A Characterisation of Strategy-Proofness for Grounded Argumentation Semantics

Iyad Rahwan[1,2], **Kate Larson**[3], **and Fernando Tohmé**[4]
[1]Faculty of Informatics, British University in Dubai, Dubai, UAE
[2](Fellow) School of Informatics, University of Edinburgh, UK
[3]Cheriton School of Computer Science, University of Waterloo, Canada
[4]LIDIA, Universidad Nacional del Sur, Bahía Blanca, CONICET, Argentina

Abstract

Recently, Argumentation Mechanism Design (ArgMD) was introduced as a new paradigm for studying argumentation among self-interested agents using game-theoretic techniques. Preliminary results showed a condition under which a direct mechanism based on Dung's grounded semantics is strategy-proof (i.e. truth enforcing). But these early results dealt with a highly restricted form of agent preferences, and assumed agents can only hide, but not lie about, arguments. In this paper, we characterise strategy-proofness under grounded semantics for a more realistic preference class (namely, focal arguments). We also provide the first analysis of the case where agents can lie.

1 Introduction

Argumentation has recently become one of the key approaches to automated reasoning and rational interaction in Artificial Intelligence [Bench-Capon and Dunne, 2007]. A key milestone has been Dung's landmark framework [Dung, 1995]. Arguments are viewed as abstract entities, with a binary defeat relation among them (resulting in a so-called *argument graph*). This view of argumentation enables high-level analysis while abstracting away from the internal structure of individual arguments. Much research has been done on defining criteria (so-called semantics) for evaluating complex argument structures [Baroni and Giacomin, 2007].

However, most research that employs Dung's approach discounts the fact that argumentation is often a multi-agent, adversarial process. Thus, the outcome of argumentation is determined not only by the rules by which arguments are evaluated, but also by the strategies employed by the agents who present these arguments. As these agents may be self-interested, they may have conflicting preferences over which arguments end up being accepted. As such, the design of the argument evaluation rule should take the mechanism design perspective [Mas-Colell *et al.*, 1995, Ch 23]: *what game rules guarantee a desirable social outcome when each self-interested agent selects the best strategy for itself?*

Recently, we introduced *Argumentation Mechanism Design* (ArgMD) as a framework for analysing the strategic incentives in argumentation and applied it to the well-known grounded semantics [Rahwan and Larson, 2008]. However, this preliminary analysis focused on a rather peculiar form of agent preferences: each agent wishes to get as many of its arguments accepted as possible. Moreover, they assumed agents can only hide, but not lie about, arguments.

In this paper, we apply the ArgMD framework to a more natural form of agent preferences, namely situations in which each agent has a single focal argument it wishes to have accepted. We provide a full characterisation of the strategy-proofness (i.e. truth-telling being a dominant strategy equilibrium) under grounded semantics when agents both hide and/or lie about arguments. We also provide intuitive, sufficient graph-theoretic conditions for strategy-proofness.

The paper advances the state-of-the-art in the computational modelling of argumentation in two major ways. Firstly, it provides the first comprehensive analysis of strategic incentives under grounded semantics when agents have focal arguments. This is a much more realistic (and common) form of agent preferences than the only other analysis undertaken to-date for grounded semantics (by [Rahwan and Larson, 2008]). Secondly, the paper provides the first analysis of incentives when agents can lie in argumentation. This is important since it shows that the ArgMD approach can be extended to such more realistic cases.

2 Background

We now briefly outline some of key elements of abstract argumentation frameworks. We begin with Dung's abstract characterisation of an argumentation system [Dung, 1995].

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* $\alpha \rightharpoonup \beta$*).*[1]

An argumentation framework can be represented as a directed graph in which vertices are arguments and directed arcs characterise defeat among arguments.

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

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]We restrict ourselves to finite sets of arguments.

- 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. 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* $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$.

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

Definition 4 (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.*

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 (hence the *least* fixed point of $\mathcal{F}$). There always exists a unique grounded extension.

Definition 5 (Indirect defeat and defence [Dung, 1995]). *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* $\alpha \nrightarrow \beta$*, if and only if there is an even-length path (with non-zero length) from* α *to* β *in the argument graph.*

Finally, the set of *initial* arguments, denoted $IN(AF)$, contains all arguments that have no defeaters.

3 Argumentation Mechanism Design

In this section we briefly define the mechanism design problem for abstract argumentation, as introduced by [Rahwan and Larson, 2008]. In particular, we specify the agents' type spaces and utility functions, what sort of strategic behaviour agents might indulge in, as well as the kinds of social choice functions we are interested in implementing.

