Argumentation Mechanism Design for Preferred Semantics

Shengying Pan [a] Kate Larson [a] Iyad Rahwan [b,c,d]
[a] *Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada*
[b] *Masdar Institute of Science & Technology, UAE*
[c] *Massachusetts Institute of Technology, USA*
[d] *University of Edinburgh, UK*

Abstract. Recently Argumentation Mechanism Design (ArgMD) was introduced as a paradigm for studying argumentation using game-theoretic techniques. To date, this framework has been used to study under what conditions a direct mechanism based on Dung's grounded semantics is strategy-proof (i.e. truth-enforcing) when knowledge of arguments is private to self-interested agents. In this paper, we study Dung's preferred semantics in order to understand under what conditions it is possible to design strategy-proof mechanisms. This is challenging since, unlike with the grounded semantics, there may be multiple preferred extensions, forcing a mechanism to select one. We show that this gives rise to interesting strategic behaviour, and we show that in general it is not possible to have a strategy-proof mechanism that selects amongst the preferred extensions in a non-biased manner. We also investigaet refinements of preferred semantics which induce unique outcomes, namely the skeptical-preferred and ideal semantics.

1. Introduction

Argumentation has become a key model for automated reasoning and rational interaction in artificial intelligence. Key to its success has been Dung's work on abstract argumentation frameworks [6]. In this model arguments are viewed as abstract entities, with a binary defeat relation among them. This abstract framework has been beneficial for such things like the study of criteria (i.e. semantics) for evaluating outcomes of complex argument structures [1]. However, this body of work assumes that all arguments are given, a priori, for evaluation by an omniscient reasoner.

Recently there has been interest in studying strategic issues which arise in a *multi-agent* view of argumentation. In this setting, each agent has knowledge of some sub-set of the arguments, which reflects the (possibly conflicting) information available to that agent. Arguments known to different agents may overlap. However, each agent is self-interested,[1] in the sense that the agent has some preference over which arguments end up being accepted. As a result, an agent may benefit from acting strategically, by misreporting its private information (i.e. the arguments it is aware of), either passively (by hiding arguments) or actively (by stating arguments it does not believe to hold).

[1] Note that self-interest does not necessarily imply selfishness. One's own interests may well happen to align with those of others.

The strategic view of argumentation raises a question akin to (game-theoretic) mechanism design. Just as an auction (or a voting rule) is a rule that maps the revealed bids (or preferences) of different agents into a social outcome by allocating resources, an argumentation semantics maps the arguments revealed by different agents into a set of accepted arguments (See Figure 1). The question then becomes: *what strategic incentives are imposed by different argument evaluation criteria, when arguments are distributed among self-interested agents?*

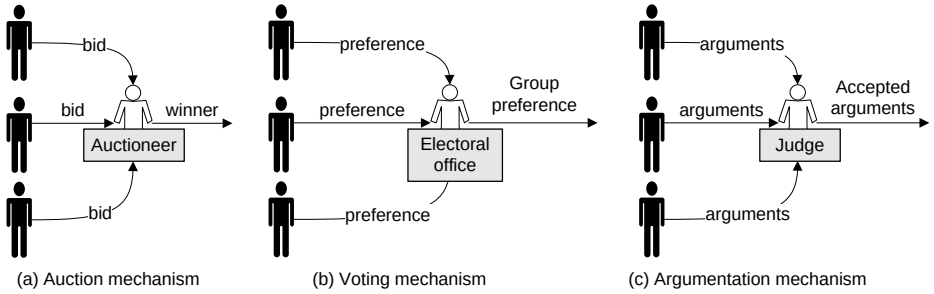

Figure 1. Argumentation mechanism (semantics) analogous to auction or voting mechanism

Rahwan and Larson proposed a new approach which they called *Argumentation Mechanism Design* (ArgMD) in which argument-evaluation procedures (or semantic criteria) are analysed to understand which strategic behaviour arises [10]. Rahwan *et al* undertook a detailed case study of the *grounded semantics* and, using the ArgMD framework, provided a full characterisation of strategy-proofness (i.e. truth-telling being a dominant strategy equilibrium) under the grounded semantics when agents can both hide and lie about their arguments [11].

In this paper we extend the analysis of strategic behaviour in argumentation frameworks to incorporate the *preferred semantics*, a more credulous semantics. While the grounded semantics induces a single outcome, the preferred semantics can result in multiple outcomes. We study whether this gives rise to new strategic-behaviour on the part of the agents and provide a graph-theoretical partial characterisation of strategy-proofness under these semantics. We also provide an analysis of two refinements of the preferred semantics: the ideal and skeptical-preferred semantics.

2. Background on Abstract Argumentation

In this section we 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 argument* α *defeats an argument* β *iff* $(\alpha, \beta) \in \rightharpoonup$ *and write this as* $\alpha \rightharpoonup \beta$*. For simplicity we restrict ourselves to finite argument sets.*

An argumentation framework can be represented as a directed graph in which the vertices are arguments and the directed edges characterise the defeat relationship among

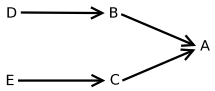

Figure 2. A simple argument graph.

