Welfare properties of argumentation-based semantics[1]

Kate Larson and Iyad Rahwan

Abstract

Since its introduction in the mid-nineties, Dung's theory of abstract argumentation frameworks has been influential in artificial intelligence. Dung viewed arguments as abstract entities with a binary defeat relation among them. This enabled extensive analysis of different (semantic) argument acceptance criteria. However, little attention has been given to comparing such criteria in relation to the preferences of self-interested agents who may have conflicting preferences over the final status of arguments. In this paper, we define a number of agent preference relations over argumentation outcomes. We then analyse different argument evaluation rules taking into account the preferences of individual agents.

1 Introduction

Negotiation is at the core of multiagent systems since it provides procedures so that agents can find beneficial agreements. While approaches based on game-theory have proved to be highly influential [9], an alternative approach for conducting negotiations is through argumentation [8]. In argumentation, the focus is on how assertions or statements are proposed and resolved in settings where agents may have different opinions and goals. Dung presented one of the most influential computational models of argument [6]. Arguments are viewed as abstract entities, with a binary defeat relation among them. This view of argumentation enables high-level analysis while abstracting away from the internal structure of individual arguments. In Dung's approach, given a set of arguments and a binary defeat relation, a rule specifies which arguments should be accepted. A variety of such rules have been analysed using intuitive *objective* logical criteria such as consistency or self-defence [2].

Most research that employs Dung's approach discounts the fact that argumentation takes place among self-interested agents, who may have conflicting preferences over which arguments end up being accepted, rejected, or undecided. As such, argumentation can (and arguably should) be studied as an economic mechanism in which determining the acceptability status of arguments is akin to allocating resources.

In any allocation mechanism involving multiple agents (be it resource allocation or argument status assignment), two complementary issues are usually studied. On one hand, we may analyse the agents' incentives in order to predict the equilibrium outcome of rational strategies. On the other hand, we may analyse the properties of the outcomes themselves in order to compare different allocation mechanisms. The above issues are the subject of study of the field of game theory and welfare economics, respectively.

The study of incentives in abstract argumentation has commenced recently [7]. To complement this work, in this paper we initiate the study of *preference* and *welfare* in abstract argumentation mechanisms. To this end, we define several new classes of agent preferences over the outcomes of an argumentation process. We then analyse different existing rules for argument status assignment in terms of how they satisfy the preferences of the agents involved. Our focus in this paper is on the property of Pareto optimality, which measures whether an outcome can be improved for one agent without harming other agents. We also discuss more refined social welfare measures.

The paper makes two distinct contributions to the state-of-the-art in computational models of argument. First, the paper extends Rahwan and Larson's definition of argumentation outcomes [7]

[1]A preliminary version of this paper appeared in AAAI 2008, with the title *Pareto optimality in abstract argumentation*.

to account for complete labellings of arguments (as opposed to accepted arguments only). This allows us to define a number of novel preference criteria that arguing agents may have.

The second contribution of this paper is the comparison of different argumentation semantics using a well-known social welfare measure, namely Pareto optimality. To our knowledge, this is the first attempt to evaluate Dung semantics in terms of the social desirability of its outcomes. In particular, we show that in many cases, these semantics fail to fully characterise Pareto optimal outcomes. Thus, when the semantics provides multiple possible argument status assignments, our analysis presents a new criterion for selecting among those.

2 Background

In this section, we briefly outline key elements of abstract argumentation frameworks. We begin with Dung's abstract characterisation of an argumentation system [6]:

Definition 1 (Argumentation framework)**.** *An* argumentation framework *is a pair* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\mathcal{A}$ *is a 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* $(\alpha, \beta) \in \rightharpoonup$ *(sometimes written* $\alpha \rightharpoonup \beta$*).*[2]

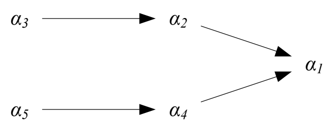


Figure 1: A simple argument graph

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

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

- S *is* conflict-free *if* $S \cap S^+ = \emptyset$.
- S defends *argument* α *if* $\alpha^- \subseteq S^+$. *We also 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 different semantics that 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 $\mathcal{F}_{AF}$.

[2]We restrict ourselves to finite sets of arguments.