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 of I self-interested agents. 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.*

An agent's type can be seen as a reflection of its expertise or domain knowledge. For example, medical experts may only be able to comment on certain aspects of forensics in a legal case, while a defendant's family and friends may be able to comment on his/her character. Also, such expertise may overlap, so agent types are not necessarily disjoint.

A social choice function f maps a type profile (agent type vector) into a subset of arguments. It specifies *the arguments the judge would wish to accept if he knew all actual arguments*.

$$f : 2^{\mathcal{A}} \times \ldots \times 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$$

We will be particularly interested in *argument acceptability* social choice functions.

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* $(\mathcal{A}_1, \ldots, \mathcal{A}_I)$*, the* argument acceptability social choice function f *is defined as the set of acceptable arguments given the semantics* $\mathcal{S}$*. That is,*

$$f(\mathcal{A}_1, \ldots, \mathcal{A}_I) = Acc(\langle A_1 \cup \ldots \cup \mathcal{A}_I, \rightharpoonup \rangle, \mathcal{S}).$$

As is standard in the mechanism design literature, we assume that agents have preferences over the outcomes $o \in 2^{\mathcal{A}}$, represented in utility functions: $u_i(o, \mathcal{A}_i)$ denotes agent i's utility for outcome o when its type is argument set $\mathcal{A}_i$.

Agents may not have incentive to reveal their true types because they may be able to influence the status of arguments and thus obtain higher utility. On one hand, an agent might *hide* some of its arguments, e.g. to break defeat chains in the argument framework, thus changing the final set of acceptable arguments. Alternatively, an agent might *lie* by making-up new arguments that it does not have in its argument set. ArgMD aims to obtain the desired outcome (as per the social choice function) despite the potential for such manipulations.

A strategy of an agent specifies a complete plan that describes what action the agent takes for every decision that a player might be called upon to take, for every piece of information that the player might have at each time that it is called upon to act. In our model, the actions available to an agent involve announcing arguments according to some protocol. Thus a strategy, $s_i \in \Sigma_i$ for agent i (where Σ_i is i's strategy space) would specify for each possible subset of arguments that could define its type, what set of arguments to reveal. An agent's strategy space specifies all its possible strategies.

Definition 8 (Argumentation Mechanism). *Given an argumentation framework* $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ *and semantics* $\mathcal{S}$*, an* argumentation mechanism *is defined as*

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

where $g : \Sigma_1 \times \ldots \Sigma_I \rightarrow 2^{\mathcal{A}}$.

Note that in the above definition, the notion of dialogue strategy is broadly construed and would depend on the argumentation protocol. In a *direct* mechanism, however, the strategy spaces of the agents are restricted so that they can only reveal a subset of arguments indicating its (alleged) type

–that is, $\Sigma_i = 2^{\mathcal{A}}$. We focus here on direct mechanisms since, according to the revelation principle [Mas-Colell *et al.*, 1995, Ch 23], any equilibrium of an indirect mechanism has an equivalent truthful direct mechanism. This approach is common in the mechanism design literature, since it greatly simplifies analysis without losing generality.

In Table 1, we summarise the mapping of multi-agent abstract argumentation as an a mechanism design problem.

We now present a direct mechanism for argumentation based on grounded semantics. The mechanism calculates the grounded extension given the arguments revealed by agents. We will refer to a specific action (i.e. set of declared arguments) as $\mathcal{A}_i^\circ \in \Sigma_i$.

Definition 9 (Grounded Direct Argumentation Mechanism). *A* grounded direct argumentation mechanism *for argumentation framework* $\langle \mathcal{A}, \rightharpoonup \rangle$ *is* $\mathcal{M}_{AF}^{grnd} = (\Sigma_1, \ldots, \Sigma_I, g(.))$ *where:*

- $\Sigma_i \in 2^{\mathcal{A}}$ *is the set of strategies available to each agent;*
- $g : \Sigma_1 \times \cdots \times \Sigma_I \rightarrow 2^{\mathcal{A}}$ *is an outcome rule defined as:* $g(\mathcal{A}_1^\circ, \ldots, \mathcal{A}_I^\circ) = Acc(\langle \mathcal{A}_1^\circ \cup \cdots \cup \mathcal{A}_I^\circ, \rightharpoonup \rangle, \mathcal{S}^{grnd})$ *where* $\mathcal{S}^{grnd}$ *denotes sceptical grounded acceptability semantics.*

4 Agents with Focal Arguments

Earlier [Rahwan and Larson, 2008], we analysed the grounded argumentation mechanism under a highly restrictive form of agent preferences, called *acceptability maximising preference*: each agent wishes to get as many of its arguments accepted as possible. This is rarely seen in practice.