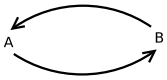

Figure 3. An argumentation framework with a cycle.

arguments. An example argument graph is shown in Figure 2. Argument A has two defeaters, B and C, which are themselves defeated by arguments D and E respectively. Cycles are also allowed in the definition of an argumentation framework, as illustrated in Figure 3. In this example there are two arguments, A and B, which defeat each other.

Let $S^+ = \{\beta \in \mathcal{A} | \alpha \rightharpoonup \beta \text{ for some } \alpha \in S\}$. Also let $\alpha^- = \{\beta \in \mathcal{A} | \beta \rightharpoonup \alpha\}$. 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 $\alpha \in \mathcal{A}$.*

1. *S is conflict-free iff $S \cap S^+ = \emptyset$.*
2. *S defends argument α iff $\alpha^- \subseteq S^+$. 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. We now look at 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 $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} | S \text{ defends } \alpha\}$.*

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

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

1. *S is admissible iff it is conflict-free and defends every element in S (i.e. if $S \subseteq \mathcal{F}(S)$).*
2. *S is a complete extension iff $S = \mathcal{F}(S)$.*
3. *S is a preferred extension iff it is a maximal (w.r.t. set-inclusion) complete extension.*
4. *S is a grounded extension iff it is a minimal (w.r.t. set-inclusion) complete extension.*

Intuitively, a set of arguments is *admissible* if it is a conflict-free set that defends itself against any defeater. An admissible set S is a *complete extension* if and only if *all* arguments defended by S are also in S. There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint. A *preferred extension* is the position that cannot be extended without causing inconsistency. The *grounded extension* only accepts arguments that are not defeated as well as arguments which are defended directly or indirectly by non-defeated arguments. We note that there always exists a unique grounded extension, but there may be multiple preferred ex-

tensions. We let $\mathcal{GE}(AF)$ represent the grounded extension of argumentation framework AF and $\mathcal{PE}(AF)$ denote the set of preferred extensions. For the argumentation framework in Figure 2 we have that $\mathcal{GE}(AF) = \{D, E, A\}$ which is also the single preferred extension, while in Figure 3 $\mathcal{GE}(AF) = \{\}$ and $\mathcal{PE}(AF) = \{\{A\}, \{B\}\}$. Finally, we formally define the notions of indirect defeat and defence.

Definition 5 (Indirect defeat and defence [6]). *Let $\alpha, \beta \in \mathcal{A}$. We say that α indirectly defeats β, written $\alpha \hookrightarrow \beta$, if and only if there is an odd-length path from α to β in the argument graph. We say that α indirectly defends β, written $\alpha \nhookrightarrow \beta$, if and only if there is an even-length path (with non-zero length) from α to β in the argument graph.*

3. Argumentation Mechanism Design

In this section we define the mechanism design problem for abstract argumentation as was introduced by Rahwan and Larson [10]. We define a mechanism with respect to an argumentation framework $\langle \mathcal{A}, \rightharpoonup \rangle$ with semantics $\mathcal{S}$, and we assume that there is a set $\{1, 2, \ldots, I\}$ of self-interested agents. A key notion in mechanism design is the *type* of an agent. An agent's type is all the information which is relevant to the agent when formulating its preferences over outcomes. In our framework, we define an agent's type to be its set of arguments.

Definition 6 (Agent Type). *Given an argumentation framework $\langle \mathcal{A}, \rightharpoonup \rangle$, the type of agent i, $\mathcal{A}_i \subseteq \mathcal{A}$, is the set of arguments that the agent is capable of putting forward.*

Note that $\alpha \in \mathcal{A}_i$ is not necessarily true, or even believed by i to be acceptable. It simply reflects a piece of information the agent has. Indeed, if α involved a contradiction, it would be self-defeating and hence never accepted by anyone.

A social choice function maps a type profile (vector of agent types) to a subset of arguments. In particular, we will interpret the set of arguments to be the arguments which are deemed to be acceptable if the actual types of the agents were known. We will determine the acceptability of arguments with respect to some specified semantics.

Definition 7 (Argument Acceptability Social Choice Functions). *Given an argumentation framework $\langle \mathcal{A}, \rightharpoonup \rangle$ with semantics $\mathcal{S}$, and given a type profile $(\mathcal{A}_1, ..., \mathcal{A}_I)$ such that $\mathcal{A}_1 \cup \ldots \cup \mathcal{A}_I \subseteq \mathcal{A}$, the argument acceptability social choice function f is defined as the set of acceptable arguments given the semantics $\mathcal{S}$. That is, $f(\mathcal{A}_1, ..., \mathcal{A}_I) = Acc(\langle \mathcal{A}_1 \cup ... \cup \mathcal{A}_I, \rightharpoonup \rangle, \mathcal{S})$*

As is common in the mechanism design literature, we assume that agents have preferences over the outcomes $o \in 2^{\mathcal{A}}$, and we represent these preferences using utility functions where $u_i(o, \mathcal{A}_i)$ denotes agent i's utility for outcome o when its type is argument set $\mathcal{A}_i$. Agent i prefers outcome o_1 to o_2 when $u_i(o_1, \mathcal{A}_i) > u_i(o_2, \mathcal{A}_i)$. In this paper, we assume that agents have *focal argument preferences*.