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

- *S is* admissible *if it is conflict-free and defends every element in S (i.e. if $S \subseteq \mathcal{F}(S)$).*
- *S is a* complete extension *if $S = \mathcal{F}(S)$.*
- *S is a* grounded extension *if it is the minimal (w.r.t. set-inclusion) complete extension (or, alternatively, if S is the least fixed-point of $\mathcal{F}(.)$).*
- *S is a* preferred extension *if it is a maximal (w.r.t. set-inclusion) complete extension (or, alternatively, if S is a maximal admissible set).*
- *S is a* stable extension *if $S^+ = \mathcal{A} \backslash S$.*
- *S is a* semi-stable extension *if S is a complete extension of which $S \cup S^+$ is maximal.*

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.

A *grounded extension* contains all the arguments which are not defeated, as well as the arguments which are defended directly or indirectly by non-defeated arguments. This can be seen as a non-committal view (characterised by the *least* fixed point of $\mathcal{F}$). As such, there always exists a unique grounded extension. Dung [6] showed that in finite argumentation systems, the grounded extension can be obtained by an iterative application of the characteristic function to the empty set. For example, in Figure 1 the grounded extension is $\{\alpha_1, \alpha_3, \alpha_5\}$, which is the only complete extension.

A *preferred extension* is a bolder, more committed position that cannot be extended – by accepting more arguments – without causing inconsistency. Thus a preferred extension can be thought of as a maximal consistent set of hypotheses. There may be multiple preferred extensions, and the grounded extension is included in all of them.

Finally, a set of arguments is a *stable extension* if it is a preferred extension that defeats every argument which does not belong to it. A *semi-stable extension* requires the weaker condition that the set of arguments defeated is maximal.

Crucial to our subsequent analysis is the notion of *argument labelling* [3], which specifies a particular *outcome* of argumentation. 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 the condition that an argument is `in` if and only if all of its defeaters are `out`. An argument is `out` if and only if at least one of its defeaters is `in`.

Definition 5 (Argument Labelling)**.** *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ 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* $L(\beta) = \texttt{out}))$

We will make use of the following notation.

Definition 6. *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, and L a labelling over AF. We define:*

- $\texttt{in}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{in}\}$

- $\texttt{out}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{out}\}$
- $\texttt{undec}(L) = \{\alpha \in \mathcal{A} \mid L(\alpha) = \texttt{undec}\}$

In the rest of the paper, by slight abuse of notation, when we refer to a labelling L as an *extension*, we will be referring to the set of accepted arguments $\texttt{in}(L)$.

Caminada [3] established a correspondence between properties of labellings and the different extensions. These are summarised in Table 1.

Extensions	Restrictions on Labellings
complete	all labellings
grounded	minimal in minimal out maximal undec
preferred	maximal in maximal out
semi-stable	minimal undec
stable	empty undec

Table 1: The relationships between extensions and labellings.

3 Agent Preferences

Abstract argumentation frameworks have typically been analysed without taking into account the agents involved. This is because the focus has mostly been on studying the logically intuitive properties of argument acceptance criteria [2]. Recently research has commenced on evaluating argument acceptance criteria taking into account agents' strategic behaviour [7]. In this paper, we focus on developing an understanding of the underlying preferences of the agents and how these can be used in refining outcomes of the argumentation process. While we assume that agents are non-strategic, this paper complements our earlier work in that strategic behaviour is often motivated by underlying preferences.

In this paper we view an outcome as an *argument labelling*, specifying not only which arguments are accepted, but also which ones are rejected or undecided. Thus the set $\mathcal{L}$ of possible outcomes is exactly the set of all possible labellings of all arguments.

We 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 the set $\mathcal{A}_i$ of arguments available to agent i, as well as the preference criterion used to evaluate outcomes. We place no restrictions on the argument sets of agents, and for $i \neq j$ it is possible that $\mathcal{A}_i \cap \mathcal{A}_j \neq \emptyset$. An agent'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 $L_2 \succeq_i L_1$.

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 of knowledge*. 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.

While many classes of preferences are possible, in this paper we focus on *self-interested* preferences. By this we mean that we are interested in preference structures where each agent i is only interested in the status (*i.e.* labelling) of its own arguments and not on the particular status of other agents' arguments. We also emphasize that we assume that all agents understand and share the underlying argumentation system. Thus, the question of merging argumentation systems is outside the scope of this paper [5].

We start with *individual acceptability maximising preferences* [7]. Under these preferences, each agent wants to maximise the number of arguments in $\mathcal{A}_i$ that end up being accepted.

Definition 7 (Acceptability maximising preferences)**.** *An agent i has* individual acceptability maximising preferences *if* $\forall L_1, L_2 \in \mathcal{L}$ *such that* $|\mathtt{in}(L_1) \cap \mathcal{A}_i| \geq |\mathtt{in}(L_2) \cap \mathcal{A}_i|$, *we have* $L_1 \succeq_i L_2$.