In many realistic dialogues, each agent i is interested in the acceptance of a particular argument $\hat{\alpha}^i \in \mathcal{A}_i$, which we call the *focal argument* of agent i. Here, other arguments in $\mathcal{A}_i \backslash \{\hat{\alpha}^i\}$ can merely be *instrumental* towards the acceptance of the focal argument. We are interested in characterising conditions under which $\mathcal{M}_{AF}^{grnd}$ is strategy-proof for scenarios in which each agent has a focal argument (Other preference criteria are also reasonable, such as wanting to win any argument from a set that support the same conclusion).

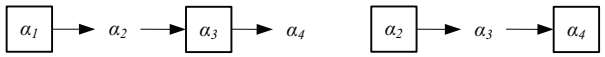
Definition 10 (Focal Argument for an Agent). *An agent i has a* focal argument $\hat{\alpha}^i \in \mathcal{A}_i$ *if and only if* $\forall o_1, o_2 \in \mathcal{O}$ *such that* $\hat{\alpha}^i \in o_1$ *and* $\hat{\alpha}^i \notin o_2$*, we have* $u_i(o_1, \mathcal{A}_i) > u_i(o_2, \mathcal{A}_i)$*, otherwise* $u_i(o_1, \mathcal{A}_i) = u_i(o_2, \mathcal{A}_i)$.

Let $o \in \mathcal{O}$ be an arbitrary outcome. If $\hat{\alpha}^i \in o$, we say that agent i *wins* in outcome o. Otherwise, i *loses* in outcome o.

5 When Agents can Hide Arguments

In this section, following our earlier work [Rahwan and Larson, 2008], we assume that there is an *external verifier* that is capable of checking whether it is possible for a particular agent to actually make a particular argument. Informally, this means that presented arguments, while still possibly defeasible, must at least be based on some sort of demonstrable 'plausible evidence.' If an agent is caught making up arguments then it will be removed from the mechanism. For example, in a court of law, any act of perjury by a witness is punished, at the very least, by completely discrediting all evidence produced by the witness. Moreover, in a court of law, arguments presented without any plausible evidence are normally discarded (e.g. "*I did not kill him, since I was abducted by aliens at the time of the crime!*"). For all intents and purposes this assumption removes the incentive for an agent to make things up.

To investigate whether mechanism $\mathcal{M}_{AF}^{grnd}$ is strategy-proof for any argumentation framework for agents with focal arguments, consider the following example.



(a) Graph with full revelation (b) Graph with α_1 withheld

Figure 1: Hiding an argument is beneficial

Example 1. *Consider a grounded direct argumentation mechanism with agents x, y and z with types $\mathcal{A}_x = \{\alpha_1, \alpha_4\}$, $\mathcal{A}_y = \{\alpha_2\}$ and $\mathcal{A}_z = \{\alpha_3\}$ respectively, and with focal arguments defined as follows: $\hat{\alpha}^x = \alpha_4$; $\hat{\alpha}^y = \alpha_2$; $\hat{\alpha}^z = \alpha_3$. Let the defeat relation be defined as follows: $\rightharpoonup = \{(\alpha_1, \alpha_2), (\alpha_2, \alpha_3), (\alpha_3, \alpha_4)\}$. If agents reveal all their arguments, we have the graph shown in Figure 1(a), with the accepted arguments marked by boxes. Here, agent z is the only winner.*

It turns out that the mechanism is susceptible to strategic manipulation, even if we suppose that agents do not lie by making up arguments (i.e., they may only withhold some arguments). In this case, for both agents y and z, revealing their true types weakly dominates revealing nothing at all (since hiding their single focal arguments can only guarantee their respective loss). However, it turns out that agent x is better off only revealing $\{\alpha_4\}$. By withholding α_1, the resulting argument network becomes as depicted in Figure 1(b). Under this outcome, x wins, which is better for x than truth-revealation.

Remark 1. *Given an arbitrary argumentation framework AF and agents with focal arguments, mechanism $\mathcal{M}_{AF}^{grnd}$ is* not *strategy-proof.*

Having established this property, the natural question to ask is whether mechanism $\mathcal{M}_{AF}^{grnd}$ is strategy-proof under some conditions. The following theorem provides a full characterisation of strategy-proof mechanisms for sceptical argumentation frameworks, for agents with focal arguments, when hiding arguments is possible. Note that $\mathcal{A}_{-i}$ denotes arguments of all agents *other than* agent i.