Definition 8 (Focal-Argument Preferences). *An agent i has* focal-argument preferences *if there exists some argument $\alpha_i^* \in \mathcal{A}_i$ such that for any outcomes o_1, $o_2 \in \mathcal{O}$ such that $\alpha_i^* \in o_1$ and $\alpha_i^* \notin o_2$ then $u_i(o_1, \mathcal{A}_i) > u_i(o_2, \mathcal{A}_i)$. Otherwise, $u_i(o_1, \mathcal{A}_i) = u_i(o_2, \mathcal{A}_i)$.*

Informally, this class of preferences can be interpreted as each agent i having a single argument, α_i^*, in which they are interested, while the other arguments are of interest only with respect to how they support α_i^*.

Agents may not have incentive to reveal their true type because they may be able to influence the final argument status assignment by lying, and thus obtain higher utility. We explicitly assume that the defeat relationship, $\rightharpoonup$, is known and understood by all agents. Then there are two ways in which an agent may lie. First, it might claim to *have* arguments which are not in its argument set (but are still part of $\mathcal{A}$). In such a case, we say that the agent *makes up* arguments. Second, it might *hide* arguments. By refusing to reveal certain arguments, an agent might be able to break defeat chains in the argument framework, thus changing the final set of acceptable arguments.

A strategy for agent i, $s_i(\mathcal{A}_i) \in \Sigma_i$, is a plan that describes what actions the agent will take for every decision that the agent might be called upon to make, for each possible piece of information that the agent may have at each time it is called to act. In our model strategies specify which arguments an agent should reveal when. The notation Σ_i denotes the *strategy space* of agent i and contains all possible legal strategies that an agent may follow.

Definition 9 (Argumentation Mechanism)**.** *Given an argumentation framework $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ and semantics $\mathcal{S}$, an argumentation mechanism is defined as*

$$\mathcal{M}_{AF}^{\mathcal{S}} = (\Sigma_1, ..., \Sigma_I, g(\cdot))$$

where Σ_i is an argumentation strategy space of agent i and $g : \Sigma_1 \times ... \times \Sigma_I \rightarrow 2^{\mathcal{A}}$.

We are particularly interested in situations where the agents' strategies are restricted so that they can only reveal sets of arguments once. Mechanisms with this particular restriction are called *direct mechanisms*.

Definition 10 (Direct Argumentation Mechanism)**.** *Given an argumentation framework $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ and semantics $\mathcal{S}$, a direct argumentation mechanism is defined as*

$$\mathcal{M}_{AF}^{\mathcal{S}} = (\Sigma_1, ..., \Sigma_I, g(\cdot))$$

where $\Sigma_i = 2^{\mathcal{A}_i}$ and $g : \Sigma_1 \times ... \times \Sigma_I \rightarrow 2^{\mathcal{A}}$.

If, given, $\mathcal{M}_{AF}^{\mathcal{S}}$ all agents are best off selecting a strategy such that $s_i(\mathcal{A}_i) = \mathcal{A}_i$ (no matter what any other agent is doing) then we say that the mechanism is *strategy-proof*. That is, agents have incentive to truthfully report their actual arguments.[2] The goal of ArgMD is to understand when and why it is possible or impossible to ensure that a mechanism is strategy-proof. The restriction to direct mechanisms is without loss of generality since the *Revelation Principle* states that if there exists a mechanism such that agents reveal their types truthfully, then there is a direct mechanism with this property [8]. Finally, we define a direct mechanism for argumentation based on the preferred semantics. We refer to a specific action of agent i as $\mathcal{A}_i^o \in \Sigma_i$.

[2]The term strategy-proof is used when ever all agents have incentive to truthfully report their types, even if truth-telling is only a weakly dominant strategy [8].

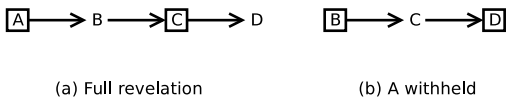

Figure 4. Hiding an argument is beneficial

Definition 11 (Preferred Direct Argumentation Mechanism)**.** *A* preferred direct argumentation mechanism *for argumentation framework* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *is* $\mathcal{M}_{AF}^{\mathcal{PE}} = (\Sigma_1, \ldots, \Sigma_I, g(\cdot))$ *where* $\Sigma_i \in 2^{\mathcal{A}}$ *is the set of strategies available to each agent, and* $g(\mathcal{A}_1^o, \ldots, \mathcal{A}_I^o) = Acc(\langle \mathcal{A}_1^o \cup \ldots \cup \mathcal{A}_I^o, \rightharpoonup \rangle, S^{\mathcal{PE}})$ *where* $S^{\mathcal{PE}}$ *denotes the preferred acceptability semantics.*

4. Implementing the Preferred Extension

We start this section with an illustrative example showing why agents may have incentive to be strategic when asked to reveal their arguments.

