie* condition which, as the name suggests, means that agents can always make a collective decision on each argument.

Definition 9 (No-Tie). *Labelling profile* $(L_1, \ldots, L_n)$ *satisfies the* no-tie condition *if for any* $\alpha \in \mathcal{A}$*, there exists a label* l *such that* $|\{i : L_i(\alpha) = l\}| > \max_{l' \neq l} |\{i : L_i(\alpha) = l'\}|$.

Next, we present the notion of Condorcet winner, which captures the plurality winner on an individual argument.

Definition 10 (Condorcet Winner). *We say that a label* $l_\alpha \in \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ *of an argument* $\alpha \in \mathcal{A}$ *is a* Condorcet Winner *with respect to a labelling profile* $(L_i)_{i=1}^n$*, denoted* $CW(\alpha, l_\alpha, (L_i)_{i=1}^n)$ *iff* $|\{i : L_i(\alpha) = l_\alpha\}| > |\{i : L_i = l'_\alpha\}|$ *for every label* $l'_\alpha \neq l_\alpha$*.*

Next, we list the *coordinated defeat* condition. Intuitively, this condition means that if an argument α is collectively rejected by the agents, then the agents must also collectively agree (via plurality) on accepting at least one of the counter-arguments against α. In other words, the agents' individual attacks on α are not arbitrary, but must exhibit some minimal degree of consensus.

Definition 11 (Coordinated Defeat). *A labelling profile* $(L_1, \ldots, L_n)$ *satisfies* coordinated defeat *if and only if the following holds:*
$CW(\alpha, \texttt{out}, (L_i)_{i=1}^n)$ *if and only if* $\exists \beta \in \mathcal{A}$*, such that* $\beta \rightharpoonup \alpha$ *and* $CW(\beta, \texttt{in}, (L_i)_{i=1}^n)$.

Finally, we need the following condition, which we call *Uncoordinated Indecision*. Intuitively, it requires that if an argument α is collectively accepted by the agents, then the agents must never collectively be undecided on or accept any of the counter-arguments against α. Notice that, unlike the existence condition in coordinated defeat, here the lack of indecision must hold for all defeaters.

Definition 12 (Uncoordinated Indecision). *A labelling profile satisfies* Uncoordinated indecision *if and only if the following holds:*
$CW(\alpha, \texttt{in}, (L_i)_{i=1}^n)$ *if and only if* $\nexists \beta \in \mathcal{A}$*, such that* $\beta \rightharpoonup \alpha$ *satisfies either* $CW(\beta, \texttt{undec}, (L_i)_{i=1}^n)$ *or* $CW(\beta, \texttt{in}, (L_i)_{i=1}^n)$.

Note that Uncoordinated Indecision implies that those who do not accept argument α lack sufficient coordination to gain plurality on the reasons behind their position.

We now define the necessary and sufficient restrictions on labelling profiles that guarantee collective rationality under argument-wise plurality voting.

Theorem 4. *The argument-wise plurality voting rule M satisfies* collective rationality *if and only if each labelling profile* $(L_1, \ldots, L_n)$ *in its domain satisfies* coordinated defeat*,* Uncoordinated Indecision *and the* No-Tie *condition.*

The careful reader will notice that the conditions of Coordinated Defeat and Uncoordinated Indecision, required for the result, actually correspond to the requirements of well-defined labellings (recall Definition 5). Indeed, this shows that collective rationality requires strong conditions on the collective structure of agents' labellings. These conditions are quite strong, in the sense that they cannot be reduced to properties of the individual labellings.

The full characterisation provided above has another consequence. In order to achieve collective rationality while only appealing to restrictions on *individual* labellings, we would need to make even stronger assumptions to those in Theorem 4. For example, we could require that whenever an agent labels an argument as `out`, then it must label each of its defeaters as `in`, and so on. While these kinds of restrictions guarantee the necessary partial consensus among agents, they are extremely unrealistic (even less realistic than the ones shown in the theorem). This reveals that satisfying collective rationality is not easily achievable in practice with a kind of argument-wise plurality vote.

7 Conclusion & Related Work

We explored the following question: *Given an argument structure and a set of agents, each with a legitimate subjective evaluation of the given arguments, how can the agents reach a collective compromise on the evaluation of those arguments?* We (1) proved that argument-wise plurality voting satisfies many well-known social choice properties, albeit not collective rationality; (2) proved the impossibility of any aggregation operator that simultaneously satisfies collective rationality together with universal domain, unanimity, anonymity and systematicity; and (3) fully characterised the space of individual judgements that guarantees collective rationality using argument-wise plurality voting.

Recently, Caminada and Pigozzi [3] presented some operators for aggregating multiple argument labellings into a single labelling. They focused on a 'compatibility' property: *that the social outcome must not go against any individual judgement*, and showed that this can be achieved together with collective rationality. However, they did not explore whether these operators could satisfy other classical social-choice properties. Our results provide an important complement to their work, by identifying bounds on what can be achieved simultaneously by *any* aggregation operator.

References

1. Kenneth J. Arrow. *Social choice and individual values.* Wiley, New York NY, USA, 1951.
2. Steven Brams and Peter Fishburn. Approval voting. *American Political Science Review*, 72(3):831–847, 1978.
3. Martin Caminada and Gabriella Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, (to appear).
4. Martin W. A. Caminada. On the issue of reinstatement in argumentation. In Michael Fisher, Wiebe van der Hoek, Boris Konev, and Alexei Lisitsa, editors, *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA)*, volume 4160 of *Lecture Notes in Computer Science*, pages 111–123. Springer, 2006.
5. Franz Dietrich. A generalised model of judgment aggregation. *Social Choice and Welfare*, 28:529–565, 2007.
6. 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.
7. Wulf Gärtner. *A Primer on Social Choice Theory.* Oxford University Press, 2006.
8. Christian List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, 2002.
9. Christian List and Clemens Puppe. Judgment aggregation: a survey. In *The Oxford handbook of rational and social choice.* Oxford University Press, Oxford, UK, 2009.
10. Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory.* Oxford University Press, New York NY, USA, 1995.
11. Iyad Rahwan and Tohmé. Collective Argument Evaluation as Judgement Aggregation. In *9th International Joint Conference on Autonomous Agents & Multi Agent Systems, AAMAS'2010, Toronto, Canada*, 2010.