Theorem 1. *Let AF be an arbitrary argumentation framework, and let $GE(AF)$ denote its grounded extension. Mechanism $\mathcal{M}_{AF}^{grnd}$ is strategy-proof for agents with focal arguments if and only if AF satisfies the following condition: $\forall i \in I, \forall S \subseteq \mathcal{A}_i$ and $\forall \mathcal{A}_{-i}$, we have $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle (\mathcal{A}_i \backslash S) \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.*

Proof. $\Rightarrow$) Let $i \in I$ be an arbitrary agent with type $\mathcal{A}_i$ and focal argument $\hat{\alpha}^i \in \mathcal{A}_i$. Suppose $\mathcal{M}_{AF}^{grnd}$ is strategy-proof. This implies that $\forall S \subseteq \mathcal{A}_i$ and $\forall \mathcal{A}_{-i}$:

$$u_i(GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i) \geq u_i(GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i)$$

MD Concept	ArgMD Instantiation
Agent type $\theta_i \in \Theta_i$	Agent's arguments $\theta_i = \mathcal{A}_i \subseteq \mathcal{A}$
Outcome $o \in \mathcal{O}$ where $\mathcal{O}$ is the set of possible outcomes.	Accepted arguments $Acc(.) \subseteq \mathcal{A}$
Utility $u_i(o, \theta_i)$	Preferences over $2^{\mathcal{A}}$ (what arguments end up being accepted)
Social choice function $f : \Theta_1 \times \ldots \times \Theta_I \rightarrow \mathcal{O}$	$f(\mathcal{A}_1, \ldots, \mathcal{A}_I) = Acc(\langle A_1 \cup \ldots \cup \mathcal{A}_I, \rightharpoonup \rangle, \mathcal{S})$. by some argument acceptability criterion
Mechanism $\mathcal{M} = (\Sigma, g(\cdot))$ where $\Sigma = \Sigma_1 \times \cdots \times \Sigma_I$ and $g : \Sigma \rightarrow \mathcal{O}$	Σ_i is an argumentation strategy, $g : \Sigma \rightarrow 2^{\mathcal{A}}$
Direct mechanism: $\Sigma_i = \Theta_i$	$\Sigma_i = 2^{\mathcal{A}}$ (every agent reveals a set of arguments)
Truth revelation	Revealing $\mathcal{A}_i$

Table 1: Abstract argumentation as a mechanism

Therefore, by definition of the focal argument: if $\hat{\alpha}^i \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ then $\hat{\alpha}^i \in GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Then, by contraposition we have that:

$\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

$\Leftarrow$) Suppose that given any $\mathcal{A}_{-i}$, we have that $\forall i \in I, \forall S \subseteq \mathcal{A}_i$, $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

We want to prove that:

$u_i(GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i) \geq u_i(GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i)$.

Suppose not. Then $\exists i$ and $\exists S^{'} \subseteq \mathcal{A}_i$ such that

$u_i(GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i) < u_i(GE(\langle S^{'} \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i)$.

But this means that $\hat{\alpha}^i \in GE(\langle S^{'} \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ while $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Contradiction. Therefore, i has no incentive to declare any arguments other than those of her type, and thus the mechanism is strategy-proof. □

This result is consistent with the literature on mechanism design. While general strategy-proof results obtain only at the cost of dropping other desirable properties (like *non-dictatorship*, as per the Gibbard-Satterthwaite theorem [Mas-Colell *et al.*, 1995, Ch 23]), positive results obtain by restricting the domain of types on which the mechanism is applied (e.g. restriction to quasi-linear preferences [Mas-Colell *et al.*, 1995, Ch 21]).

Although the above theorem gives us a full characterisation, it is difficult to apply in practice. In particular, the theorem does not give us an indication of how agents (or the mechanism designer) can identify whether the mechanism is strategy-proof for a class of argumentation frameworks by appealing to explicit graph-theoretic properties. Below, we provide such analysis. But before we can do this, we present the following lemma. This lemma explores what happens when we add a new argument (and its associated defeats) to a given argumentation framework, thus resulting in a new argumentation framework. In particular, we are interested in conditions under which arguments acceptable in the first framework are also accepted in the second. We show that this is true under the condition that the new argument does not indirectly defeat arguments acceptable in the first framework.