Example 1. *Assume that there are three agents where* $\mathcal{A}_1 = \{A, D\}, \mathcal{A}_2 = \{B\}$ *and* $\mathcal{A}_3 = \{C\}$*, and with focal arguments* ${\alpha^*}_1 = D, {\alpha^*}_2 = B$ *and* ${\alpha^*}_3 = C$*. Assume also that the defeat relationship* $\rightharpoonup = \{(A, B), (B, C), (C, D)\}$*. If all agents reveal their arguments then the resulting argument graph is shown in Figure 4(a). There is a single preferred extension where the arguments marked by boxes in Figure 4(a) are accepted. However, if agent 1 does not reveal argument* A *then the unique preferred extension is shown in Figure 4(b). Note that in this outcome, agent 1's focal argument is accepted, while in the original outcome, agent 1's focal argument was not accepted. Thus, agent 1 has incentive to hide its argument* A*.*

This example is also illustrative of another property of argumentation frameworks. If the underlying argumentation graph is *acyclic* then the unique preferred extension is equal to the grounded extension. Thus, it immediately follows that all ArgMD results for the grounded extension also apply to preferred extensions when the underlying argumentation graph is acyclic [11].

Theorem 1. *Suppose each agent* $i \in \{1, \ldots, I\}$ *has a focal argument* $\alpha_i^* \in \mathcal{A}_i$*, and suppose that the underlying argumentation graph is acyclic. If the following conditions hold:*

- *no agent type contains an (in)direct defeat against its focal argument*
- *no argument outside any agent's type (in)directly defends its focal argument*

then $\mathcal{M}_{AF}^{\mathcal{PE}}$ *is strategy-proof.*

In the rest of this paper we focus our attention to argumentation frameworks where the associated argumentation graph contains at least one cycle. One challenge is that for these argumentation frameworks, $\mathcal{PE}$ may contain more than one set of acceptable arguments, and thus the social choice function must have some principled way to select from amongst the elements of $\mathcal{PE}$. We propose two minimal standard properties the social choice function must exhibit when making such a selection.

Definition 12 (Agent-Anonymous)**.** *[8] A social choice function f is* agent-anonymous *if for any onto function $\pi : \{1,\ldots,I\} \mapsto \{1,\ldots,I\}$, and for any type profile $(\mathcal{A}_1,\ldots,\mathcal{A}_I)$ we have $f(\mathcal{A}_1,\ldots,\mathcal{A}_I) = f(\mathcal{A}_{\pi(1)},\ldots,\mathcal{A}_{\pi(I)})$.*

Definition 13 (Argument-Anonymous)**.** *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be any argumentation framework and let $\pi : \{1,\ldots,I\} \mapsto \{1,\ldots,I\}$ be an onto function. Define $AF^{\pi} = \langle \mathcal{A}^{\pi}, \rightharpoonup_{\pi} \rangle$ such that for any $\alpha_i \in \mathcal{A}$ then $\alpha_{\pi(i)} \in \mathcal{A}^{\pi}$ and if $\alpha_i \rightharpoonup \alpha_j$ then $\alpha_{\pi(i)} \rightharpoonup_{\pi} \alpha_{\pi(j)}$. A social choice function is* argument anonymous *if for any type profile $(\mathcal{A}_1,\ldots,\mathcal{A}_I)$ we have $f(\mathcal{A}_1,\ldots,\mathcal{A}_I) = f(\mathcal{A}_1^{\pi},\ldots,\mathcal{A}_I^{\pi})$.*

The first property affirms that the names of the agents should not matter, while the second property states the names of the arguments should not matter. Unfortunately, there is an immediate problem when trying to enforce these properties when applying them in the preferred semantics framework.

Theorem 2. *No deterministic social choice function which selects an outcome amongst the preferred extensions is both agent- and argument-anonymous.*

Proof. (Sketch) Consider the argumentation framework shown in Figure 3, and assume that there are two agents such that $\mathcal{A}_1 = \{A\}$ and $\mathcal{A}_2 = \{B\}$ (and each agent's focal argument is its single argument). $\mathcal{PE} = \{\{A\},\{B\}\}$ and there is no justification which would respect the agent- and argument-anonymity properties to select one extension over the other.

□

Theorem 2, as stated, only applies to deterministic social choice functions. It has been suggested that allowing for *randomization* with respect to selecting outcomes may circumvent certain impossibilities [5,9]. We investigate this observation as it applies to ArgMD by defining a *preferred randomized mechanism*.