An agent may, instead, aim to minimise the number of arguments in $\mathcal{A}_i$ that end up rejected.

Definition 8 (Rejection minimising preferences)**.** *An agent i has* individual rejection minimising preferences *if* $\forall L_1, L_2 \in \mathcal{L}$ *such that* $|\mathtt{out}(L_1) \cap \mathcal{A}_i| \leq |\mathtt{out}(L_2) \cap \mathcal{A}_i|$, *we have* $L_1 \succeq_i L_2$.

An agent may prefer outcomes which minimise uncertainty by having as few undecided arguments as possible.

Definition 9 (Decisive preferences)**.** *An agent i has decisive preferences if* $\forall L_1, L_2 \in \mathcal{L}$ *if* $|\mathtt{undec}(L_1) \cap \mathcal{A}_i| \leq |\mathtt{undec}(L_2) \cap \mathcal{A}_i|$ *then* $L_1 \succeq_i L_2$.

An agent may only be interested in getting *all* of its arguments collectively accepted.

Definition 10 (All-or-nothing preferences)**.** *An agent i has* all-or-nothing preferences *if and only if* $\forall L_1, L_2 \in \mathcal{L}$, *if* $\mathcal{A}_i \subseteq \mathtt{in}(L_1)$ *and* $\mathcal{A}_i \nsubseteq \mathtt{in}(L_2)$, *then* $L_1 \succ_i L_2$, *otherwise* $L_1 \sim_i L_2$.

Instead of having all of its arguments collectively accepted, an agent may be interested in having one particular *focal* argument accepted.

Definition 11 (Focal-argument preferences)**.** *An agent i has* focal-argument preferences *if and only if there exists some argument* $\alpha_i^* \in \mathcal{A}_i$ *such that* $\forall L_1, L_2 \in \mathcal{L}$ *if* $\alpha_i^* \in \mathtt{in}(L_1)$ *and* $\alpha_i^* \notin \mathtt{in}(L_2)$ *then* $L_1 \succ_i L_2$, *otherwise,* $L_1 \sim_i L_2$.

Finally, we analyse a preference structure which is not strictly self-interested. In *aggressive preferences* an agent is interested in defeating as many arguments of other all agents' as possible, and thus does care about the labelling of arguments of others.

Definition 12 (Aggressive preferences)**.** *An agent i has* aggressive preferences *if* $\forall L_1, L_2 \in \mathcal{L}$, *if* $|\mathtt{out}(L_1) \setminus \mathcal{A}_i| \geq |\mathtt{out}(L_2) \setminus \mathcal{A}_i|$ *then* $L_1 \succeq_i L_2$.

4 Pareto Optimality

Welfare economics provides a formal tool for assessing outcomes in terms of how they affect the well-being of society as a whole [1]. Often these outcomes are allocations of goods or resources. In the context of argumentation, however, an outcome specifies a particular labelling. In this section, we analyse the Pareto optimality of the different argumentation outcomes. Since labellings coincide exactly with all complete extensions, in the subsequent analysis, all `in` arguments in our outcomes are conflict-free, self-defending, and contain all arguments they defend.

A key property of an outcome is whether it is *Pareto optimal*. This relies on the notion of Pareto dominance.

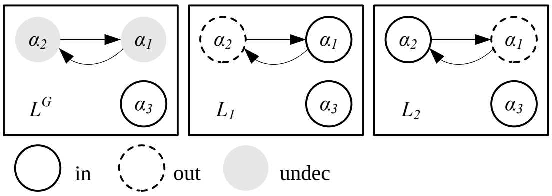
Definition 13 (Pareto Dominance)**.** *An outcome* $o_1 \in \mathcal{O}$ Pareto dominates *outcome* o_2 *if* $\forall i \in I$, $o_1 \succeq_i o_2$ *and* $\exists j \in I$, $o_1 \succ_j o_2$.

An outcome is Pareto optimal if it is not Pareto dominated by any other outcome – or, equivalently, if it cannot be improved upon from one agent's perspective without making another agent worse off. Formally:

Definition 14 (Pareto Optimality)**.** *An outcome* $o_1 \in \mathcal{O}$ *is* Pareto optimal *(or* Pareto efficient*) if there is no other outcome* $o_2 \neq o_1$ *such that* $\forall i \in I$*,* $o_2 \succeq_i o_1$ *and* $\exists j \in I$*,* $o_2 \succ_j o_1$*.*

It is interesting to see that the grounded extension is *not* Pareto optimal for a population of individual acceptability maximising agents. Consider the following example.