Lemma 1 ([Rahwan and Larson, 2008]). *Let $AF_1 = \langle \mathcal{A}, \rightharpoonup_1 \rangle$ 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 grounded extension of AF_1 and α' does not indirectly defeat α, then α is also in the grounded extension of AF_2.*

With the above lemma in place, we now provide an intuitive, graph-theoretic condition that is sufficient to ensure that $\mathcal{M}_{AF}^{grnd}$ is strategy-proof when agents have focal arguments.

Theorem 2. *Suppose every agent $i \in I$ has a focal argument $\hat{\alpha}^i \in \mathcal{A}_i$. If each agent's type contains no (in)direct defeat against $\hat{\alpha}^i$ (formally $\forall i \in I, \nexists \alpha \in \mathcal{A}_i$ such that $\alpha \hookrightarrow \hat{\alpha}^i$), then $\mathcal{M}_{AF}^{grnd}$ is strategy-proof.*

Proof. Let $\mathcal{A}'_{-i} = (\mathcal{A}'_1, \ldots, \mathcal{A}'_{i-1}, \mathcal{A}'_{i+1}, \ldots, \mathcal{A}'_I)$ be arbitrary revelations from all agents not including i. We will show that agent i is always best off revealing $\mathcal{A}_i$. That is, no matter what sets of arguments the other agents reveal, agent i is best off revealing its full set of arguments. Formally, we will show that $\forall i \in I$ $u_i(Acc(\langle \mathcal{A}'_1 \cup \cdots \cup \mathcal{A}_i \cup \cdots \cup \mathcal{A}'_I, \rightharpoonup \rangle, \mathcal{S}^{grnd}), \mathcal{A}_i) \geq u_i(Acc(\langle \mathcal{A}'_1 \cup \cdots \cup \mathcal{A}^{\circ}_i \cup \cdots \cup \mathcal{A}'_I, \rightharpoonup \rangle, \mathcal{S}^{grnd}), \mathcal{A}_i)$ for any $\mathcal{A}^{\circ}_i \subset \mathcal{A}_i$.

We use induction over the sets of arguments agent i may reveal, starting from the focal argument $\hat{\alpha}^i$ (note that any strategy that does not reveal $\hat{\alpha}^i$ can be safely ignored). 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 $\hat{\alpha}^i$ accepted, i.e. it (weakly) improves 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 additional argument can increase, but never decrease the agent's utility. In other words, 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 $\mathcal{A}''_i$.

Let α' where $\{\alpha'\} = \mathcal{A}'_i - \mathcal{A}''_i$ be the new argument.

Suppose the focal argument $\hat{\alpha}^i$ is in the grounded extension when revealing $\mathcal{A}''_i$ (formally $\hat{\alpha}^i \in Acc(\langle \mathcal{A}'_1 \cup \cdots \cup \mathcal{A}''_i \cup \cdots \cup \mathcal{A}'_I, \rightharpoonup \rangle, \mathcal{S}^{grnd})$). We need to show that after adding α', argument $\hat{\alpha}^i$ remains in the grounded extension. Formally, we need to show that $\hat{\alpha}^i \in \mathcal{A}'_i \cap Acc(\langle \mathcal{A}'_1 \cup \cdots \cup \mathcal{A}'_i \cup \cdots \cup \mathcal{A}'_I, \rightharpoonup \rangle, \mathcal{S}^{grnd})$. This is true from Lemma 1, and from the fact that $\mathcal{A}_i$ does not include indirect defeats against $\hat{\alpha}^i$.

Thus, by induction, revealing the full set $\mathcal{A}_i$ weakly dominates revealing any sub-set thereof. □

Note that in the theorem, $\hookrightarrow$ is over all arguments in $\mathcal{A}$. Intuitively, to guarantee the strategy-proof property for agents with focal arguments, it suffices that no (in)direct defeats exist from an agent's own arguments to its focal argument. Said differently, each agent i's arguments must *not* undermine its

own focal argument, neither *explicitly* and *implicitly*. By 'explicitly,' we mean that none of i's own arguments can defeat its focal argument. By 'implicitly,' we mean that other agents cannot possibly present a sequence of arguments that reveal an indirect defeat between i's own arguments and its focal argument. More concretely, in Example 1 and Figure 1(a), while agent x's argument set $\mathcal{A}_x = \{\alpha_1, \alpha_4\}$ is conflict-free, when agents y and z presented their own arguments α_2 and α_3, they revealed an implicit conflict between x's arguments and x's focal argument. In other words, they showed that x contradicts himself (i.e. committed a *fallacy* of some kind).