Definition 14 (Preferred Randomized Mechanism)**.** *A preferred randomized mechanism for argumentation framework $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ is $\mathcal{RM}_{AF}^{\mathcal{PE}} = (\Sigma_1,...,\Sigma_I, g(\cdot))$ where:*

- *$\Sigma_i \in 2^{\mathcal{A}}$ is the set of strategies available to each agent;*
- *$g : \Sigma_1 \times ... \times \Sigma_I \rightarrow \Delta\mathcal{PE}$ where $\Delta\mathcal{PE}$ is a distribution with full support over $\mathcal{PE}$.*

Given the revealed arguments of the agents, the mechanism selects a preferred extension at random according to some pre-specified distribution. The full-support requirement implies that any preferred extension can potentially be selected. Using $\mathcal{RM}_{AF}^{\mathcal{PE}}$ we study settings where agents' types do not contain (in)direct defeats against their own focal argument since this case was strategy-proof for acyclic argumentation frameworks.

Example 2. *Assume there are 7 agents with argument sets $\mathcal{A}_1 = \{B,F\}, \mathcal{A}_2 = \{A\}, \mathcal{A}_3 = \{C\}, \mathcal{A}_4 = \{D\}, \mathcal{A}_5 = \{E\}, \mathcal{A}_6 = \{G\}$ and $\mathcal{A}_7 = \{H\}$. For agents with only one argument their sole argument is their focal argument, α_i^*. For agent 1, $\alpha_1^* = F$. Assume the full defeat relationship (when all arguments are revealed) is shown in Figure 5. Note that there is* no *odd-length directed path between arguments B and F, which is equivalent to stating that there is no (in)direct defeat between them. In fact, there is no odd-length undirected path between the two arguments.*

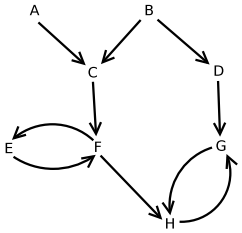

Figure 5. Argumentation graph if all agents reveal their arguments.

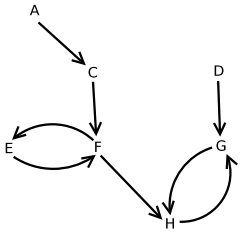

Figure 6. Argumentation graph if agent 1 hides argument B.

Assume the mechanism selects a preferred extension uniformly at random. All agents with only one argument (i.e. their focal argument) are best off revealing it. If all agents, including agent 1, reveal their arguments then $\mathcal{PE} = \{\{A, B, F, G\}, \{A, B, E, G\}, \{A, B, E, H\}\}$. *If agent 1 hides argument* B *then the resulting argumentation graph is shown in Figure 6 and* $\mathcal{PE} = \{\{A, D, E, H\}, \{A, D, F\}\}$. *If agent 1 revealed both its arguments, then the probability that its focal argument was in the chosen outcome is* $\frac{1}{3}$. *However, if agent 1 hid argument* B*, then the probability that its focal argument was in the chosen outcome increases to* $\frac{1}{2}$. *Therefore, agent 1 is best off hiding argument* B.

The restriction that the distribution must have full support, and our requirements of agent and argument anonymity mean that no matter what distribution is used agent 1 can always increase the probability of having its focal argument in the selected outcome by hiding its other argument. Thus, the properties which induced truth-telling for acyclic argumentation frameworks are not sufficient for non-acyclic frameworks. We now investigate a sufficient condition for truth-telling. It relies on two structural results for preferred extensions. Due to space limitation we are unable to include all proofs.

Lemma 1 characterizes the relationship between the preferred extensions of an argumentation framework, $AF = \langle \mathcal{A}, \rightharpoonup \rangle$, and the preferred extensions of argumentation frameworks induced by particular partitions of $\mathcal{A}$.

Lemma 1. *Let* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *be an arbitrary argumentation framework. Let* $S \subseteq \mathcal{A}, R = \mathcal{A} \setminus S$ *be two non-empty subsets of* $\mathcal{A}$ *such that for any* $\alpha \in S$ *and* $\beta \in R$ $(\alpha, \beta) \notin \rightharpoonup$ *and* $(\beta, \alpha) \notin \rightharpoonup$. *Define* $AF_S = \langle S, \rightharpoonup_S \rangle$ *where* $\rightharpoonup_S = \{(\alpha, \beta) | \alpha, \beta \in S \wedge (\alpha, \beta) \in \rightharpoonup\}$. *Define* $AF_R = \langle R, \rightharpoonup_R \rangle$ *similarly. Then*

1. *For any preferred extension* P *in* AF, $P \cap S$ *is a preferred extension in* AF_S *and* $P \cap R$ *is a preferred extension in* AF_R.
2. *For any preferred extensions* P_S *in* AF_S *and* P_R *in* AF_R, $P_S \cup P_R$ *is a preferred extension in* AF.

Lemma 2. *Let* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\alpha, \alpha' \in \mathcal{A}$ *and* $AF' = \langle \mathcal{A} \setminus \{\alpha'\}, \rightharpoonup' \rangle$ *such that* $\rightharpoonup'$ *is the restriction of* $\rightharpoonup$ *to* $\mathcal{A} \setminus \{\alpha'\}$. *Let* $Pr_D(\alpha | \mathcal{PE}(AF))$ *denote the probability that an extension containing* α *is selected at random under a distribution,* D*, with full support over* $\mathcal{PE}(AF)$ *and which satisfies the anonymity criteria. Assume that there is no undirected path between* α *and* α'. *Then* $Pr_D(\alpha | \mathcal{PE}(AF)) = Pr_{D'}(\alpha | \mathcal{PE}(AF'))$ *where* D' *is the restriction of* D *to* $\mathcal{PE}(AF')$.