Example 1. *Consider the graph below with three outcomes.*

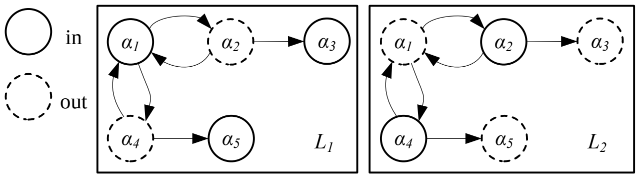


Suppose we have two agents with types $\mathcal{A}_1 = \{\alpha_1, \alpha_3\}$ *and* $\mathcal{A}_2 = \{\alpha_2\}$*. The grounded extension is the labelling* L^G*, which is not Pareto optimal. Agent 1 strictly prefers* L_1 *and is indifferent between* L^G *and* L_2*, while agent 2 strictly prefers outcome* L_2 *and is indifferent between* L^G *and* L_1*.*

The above observation is caused by the fact that the grounded extension is the *minimal* complete extension with respect to set inclusion. Thus, it is possible to accept more arguments without violating the fundamental requirement that the outcome is a complete extension (*i.e.* conflict-free, admissible, and includes everything it defends).

One might expect that all preferred extensions are Pareto optimal outcomes, since they are maximal with respect to set inclusion. However, as the following example demonstrates, this is not necessarily the case.

Example 2. *Consider the graph below, in which the graph has two preferred extensions.*



Suppose we have three individual acceptability maximising agents with types $\mathcal{A}_1 = \{\alpha_3, \alpha_4\}$*,* $\mathcal{A}_2 = \{\alpha_1\}$ *and* $\mathcal{A}_3 = \{\alpha_2, \alpha_5\}$*. Agents* $\mathcal{A}_1$ *and* $\mathcal{A}_3$ *are indifferent between the two extensions (they get a single argument accepted in either) but agent* $\mathcal{A}_2$ *strictly prefers outcome* L_1*. Thus* L_2 *is not Pareto optimal.*

However, it is possible to prove that every Pareto optimal outcome is a preferred extension (*i.e.* all non-preferred extensions are Pareto dominated by some preferred extension).

Theorem 1. *If agents have acceptability-maximising preferences and if an outcome is Pareto optimal then it is a preferred extension.*

Proof. Let $L \in \mathcal{L}$ be a Pareto optimal outcome. Assume that L is not a preferred extension. Since L is not a preferred extension, then there must exist a preferred extension $L^P \in \mathcal{L}$ such that $\texttt{in}(L) \subset \texttt{in}(L^P)$. Thus, for all i, $\texttt{in}(L) \cap \mathcal{A}_i \subseteq \texttt{in}(L^P) \cap \mathcal{A}_i$ and $|\texttt{in}(L) \cap \mathcal{A}_i| \leq |\texttt{in}(L^P) \cap \mathcal{A}_i|$ which implies that $L^P \succeq_i L$. Additionally, there exists an argument $\alpha' \in \mathcal{A}_j$ for some agent j such that $\alpha' \notin L$ and $\alpha' \in L^P$. Therefore, $|\texttt{in}(L) \cap \mathcal{A}_j| < |\texttt{in}(L^P) \cap \mathcal{A}_j|$ and so $L^P \succ_j L$. That is, L^P Pareto dominates L. Contradiction. □

The grounded extension turns out to be Pareto optimal for a different population of agents.

Theorem 2. *If agents have rejection-minimising preferences then the grounded extension is Pareto optimal.*

Proof. This follows from the fact that the grounded extension coincides with labellings with minimal $\texttt{out}$ labellings [3]. Thus any other outcome would have strictly more $\texttt{out}$ labels, resulting in at least one agent being made worse-off. □

It is also possible to prove the following.

Theorem 3. *If agents have rejection-minimising preferences, then for any outcome $L \in \mathcal{L}$, either L is the grounded extension, or L is Pareto dominated by the grounded extension.*