An important observation is that under the condition in Theorem 2, we need not assume that the actual set of possible presentable arguments is common knowledge. To ensure the strategy-proof property, each agent only needs to know that indirect defeat chains cannot arise from any of its arguments to its focal argument.

One may reasonably ask if the *sufficient* condition in Theorem 2 is also *necessary* for agents to reveal all their arguments truthfully. As Example 2 shows, this is not the case. In particular, for certain argumentation frameworks, an agent may have truthtelling as a dominant strategy despite the presence of indirect defeats among its own arguments.

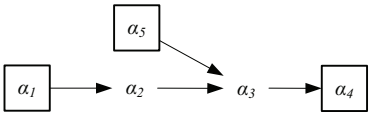


Figure 2: Strategy-proofness despite indirect self-defeat

Example 2. *Consider the variant of Example 1 with the additional argument α_5 and defeat (α_5, α_3). Let the agent types be $\mathcal{A}_x = \{\alpha_1, \alpha_4, \alpha_5\}$, $\mathcal{A}_y = \{\alpha_2\}$ and $\mathcal{A}_z = \{\alpha_3\}$ respectively. The full argument graph is depicted in Figure 2. With full revelation, the mechanism outcome rule produces the outcome $o = \{\alpha_1, \alpha_4, \alpha_5\}$.*

Note that in Example 2, truth revelation is now a dominant strategy for x despite the fact that $\alpha_1 \hookrightarrow \alpha_4$ (note that here, x gains nothing by hiding α_1). This hinges on the presence of an argument (namely α_5) that cancels out the negative effect of the (in)direct self-defeat among x's own arguments.

6 When Agents can Hide or Lie

In the previous section, we restricted agent strategies to showing or hiding arguments in their own type. We did not allow agents to reveal arguments that are outside of their types. That is, agents were not allowed to lie by stating something they did not know, but only by hiding something they do know. This is the approach taken originally by us [Rahwan and Larson, 2008].

In this section, we investigate (for the first time) strategy-proofness of grounded mechanisms without this assumption. We first show that the characterization of strategy-proofness is identical to that when agents could only hide arguments (the only difference is that S ranges over $\mathcal{A}$ instead of $\mathcal{A}_i$).

Theorem 3. *Let AF be an arbitrary argumentation framework, and let $GE(AF)$ denote its grounded extension. Mechanism $\mathcal{M}_{AF}^{grnd}$ is strategy-proof for agents with focal arguments if and only if AF satisfies the following condition: $\forall i \in I, \forall S \subseteq \mathcal{A}$ and $\forall \mathcal{A}_{-i}$, we have $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.*

Proof. The proof is essentially the same as the proof of Theorem 1, such that S ranges over $\mathcal{A}$ instead of $\mathcal{A}_i$. □

As we did in the case of hiding, this result can be weakened to yield a more intuitive sufficient condition for strategy-proofness.

Theorem 4. *Suppose every agent $i \in I$ has a focal argument $\hat{\alpha}^i \in \mathcal{A}_i$. If the following conditions hold:*

- **(A)** *no agent type contains (in)direct defeat against its focal argument (formally $\forall i \in I, \nexists \beta \in \mathcal{A}_i$ such that $\beta \hookrightarrow \hat{\alpha}^i$);*
- **(B)** *no argument outside any agent's type (in)directly defends its focal argument (formally $\forall i \in I, \nexists \beta \in \mathcal{A} \backslash \mathcal{A}_i$ such that $\beta \hookrightarrow \hat{\alpha}^i$);*

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

Proof. What we want to prove is that $\forall i, \forall \mathcal{A}_{-i}$ and $\forall S \neq \mathcal{A}_i$, if **(A)** and **(B)** hold, then

$$u_i(GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i) \geq u_i(GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle), \mathcal{A}_i).$$

Following the definition of focal arguments, our goal above can be rephrased as proving, for any arbitrary $S \neq \mathcal{A}_i$, that:

$$\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle) \text{ implies } \hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle).$$

Suppose $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$, and let us show that $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

We do this by showing, recursively, that for a given $\hat{\alpha}^i$, there must exist a $\beta \rightharpoonup \alpha$ such that for all $z \rightharpoonup \beta$:
$z \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ then $z \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$
implies $\alpha \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ then $\alpha \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

Base Step: From $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$, it follows that $\exists \beta \in \mathcal{A}_i \cup \mathcal{A}_{-i}, \beta \rightharpoonup \hat{\alpha}^i$ for which $\nexists z \in GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z \rightharpoonup \beta$. Let $\beta^1 = \beta$ be such a defeater.