Lemma 2 states that as long as there is no path between two arguments, then whether or not one argument is revealed can not influence the probability that a preferred exten-

sion containing the other argument will be chosen. We are now able to provide a partial characterization of strategy-proofness for the preferred semantics, if we assume agents will only *hide* arguments.

Theorem 3. *Suppose every agent $i \in \{1, \ldots, I\}$ has a focal argument $\alpha_i^* \in \mathcal{A}_i$. If for each agent i, $\mathcal{A}_i$ contains no argument with an undirected path to α_i^*, then $\mathcal{RM}_{AF}^{\mathcal{PE}}$ is strategy-proof.*

Proof. Let $p_i(\langle \mathcal{A}, \rightharpoonup \rangle)$ be the probability of a preferred extension containing agent i's focal argument to be chosen randomly from preferred extensions in argumentation framework $\langle \mathcal{A}, \rightharpoonup \rangle$.

Suppose the randomized mechanism is not strategy-proof, then there exists an argumentation framework $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ such that $\exists i$, for $\mathcal{A}'_i \subset \mathcal{A}_i$ a revelation of agent i and $\mathcal{A}'_{-i} = (\mathcal{A}'_1, ..., \mathcal{A}'_{i-1}, \mathcal{A}'_{i+1}, ..., \mathcal{A}'_I)$ a revelation of all agents not including i, $p_i(\langle \mathcal{A}'_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) > p_i(\langle \mathcal{A}_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle)$.

Let $n = |\mathcal{A}_i| - |\mathcal{A}'_i|, \{\beta_1, \beta_2, ..., \beta_n\} = \mathcal{A}_i \setminus \mathcal{A}'_i, \forall 1 \leq j \leq n$, there is no path (disregard direction) from β_j to i's focal argument $\hat{\alpha}^i$. By **Lemma 2**, $p_i(\langle \mathcal{A}_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) = p_i(\langle (\mathcal{A}_i \setminus \{\beta_1\}) \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) = p_i(\langle (\mathcal{A}_i \setminus \{\beta_1, \beta_2\}) \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) = \cdots = p_i(\langle (\mathcal{A}_i \setminus \{\beta_1, \beta_2, ..., \beta_n\}) \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) = p_i(\langle (\mathcal{A}'_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle)$, which contradicts $p_i(\langle \mathcal{A}'_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle) > p_i(\langle \mathcal{A}_i \cup \mathcal{A}'_{-i}, \rightharpoonup \rangle)$. Therefore, $\mathcal{RM}_{AF}^{\mathcal{PE}}$ is strategy-proof.

□

Theorem 3 states that agents have no incentive to hide arguments when their focal arguments are in subgraphs disconnected from their other arguments. This is a very strong condition and is significantly stronger than the required condition for strategy-proofness for acyclic frameworks. We also note that this is only a sufficient condition. There may be other topological restrictions which would allow for strategy-proofness.

5. Refinements of the Preferred Semantics

One way of handling the multiplicity of preferred extensions is to provide additional refinements to the semantics. In this section we look at two such refinements, the *skeptical-preferred semantics* and the *ideal semantics*, both of which provide a unique extension, thus avoiding the problem faced in the last section.

Definition 15 (Skeptical-Preferred Semantics)**.** *[2] Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework and let $\mathcal{PE}(AF)$ be the set of preferred extensions. The* skeptical-preferred extension *is $\mathcal{SP}(AF) = \cap_{S \in \mathcal{PE}(AF)} S$.*

Clearly the skeptical-preferred extension is unique, but it may not be admissible, and thus not a complete extension.

Definition 16 (Ideal Semantics)**.** *[7] Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. The* ideal extension*, $\mathcal{I}(AF)$, is the maximal (w.r.t. set-inclusion) admissible set that is a subset of each preferred extension.*

As shown by Caminada, the ideal extension always exists and is unique [3].

Proposition 1. *[3] Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. There exists exactly one maximal (w.r.t. set-inclusion) admissible set that is a subset of each preferred extension.*

The above proposition implies that $\mathcal{I}(AF) \subseteq \mathcal{SP}(AF)$.

Given these two new semantics, we define argumentation mechanisms in a similar way as we did for the preferred semantics. In particular $\mathcal{M}_{AF}^{\mathcal{SP}}$ is an argumentation mechanism where the outcome is selected using the skeptical-preferred semantics, while $\mathcal{M}_{AF}^{\mathcal{I}}$ is an argumentation mechanism where the outcome is selected using the ideal semantics. Since both the skeptical-preferred and ideal semantics result in a unique extension, we do not require randomization.