Proof. Let L^G denote the grounded extension, and let $L \in \mathcal{L}$ be any outcome. If $L = L^G$ then we are done. Assume that $L \neq L^G$. Since L^G has minimal $\texttt{out}$ among all outcomes in $\mathcal{L}$, then $\texttt{out}(L^G) \subset \texttt{out}(L)$. Thus, for each agent i, if argument $\alpha \in \mathcal{A}_i$ and $\alpha \in \texttt{out}(L^G)$ then $\alpha \in \texttt{out}(L)$. Therefore, $\texttt{out}(L^G) \cap \mathcal{A}_i \subset \texttt{out}(L) \cap \mathcal{A}_i$, and so $|\texttt{out}(L^G) \cap \mathcal{A}_i| \leq |\texttt{out}(L) \cap \mathcal{A}_i|$ which implies that $L^G \succeq_i L$. In addition, there also exists some agent j and argument α' such that $\alpha' \in \mathcal{A}_j$, $\alpha' \notin \texttt{out}(L^G)$ and $\alpha' \in \texttt{out}(L)$. Therefore, $|\texttt{out}(L^G) \cap \mathcal{A}_i| < |\texttt{out}(L) \cap \mathcal{A}_i|$ which implies that $L^G \succ_j L$. That is, L^G Pareto dominates L. □

The two previous theorems lead to a corollary.

Corollary 1. *The grounded extension characterises exactly the Pareto optimal outcome among a rejection minimising population.*

The following result relates to decisive agents.

Theorem 4. *If agents have decisive preferences, then all Pareto optimal outcomes are semi-stable extensions.*

Proof. This follows from the fact that any semi-stable extension coincides with a labelling in which $\texttt{undec}$ is minimal with respect to set inclusion [3]. The actual proof is similar in style to Theorem 1 and so due to space constraints we do not include the details. □

Note that any finite argumentation framework must have at least one semi-stable extension [4]. Moreover, when at least one stable extension exists, the semi-stable extensions are equal to the stable extensions, which themselves coincide with an empty $\texttt{undec}$ [4], which is ideal for decisive agents.

Corollary 2. *For agents with decisive preferences, if there exists a stable extension, then the stable extensions fully characterise the Pareto optimal outcomes for agents with decisive preferences.*

If a population of agents have all-or-nothing preferences then we can provide a partial characterisation of the Pareto optimal outcomes.

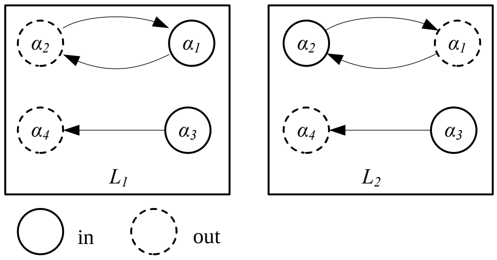
Theorem 5. *If agents have all-or-nothing preferences, then there exists a Pareto optimal preferred extension.*

Proof. We can prove this theorem by studying the possible cases. Let $\mathcal{L}$ be the set of all labellings.
Case 1: If for all $L \in \mathcal{L}$, it is the case that for all $i \in I$, $\mathcal{A}_i \nsubseteq \texttt{in}(L)$, then all agents are indifferent between all labellings, and thus all are Pareto optimal, including all preferred extensions.
Case 2: Assume there exists labelling L such that there exists an agent i with $\mathcal{A}_i \subseteq \texttt{in}(L)$ and which is Pareto optimal. If L is also a preferred extension then we are done. If L is not a preferred extension, then there must exist a preferred extension L' such that $\texttt{in}(L) \subseteq \texttt{in}(L')$. Since L was Pareto optimal, then for all agents j, it must be the case that $L \sim_j L'$ and so L' is Pareto optimal.
Case 3: Assume there exists a labelling L such that there exists an agent i with $\mathcal{A}_i \subseteq \texttt{in}(L)$ and which is not Pareto optimal. Thus, L is Pareto dominated by some labelling L^* and so there must

exist an agent j such that $\mathcal{A}_j \not\subseteq \texttt{in}(L)$ and $\mathcal{A}_i, \mathcal{A}_j \subseteq \texttt{in}(L^*)$. If L^* is not Pareto optimal then there must exist an agent k and a labelling L^{**} such that $\mathcal{A}_k \not\subseteq L^*$ and $\mathcal{A}_i, \mathcal{A}_j, \mathcal{A}_k \subseteq L^{**}$. Continue this process until the final labelling is Pareto optimal. This is guaranteed to terminate since we have a finite set of agents and labellings. Apply Case 2. □

If agents have all-or-nothing preferences, then it is possible that a preferred extension can Pareto dominate another preferred extension.

Example 3. *Consider the graph below, in which there are two preferred extensions.*



Suppose we have two agents with all-or-nothing preferences and with $\mathcal{A}_1 = \{\alpha_2, \alpha_3\}$ *and* $\mathcal{A}_2 = \{\alpha_1, \alpha_4\}$. *Outcome* L_2 *Pareto dominates outcome* L_1.