By assumption **(A)**, $\beta^1 \notin \mathcal{A}_i$ (since otherwise, we would have $\beta^1 \rightharpoonup \hat{\alpha}^i$ and therefore $\beta^1 \hookrightarrow \hat{\alpha}^i$ for some $\beta^1 \in \mathcal{A}_i$). But since $\beta^1 \in \mathcal{A}_i \cup \mathcal{A}_{-i}$, we conclude that $\beta^1 \in \mathcal{A}_{-i}$. This in turn implies that $\beta^1 \in S \cup \mathcal{A}_{-i}$. In other words, the defeaters of $\hat{\alpha}^i$ given action profile $\mathcal{A}_i \cup \mathcal{A}_{-i}$ are preserved when the agent changes the action profile to $S \cup \mathcal{A}_{-i}$.

We will now show that $\nexists z \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z \rightharpoonup \beta^1$.

Let $z^1 \in S \cup \mathcal{A}_{-i}$, such that $z^1 \rightharpoonup \beta^1$, be an arbitrary defeater of β^1 when agent i lies. By assumption **(B)**, we conclude that $z^1 \notin \mathcal{A} \backslash \mathcal{A}_i$ (since otherwise, we would have $z^1 \rightharpoonup \beta^1 \rightharpoonup \hat{\alpha}^i$ and therefore $z \hookrightarrow \hat{\alpha}^i$ for some $z \in \mathcal{A} \backslash \mathcal{A}_i$). This in turn implies that $z^1 \in \mathcal{A}_i$. In other words, no new defenders of $\hat{\alpha}^i$ can be introduced when the agent moves from action profile $\mathcal{A}_i \cup \mathcal{A}_{-i}$ to action profile $S \cup \mathcal{A}_{-i}$. Therefore, since we already know that $z^1 \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$, z^1 cannot be in $GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Then:

$$z^1 \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle) \text{ implies } z^1 \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle) \quad (*)$$

Suppose now that $\hat{\alpha}^i \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. This would mean that for every $\beta \in \mathcal{A}_i \cup \mathcal{A}_{-i}$, $\beta \rightharpoonup \hat{\alpha}^i$, $\exists z \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z \rightharpoonup \beta$, but then, for β^1 there should exist a z^1 satisfying this condition. But by $(*)$, no $z^1 \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Therefore, we have from $(*)$ that:

$\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

Recursive Step: Let $z^k \in \mathcal{A}_i \cup \mathcal{A}_{-i}$ be an arbitrary argument such that $z^k \rightharpoonup \beta^k$ for some β^k with $\beta^k \hookrightarrow \hat{\alpha}^i$. Assume that $z^k \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. We will show that $z^k \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

From $z^k \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$, it follows that $\exists \beta \in \mathcal{A}_i \cup \mathcal{A}_{-i}$ such that $\beta \rightharpoonup z^k$ while $\nexists z^{k+1} \in GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Let $\beta^{k+1} = \beta$ be such a defeater.

By assumption **(A)**, $\beta^{k+1} \notin \mathcal{A}_i$ (since otherwise, we would have $\beta^{k+1} \hookrightarrow \hat{\alpha}^i$ for some $\beta^{k+1} \in \mathcal{A}_i$). But since $\beta^{k+1} \in \mathcal{A}_i \cup \mathcal{A}_{-i}$, we conclude that $\beta^{k+1} \in \mathcal{A}_{-i}$. This in turn implies that $\beta^{k+1} \in S \cup \mathcal{A}_{-i}$. In other words, the defeaters of z^{k+1} given action profile $\mathcal{A}_i \cup \mathcal{A}_{-i}$ are preserved when the agent changes the action profile to $S \cup \mathcal{A}_{-i}$.

We will now show that $\nexists z \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z \rightharpoonup \beta^{k+1}$.