We first look at the case where agents' strategies are restricted so that they can only *hide* their arguments. While, due to space limitations, we focus on $\mathcal{M}_{AF}^{\mathcal{I}}$, we first present Lemma 3 which states what happens to the skeptical-preferred extension when a new argument is added to an argumentation framework. Any argument that had been initially acceptable w.r.t. the skeptical-preferred extension, remains acceptable in the new argumentation framework as long as the new argument did not (in)directly defeat it.

Lemma 3. *For $AF_1 = \langle \mathcal{A}, \rightharpoonup_1 \rangle$ and $AF_2 = \langle \mathcal{A} \cup \{\alpha'\}, \rightharpoonup_2 \rangle$ such that $\rightharpoonup_1 \subseteq \rightharpoonup_2$ and $(\rightharpoonup_2 \setminus \rightharpoonup_1) \subseteq (\{\alpha'\} \times \mathcal{A}) \cup (\mathcal{A} \times \{\alpha'\})$. If α is in the skeptical preferred extension of AF_1, and α' doesn't indirectly defeat α, then α is still in the skeptical preferred extension in AF_2.*

A similar result can be extended for the ideal extensions.

Lemma 4. *For $AF_1 = \langle \mathcal{A}, \rightharpoonup_1 \rangle$ and $AF_2 = \langle \mathcal{A} \cup \{\alpha'\}, \rightharpoonup_2 \rangle$ such that $\rightharpoonup_1 \subseteq \rightharpoonup_2$ and $(\rightharpoonup_2 \setminus \rightharpoonup_1) \subseteq (\{\alpha'\} \times \mathcal{A}) \cup (\mathcal{A} \times \{\alpha'\})$. If α is in the ideal extension of AF_1, and α' doesn't indirectly defeat α, then α is still in the ideal extension in AF_2.*

Proof. In an argumentation framework, the ideal extension is always a subset of the skeptical preferred extension.

Let S be the set of arguments in the ideal extension of AF_1 which are either α or (in)direct defenders of α. Then S must be admissible. Moreover, S is a subset of the ideal extension of AF_1 thus a subset of the skeptical preferred extension. $\forall \beta \in S, \beta$ is in the skeptical preferred extension of AF_1 and α' doesn't indirectly defeat β since in such case α' will indirectly defeat α. Therefore, β is in the skeptical preferred extension of AF_2 by Lemma 3. Thus S is a subset of the skeptical preferred extension of AF_2. Clearly S is still conflict-free in AF_2. Since $\alpha \in S$, if S is admissible in AF_2, by Proposition 1 and Definition 16, the ideal extension in AF_2 is a superset of S thus contains α, a contradiction. If S is not admissible in AF_2, the only possible way to break the admissibility is to have α' defeat one argument $\beta \in S$ in AF_2. But since β is an indirect defender of α, α' therefore indirectly defeats α, a contradiction. Hence, α is still in the ideal extension of AF_2. □

Theorem 4. *Suppose every agent $i \in \{1, \ldots, I\}$ has a focal argument $\alpha_i^* \in \mathcal{A}_i$. If each agent's type contains no (in)direct defeat against α_i^*, then $\mathcal{M}_{AF}^{\mathcal{I}}$ is strategy-proof.*

Proof. (Sketch) Due to space limitations we only provide a sketch of the induction proof by describing the base case and induction step. The goal is to show formally that $\forall i \in$

$\{1, \ldots, I\}, u_i(Acc(\langle \mathcal{A'}_1 \cup \ldots \cup \mathcal{A}_i \cup \ldots \cup \mathcal{A'}_I, \rightharpoonup \rangle, \mathcal{S}^{\mathcal{I}}), \mathcal{A}_i) \geq u_i(Acc(\langle \mathcal{A'}_1 \cup \ldots \cup \mathcal{A}_i^{\circ} \cup \ldots \cup \mathcal{A'}_I, \rightharpoonup \rangle, \mathcal{S}^{\mathcal{I}}), \mathcal{A}_i)$ for any $\mathcal{A}_i^{\circ} \subset \mathcal{A}_i$ and $\mathcal{A}'_j \subseteq \mathcal{A}_j$.

We use induction over the sets of arguments agent i may reveal, starting from the focal argument α_i^*. We show that, considering any strategy $\mathcal{A}_i'' \subseteq \mathcal{A}_i$, revealing one more argument can only increase i's chance of getting α_i^* accepted, $i.e.$ it (weakly) improve i's utility. **Base Step:** If $\mathcal{A}_i = \{\hat{\alpha}^i\}$, then trivially, revealing $\mathcal{A}_i$ weakly dominates revealing $\emptyset$.