If agents have focal-argument preferences, then we can also provide a partial characterization of the Pareto optimal outcomes.

Theorem 6. *If agents have focal-argument preferences, then there exists a Pareto optimal preferred extension.*

The proof is similar to Theorem 5 and so due to space constraints we do not include the details.

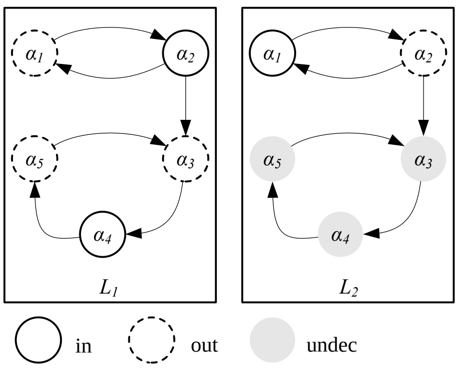
Theorem 7 says that if the population of agents have aggressive preferences, then every Pareto optimal outcome is a preferred extension.

Theorem 7. *If agents have aggressive preferences then all Pareto optimal outcomes are preferred extensions.*

Proof. Let L be a Pareto optimal outcome. Assume that L is not a preferred extension. Since L is not a preferred extension, then there must exist a preferred extension L' such that $\texttt{out}(L) \subset \texttt{out}(L')$. Thus, there must exist an agent i with $\mathcal{A}_i$ and $|\texttt{out}(L') \cap \mathcal{A}_i| > |\texttt{out}(L) \cap \mathcal{A}_i|$, and for all agents j such that $\mathcal{A}_j \in \texttt{out}(L)$, $|\texttt{out}(L') \cap \mathcal{A}_j| \geq |\texttt{out}(L) \cap \mathcal{A}_j|$ and so L' Pareto dominates L. Contradiction. □

However, not all preferred extensions are Pareto optimal, as is demonstrated in the following example.

Example 4. *Consider the graph below, in which there are two preferred extensions.*



Population Type	Pareto Optimality
Individual acceptance maximisers	Pareto optimal outcomes $\subseteq$ preferred extensions (Theorem 1)
Individual rejection minimisers	Pareto optimal outcome $=$ grounded extension (Theorem 2, 3, and Corollary 3)
Decisive	Pareto optimal outcomes $\subseteq$ semi-stable extensions (Theorem 4); if a stable extension exists, then Pareto optimal outcomes = stable extensions (Corollary 2)
All-or-nothing	Some preferred extension (Theorem 5) and possibly other complete extensions
Focal argument	Some preferred extension (Theorem 6) and possibly other complete extensions
Aggressive	Pareto optimal outcomes $\subseteq$ preferred extensions (Theorem 7)

Table 2: Classical extensions & Pareto optimality

Suppose we have three agents with aggressive preferences such that $\mathcal{A}_1 = \{\alpha_2, \alpha_4\}$*,* $\mathcal{A}_2 = \{\alpha_1, \alpha_3\}$ *and* $\mathcal{A}_3 = \{\alpha_5\}$*. Then* $L_1 \succ_1 L_2$*,* $L_1 \succ_3 L_2$ *and* $L_1 \sim_2 L_2$*. That is,* L_1 *Pareto dominates* L_2*.*

We summarise the results from this section in Table 2. These results are important since they highlight a limitation in the definitions of extensions in classical argumentation. In some cases, Pareto optimal outcomes are fully characterised by an extension (*e.g.* grounded extension and rejection minimising agents). In other cases, however, classical extensions do not provide a full characterisation (*e.g.* for acceptance maximising agents, every Pareto optimal outcome is a preferred extension but not vice versa). In such cases, we need to explicitly refine the set of extensions in order to select the Pareto optimal outcomes (*e.g.* generate all preferred extensions, then iteratively eliminate dominated ones).

5 Restrictions on the Argumentation Framework

In Section 4 we placed no restrictions on the topological structure of the argumentation framework, nor on the structure of the argument sets of agents. In this section we impose a restriction on the argumentation framework which induces *coherency* in the framework, and then show that this provides refined characterizations of the Pareto optimal outcomes.

Definition 15 (Coherent [6])**.** *An argumentation framework, AF, is* coherent *if each preferred extension of AF is stable.*

We introduce an extended definition of defeat.

Definition 16 (Indirect defeat [6])**.** *Let $\alpha, \beta \in \mathcal{A}$. We say that α* indirectly defeats *β if and only if there is an odd length path from α to β in the argument graph.*

Dung introduced the notion of an argumentation framework being *limited-controversial*, which is equivalent to there being no odd-length cycles in the argumentation graph. That is, an argumentation framework is limited-controversial if no argument indirectly defeats itself. Given this restriction on the argumentation framework, the following result is obtained.