Let $z^{k+1} \in S \cup \mathcal{A}_{-i}$ such that $z^{k+1} \rightharpoonup \beta^{k+1}$ be an arbitrary defeater of β^{k+1} when agent i lies. By assumption **(B)**, we conclude that $z^{k+1} \notin \mathcal{A} \backslash \mathcal{A}_i$ (since otherwise, we would have $z^{k+1} \rightharpoonup \beta^{k+1} \hookrightarrow \hat{\alpha}^i$ and therefore $z \nrightarrow \hat{\alpha}^i$ for some $z \in \mathcal{A} \backslash \mathcal{A}_i$). This in turn implies that $z^{k+1} \in \mathcal{A}_i$. In other words, no new defenders of $\hat{\alpha}^i$ can be introduced when the agent moves from action profile $\mathcal{A}_i \cup \mathcal{A}_{-i}$ to action profile $S \cup \mathcal{A}_{-i}$. Therefore, since we already know that $z^{k+1} \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$, z^{k+1} cannot be either in $GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Then:

$$z^{k+1} \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle) \text{ implies } z^{k+1} \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle) \quad (**)$$

Suppose now that $z^k \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. This would mean that for every $\beta \in \mathcal{A}_i \cup \mathcal{A}_{-i}$, $\beta \rightharpoonup \hat{\alpha}^i$, $\exists z \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z \rightharpoonup \beta$, but then, for β^k there should exist a z^k satisfying this condition. But by $(**)$, no $z^{k+1} \in GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Therefore, we have that from $(**)$ it follows that:

$z^k \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $z^k \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

Recursion Termination: The recursion must eventually reach some $\beta^K \in IN(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ with $\beta^K \hookrightarrow \hat{\alpha}^i$. Let $z^K \in S \cup \mathcal{A}_{-i}$ such that $z^K \rightharpoonup \beta^K$ be an arbitrary defeater of β^K when agent i lies. By assumption **(B)**, we conclude that $z^K \notin \mathcal{A} \backslash \mathcal{A}_i$ (since otherwise, we would have $z^K \rightharpoonup \beta^K \hookrightarrow \hat{\alpha}^i$ and therefore $z \nrightarrow \hat{\alpha}^i$ for some $z \in \mathcal{A} \backslash \mathcal{A}_i$). But this in turn implies that $z^K \in \mathcal{A}_i$. But this contradicts with the fact that $\beta^K \in IN(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Hence, no such z^K exists and therefore $z^K \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Furthermore, we conclude that $\beta^K \notin IN(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ and thus $\beta^K \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Therefore, $z^K \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ such that $z^K \rightharpoonup \beta^K$. In summary:

$z^K \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ implies $z^K \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$.

By the above recursion, we conclude that, since $\hat{\alpha}^i \notin GE(\langle \mathcal{A}_i \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$ it follows that $\hat{\alpha}^i \notin GE(\langle S \cup \mathcal{A}_{-i}, \rightharpoonup \rangle)$. Therefore, $\mathcal{M}_{AF}^{grnd}$ is strategy-proof. □

Let us interpret the above theorem intuitively. The theorem rests on two key conditions: (a) that an agent cannot benefit from hiding any of its own arguments, because its arguments cannot "harm" its focal argument; and (b) that an agent cannot benefit from revealing any argument it does not have, because these arguments cannot "benefit" its focal argument. For the theorem to hold, these conditions must be satisfied for every agent, no matter what the other agents reveal. While this may appear obvious in hindsight, the precise proof is rather involved, and requires careful attention to the intricate aspects of the grounded semantics when applied to different related argument graphs – namely graphs corresponding to different strategy profiles that agents may choose to play. This level of attention to detail is essential for any thorough analysis of strategic argumentation.

7 Conclusion

ArgMD is a new paradigm for studying argumentation among self-interested agents using game-theoretic techniques. It provides a fresh perspective on the study of semantics for conflicting knowledge bases, especially when those are distributed among different entities (e.g. knowledge-based agents on the Semantic Web). While game-theoretic approaches have been applied extensively to resource allocation among agents (e.g. through auctions), no similar development has yet taken place for strategic aspects of interaction among *knowledge*-based agents.

The only other paper that uses ArgMD to-date [Rahwan and Larson, 2008] dealt with a highly restricted form of agent preferences, and assumed agents can only hide, but not lie about, arguments. In this paper, we showed how ArgMD can be applied to more realistic preferences and action spaces.

Future work includes analysing incentives under other varieties of agent preferences and other argumentation semantics.

Acknowledgments

We are grateful to the anonymous reviewers for the valuable advice, which helped us improve the paper.

References

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

[Bench-Capon and Dunne, 2007] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10–15):619–641, 2007.

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

[Mas-Colell *et al.*, 1995] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York NY, USA, 1995.

[Rahwan and Larson, 2008] Iyad Rahwan and Kate Larson. Mechanism design for abstract argumentation. In *Proceedings of AAMAS'2008*, pages 1031–1038, 2008.