Induction Step: Suppose that revealing argument set $\mathcal{A}_i'' \subseteq \mathcal{A}_i$ weakly dominates revealing any subset of $\mathcal{A}_i''$. We need to prove that revealing any set $\mathcal{A}_i'$, where $\mathcal{A}_i'' \subset \mathcal{A}_i' \subseteq \mathcal{A}_i$ and $|\mathcal{A}_i'| = |\mathcal{A}_i''| + 1$, weakly dominates revealing $\mathcal{A}_i''$. This follows from Lemma 4. □

We note that a similar characterisation is possible under the skeptical-preferred semantics. Interestingly, Theorem 4 provides the same characterization for when argument hiding is not beneficial for agents as for the grounded semantics [11]. This is true even though the underlying *structure* of the extensions is quite different, and the properties required for the grounded semantics characterisation do not immediately translate to the skeptical-preferred and ideal semantics due to the difference in their definitions. It is also interesting to note that if we study the situation where agents may also *make up* arguments, then we again obtain a similar characterisation as for the grounded semantics. We state the theorem for the ideal semantics, but an identical theorem also holds for the skeptical-preferred semantics.

Theorem 5. *Suppose every agent $i \in \{!, \ldots, I\}$ has a focal argument $\alpha_i^* \in \mathcal{A}_i$, and that agents can both hide or lie about arguments. If the following conditions hold:*

1. *each agent's type contains no (in)direct defeat against α_i^* (formally $\forall i \in I, \nexists \beta \in \mathcal{A}_i$ such that $\beta \hookrightarrow \alpha_i^*$);*
2. *for any agent i, no argument outside i's type (in)directly defends α_i^* (formally $\forall i \in I, \nexists \beta \in \mathcal{A} \setminus \mathcal{A}_i$ such that $\beta \not\hookrightarrow \alpha_i^*$);*

then $\mathcal{M}_{AF}^{\mathcal{I}}$ is strategy-proof.

Proof. (Sketch)

What we want to prove is that for an arbitrary $S \neq \mathcal{A}_i$: $\alpha_i^* \notin \mathcal{I}(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\alpha_i^* \notin \mathcal{I}(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. where $\mathcal{I}(AF)$ is the ideal extension of argumentation framework AF. The rest of the proof follows a similar logic as that for the grounded extension. Due to space restrictions we refer the reader to [11].

□

6. Conclusion

ArgMD is a useful paradigm for reasoning about argumentation among self-interested agents using game-theoretic techniques. To date it has been applied only to the grounded semantics, which is often criticized as taking an overly skeptical stance with respect to argument acceptability. In this paper we applied the ArgMD framework to the preferred semantics.

Unlike grounded semantics which yield a unique extension, multiple preferred extensions may exist for arbitrary argumentation frameworks. We proposed some minimal

properties which ensured non-bias with respect to agents and arguments when selecting from amongst the preferred extensions. We illustrated that it was impossible to satisfy our minimal anonymity properties with a deterministic social choice function. By incorporating randomization into our mechanism, we determined conditions under which agents had incentive to reveal all their arguments. We intend to investigate less restrictive requirements on agents types, or other possible restrictions or extensions of the mechanism, so as to ensure strategy-proofness.

We also studied refinements of the preferred semantics which result in unique extensions. In particular, we were able to provide a similar characterization of strategy-proofness for the skeptical-preferred and ideal semantics, as had previously been provided for the grounded semantics. We found this interesting since the underlying structure of the extensions is quite different. We conjecture that the *uniqueness* of the extensions is important in the characterization, and intend to investigate other unique extensions (for example, the eager extension [3]).

We assumed, throughout this paper, that agents' preferences had a particular structure, that is agents had *focal-argument preferences*. One obvious question is whether the results in this paper are also applicable if agents have different preference structures, which opens up another line of future research.

It is worth noting how work reported in this paper differs from recent work on judgement aggregation in argumentation [12,4]. In judgement aggregation, all arguments are given, and each agent has preferences over how these arguments should be evaluated. In our work, on the other hand, the arguments themselves are distributed among the agents, and different argument graphs emerge based on what they choose to reveal.

References

[1] P. Baroni and M. Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171:675–700, 2007.

[2] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract argumentation-theoretic framework for default reasoning. *Artificial Intelligence*, 93(1-2):63–101, 1997.

[3] M. Caminada. Comparing two unique extension semantics for formal argumentation: Ideal and eager. In *Proceedings of 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC)*, pages 81–87, 2007.

[4] M. Caminada and G. Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, (to appear).

[5] V. Conitzer and T. Sandholm. Nonexistence of voting rules that are usually hard to manipulate. In *Proceedings of the 21st AAAI*, pages 627–634, 2006.

[6] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[7] P. M. Dung, P. Mancarella, and F. Toni. A dialectic procedure for sceptical, assumption-based argumentation. In *Proceedings of Computational Models of Argument (COMMA)*, pages 145–156, 2006.

[8] A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[9] A. Procaccia. Can approximation circumvent Gibbard-Satterthwaite? In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI 2010)*, 2010.

[10] I. Rahwan and K. Larson. Mechanism design for abstract argumentation. In *Proceedings of AAMAS 2008*, pages 1031–1038, 2008.

[11] I. Rahwan, K. Larson, and F. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI 2009)*, pages 251–256, Pasadena, CA, July 2009.

[12] I. Rahwan and F. Tohmé. Collective Argument Evaluation as Judgement Aggregation. In *9th International Joint Conference on Autonomous Agents & Multi Agent Systems, AAMAS'2010, Toronto, Canada*, 2010.