Theorem 8. *Every limited-controversial argumentation framework is coherent [6].*

Theorem 8 and Definition 15 together imply Corollary 3.

Corollary 3. *If an argumentation framework, AF, contains no odd-length cycles, then all of its preferred extensions are stable. That is, if L^P is a preferred extension of AF then* $\texttt{undec}(L^P) = \emptyset$.

We now introduce a restriction on the sets of arguments that agents can maintain. In particular, we assume that for each agent i, the set of arguments, $\mathcal{A}_i$, contains no arguments which indirectly defeat each other, given the argumentation framework, AF. For example, referring to the figure in Example 2, argument α_4 indirectly defeats α_1, α_3 and α_5. Thus, we assume that no agent's argument set contains both α_4 and either α_1, α_3 or α_5. Intuitively, this property implies that each agent's arguments must be conflict-free (*i.e.* consistent), both explicitly and implicitly. Explicit consistency implies that no argument defeats another. Implicit consistency implies that other agents cannot possibly present a set of arguments that reveal an indirect defeat among one's own arguments. More concretely, exposing an indirect defeat chain can be seen as exposing a fallacy in one's arguments.

If the argument sets of the agents contain no indirect defeats with respect to the argumentation framework, then there are no odd-length cycles in the entire argument graph since, otherwise, at least one agent would have an argument that indirectly defeats itself. This allows us to provide a further characterization of the Pareto optimal outcomes for certain classes of agents' preferences.

Theorem 9. *Assume that agents have decisive preferences and that no agent has an argument set that contains indirect defeats. Then the set of stable extensions completely characterises the Pareto optimal outcomes.*

Proof. From Corollary 3 any labelling L^P that corresponds to a preferred extension must be a stable extension. From Corollary 2 if stable extensions exists then they fully characterise the set of Pareto optimal outcomes for decisive agents. □

Theorem 10. *Assume that agents have acceptability-maximising preferences, and that no agent has an argument set that contains indirect defeats. Then,*

- *there exists at least one stable extension, and*
- *every Pareto optimal outcome is a stable extension.*

Proof. Dung proved that every argumentation framework has at least one preferred extension (Corollary 12 [6]). Given the restriction on the agents' argument sets, there are no odd-length cycles and so all preferred extensions are stable (Corollary 3). By Theorem 1 if an outcome is Pareto optimal then it must be a preferred extension, and, thus, a stable extension. □

Finally, Theorems 9 and 10 allow us to characterize the Pareto optimal outcomes even when the agent population contains different preferences.

Corollary 4. *For agent populations consisting of both acceptability-maximising and decisive preferences, if agents' argument sets contain no indirect defeats with respect to the argumentation framework, then every Pareto optimal outcome is a stable extension.*

6 Further Refinement using Social Welfare

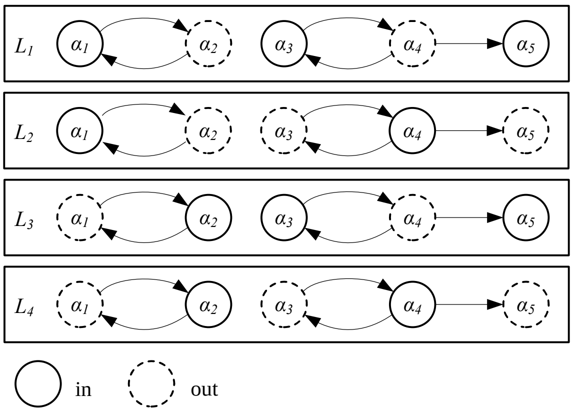
While Pareto optimality is an important way of evaluating outcomes, it does have some limitations. First, as highlighted above, there may be many Pareto optimal outcomes, and it can be unclear why one should be chosen over another. Second, sometimes Pareto optimal outcomes may be *undesirable* for some agents. For example, in a population of individual acceptability maximising agents, a preferred extension which accepts all arguments of one agent while rejecting all other arguments is Pareto optimal.

Social welfare functions provide a way of combining agents' preferences in a systematic way in which to compare different outcomes, and in particular, allow us to compare Pareto optimal extensions. We assume that an agent's preferences can be expressed by a utility function in the standard way and that it is possible to compare utility functions of the agents in a meaningful way. A social welfare function is an increasing function of individual agents' utilities and is related to the notion of Pareto optimality in that any outcome that *maximises* social welfare is also Pareto optimal.

Thus by searching for social-welfare maximising outcomes we select outcomes from among the set of Pareto optimal ones.

While there are many types of social welfare functions, two important ones are the utilitarian and egalitarian social welfare functions.[3] Example 5 illustrates how these functions can be used to compare different Pareto optimal outcomes.

Example 5. *Consider the graph below with four preferred extensions.*



Assume that there are two agents with $\mathcal{A}_1 = \{\alpha_1, \alpha_3, \alpha_5\}$ *and* $\mathcal{A}_2 = \{\alpha_2, \alpha_4\}$*, and that these agents have acceptability maximising preferences with utility functions* $u_i(L, \mathcal{A}_i) = |\mathtt{in}(L) \cap \mathcal{A}_i|$. L_1, L_3 *and* L_4 *are all Pareto optimal.* L_1 *and* L_3 *both maximise the utilitarian social welfare, while* L_3 *also maximises the egalitarian social welfare function.*

The above analysis shows that by taking into account welfare properties, it is possible to provide more fine grained criteria for selecting among classical extensions (or labellings) in argumentation frameworks. Such refined criteria can be seen as a sort of *welfare semantics* for argumentation.

7 Discussion and Conclusion

Until recently, argumentation-based semantics have been compared mainly on the basis of how they deal with specific benchmark problems (argument graph structures with odd-cycles *etc.*). Recently, it has been argued that argumentation semantics must be evaluated based on more general intuitive principles [2]. Our work can be seen to be a contribution in this direction. We introduced a new perspective on analysing and designing argument acceptability criteria in abstract argumentation frameworks. Acceptability criteria can now be evaluated not only based on their logically intuitive properties, but also based on their welfare properties in relation to a society of agents.

Our framework and results can be used to decide which argument evaluation rule to use given the type of agent population involved. While we formulated the problem as being *multiagent* in nature, our findings can also be extended to single-agent settings. In situations where there are several extensions, the agent can be consulted as to its preferences, in order to select the extension that the agent prefers.

The results are also of key importance to argumentation mechanism design (ArgMD) [7] where agents may argue strategically – *e.g.* possibly hiding arguments. ArgMD aims to design rules of interaction such that self-interested agents produce, in equilibrium, a particular desirable social outcome (*i.e.* the rules *implement* a particular social choice function). Understanding what social

[3]Given some outcome o, the utilitarian social welfare function returns the sum of the agents' utilities for o, while the egalitarian social welfare function returns $\min_i u_i(o, \theta_i)$.

outcomes are desirable (in this case, Pareto optimal) for different kinds of agents is an important step in the ArgMD process. Indeed, a major future research direction, opened by this paper, is the design of argumentation mechanisms that implement Pareto optimal social choice functions under different agent populations.

Acknowledgements

We thank Martin Kiefel for his contributions. He played a significant role in motivating Section 5.

References

[1] Kenneth J. Arrow, A. K. Sen, and K. Suzumura, editors. *Handbook of Social Choice and Welfare*, volume 1. Elsevier Science Publishers (North-Holland), 2002.

[2] Pietro Baroni and Massimiliano Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10–15):675–700, 2007.

[3] Martin W. A. Caminada. On the issue of reinstatement in argumentation. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA 2006)*, volume 4160 of *Lecture Notes in Computer Science*, pages 111–123. Springer, 2006.

[4] Martin W. A. Caminada. Semi-stable semantics. In *Proceedings of the 1st International Conference on Computational Models of Argument (COMMA)*, pages 121–130, Amsterdam, Netherlands, 2006.

[5] 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.

[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] Iyad Rahwan and Kate Larson. Mechanism Design for Abstract Argumentation. In *Proceedings of the 7th International Joint Conference on Autonomous Agents & Multi Agent Systems, AAMAS'2008, Estoril, Portugal*, 2008.

[8] Iyad Rahwan, Sarvapali D Ramchurn, Nicholas R. Jennings, Peter McBurney, Simon Parsons, and Liz Sonenberg. Argumentation based negotiation. *Knowledge Engineering Review*, 18(4):343–375, 2003.

[9] Jeff Rosenschein and Gilad Zlotkin. *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*. MIT Press, Cambridge MA, USA, 1994.

Kate Larson
Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada
Email: `klarson@cs.uwaterloo.ca`

Iyad Rahwan
Faculty of Informatics, British University in Dubai
P.O. Box 502216, Dubai, UAE
(Fellow) School of Informatics, University of Edinburgh
Edinburgh, EH8 9LE, UK
Email: `irahwan@acm.org`