A Characterisation of Strategy-Proofness for Grounded Argumentation Semantics

Iyad Rahwan\textsuperscript{1,2}, Kate Larson\textsuperscript{3}, and Fernando Töhmé\textsuperscript{4}
\textsuperscript{1}Faculty of Informatics, British University in Dubai, Dubai, UAE
\textsuperscript{2}(Fellow) School of Informatics, University of Edinburgh, UK
\textsuperscript{3}Cheriton School of Computer Science, University of Waterloo, Canada
\textsuperscript{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 $AF = \langle A, \rightarrow \rangle$ where $A$ is a set of arguments and $\rightarrow \subseteq A \times A$ is a defeat relation. We say that an argument $\alpha$ defeats an argument $\beta$ if $(\alpha, \beta) \in \rightarrow$ (sometimes written $\alpha \rightarrow_{\rightarrow} \beta$).

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

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

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

\footnote{We restrict ourselves to finite sets of arguments.}
• \( S \) is conflict-free if \( S \cap S^+ = \emptyset \).
• \( S \) defends argument \( \alpha \) if \( \alpha^- \subseteq S^+ \). We also say that argument \( \alpha \) 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 \( AF = \langle A, \vdash \rangle \) be an argumentation framework. The characteristic function of \( AF \) is \( F_{AF} : 2^A \rightarrow 2^A \) such that, given \( S \subseteq A \), we have \( F_{AF}(S) = \{ \alpha \in A \mid S \) defends \( \alpha \} \).

When there is no ambiguity about the argumentation framework in question, we will use \( F \) instead of \( F_{AF} \).

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

• \( S \) is admissible if it is conflict-free and defends every element in \( S \) (i.e. if \( S \subseteq F(S) \)).
• \( S \) is a complete extension if \( S = F(S) \).
• \( S \) 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 \( F \)). 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 \( F \)). There always exists a unique grounded extension.

**Definition 5 (Indirect defeat and defence [Dung, 1995]).** Let \( \alpha, \beta \in \mathcal{A} \). We say that \( \alpha \) indirectly defeats \( \beta \), written \( \alpha \leftarrow \beta \), if and only if there is an odd-length path from \( \alpha \) to \( \beta \) in the argument graph. We say that \( \alpha \) indirectly defends \( \beta \), written \( \alpha \rightarrow \beta \), if and only if there is an even-length path (with non-zero length) from \( \alpha \) to \( \beta \) 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 \( \langle A, \vdash \rangle \) with semantics \( S \), 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 \( \langle A, \vdash \rangle \), the type of agent \( i, A_i \subseteq A \), 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.

\[
f : 2^A \times \ldots \times 2^A \rightarrow 2^A
\]

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

**Definition 7 (Argument Acceptability Social Choice Functions).** Given an argumentation framework \( \langle A, \vdash \rangle \) with semantics \( S \), and given a type profile \( \langle A_1, \ldots, A_I \rangle \), the argument acceptability social choice function \( f \) is defined as the set of acceptable arguments given the semantics \( S \). That is,

\[
f(A_1, \ldots, A_I) = Acc(A_1 \cup \ldots \cup A_I , \vdash, S).
\]

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

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 \( \Sigma_i \) 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 (Argument Mechanism).** Given an argumentation framework \( AF = \langle A, \vdash \rangle \) and semantics \( S \), an argumentation mechanism is defined as

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

where \( g : \Sigma_1 \times \ldots \times \Sigma_I \rightarrow 2^A \).

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, \( \Sigma_i = 2^4 \). 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 \( A^i_{grnd} \).

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

- \( \Sigma_i \in 2^4 \) is the set of strategies available to each agent;
- \( g : \Sigma_1 \times \cdots \times \Sigma_I \rightarrow 2^4 \) is an outcome rule defined as: \( g(A^1_{grnd}, \ldots, A^I_{grnd}) = Acc(A^1 \cup \cdots \cup A^I, \rightarrow) \), \( S^{grnd} \) where \( S^{grnd} \) 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 \( \hat{\alpha} \in A_i \), which we call the focal argument of agent \( i \). Here, other arguments in \( A_i \setminus \{\hat{\alpha}\} \) can merely be instrumental towards the acceptance of the focal argument. We are interested in characterising conditions under which \( M^{grnd}_{AF} \) 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 conclusion).

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

Let \( o \in O \) be an arbitrary outcome. If \( \hat{\alpha} \in o \), 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 \( M^{grnd}_{AF} \), for any argumentation framework for agents with focal arguments, consider the following example.

![Figure 1](image_url)

**Figure 1:** Hiding an argument is beneficial

**Example 1.** Consider a grounded direct argumentation mechanism with agents \( x, y \) and \( z \) with types \( A_x = \{o_1, o_2\} \), \( A_y = \{o_2\} \) and \( A_z = \{o_3\} \) respectively, and with focal arguments defined as follows: \( \hat{\alpha}^x = o_1; \hat{\alpha}^y = o_2; \hat{\alpha}^z = o_3 \).

Let the defeat relation be defined as follows: \( \rightarrow = \{(o_1, o_2), (o_2, o_3), (o_3, o_4)\} \). 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 \( \{o_1\} \). By withholding \( o_1 \), the resulting argument network becomes as depicted in Figure 1(b). Under this outcome, \( x \) wins, which is better for \( x \) than truth-revelation.

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

Having established this property, the natural question to ask is whether mechanism \( M^{grnd}_{AF} \) 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 \( A_{\sim i} \) 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 \( M^{grnd}_{AF} \) is strategy-proof for agents with focal arguments if and only if \( AF \) satisfies the following condition:

\[ \forall i \in I, \forall S \subseteq A_i \text{ and } \forall A_{\sim i}, \text{ we have } \hat{\alpha}^i \notin GE((A_i \cup A_{\sim i}, \rightarrow)) \implies \hat{\alpha}^i \notin GE(((A_i \setminus S) \cup A_{\sim i}, \rightarrow)). \]

**Proof:** Let \( i \in I \) be an arbitrary agent with type \( A_i \) and focal argument \( \hat{\alpha}^i \in A_i \). Suppose \( M^{grnd}_{AF} \) is strategy-proof. This implies that \( \forall S \subseteq A_i \) and \( \forall A_{\sim i}, \)

\[ u_i(GE((A_i \cup A_{\sim i}, \rightarrow)), A_i) \geq u_i(GE((S \cup A_{\sim i}, \rightarrow)), A_i) \]
Table 1: Abstract argumentation as a mechanism

<table>
<thead>
<tr>
<th>MD Concept</th>
<th>ArgMD Instantiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent type $\theta_i \in \Theta_i$</td>
<td>Agent’s arguments $\theta_i = A_i \subseteq A$</td>
</tr>
<tr>
<td>Outcome $\alpha \in \Theta$ where $\Theta$ is the set of possible outcomes.</td>
<td>Accepted arguments $Acc(\theta_i) \subseteq A$</td>
</tr>
<tr>
<td>Utility $u_i(\theta_i, \theta_i)$</td>
<td>Preferences over $2^A$ (what arguments end up being accepted)</td>
</tr>
<tr>
<td>Social choice function $f : \Theta_1 \times \ldots \times \Theta_n \rightarrow \Theta$</td>
<td>by some argument acceptability criterion</td>
</tr>
<tr>
<td>Mechanism $\mathcal{M} = (\Sigma, g(\cdot))$ where $\Sigma = \Sigma_1 \times \ldots \times \Sigma_n$ and $g : \Sigma \rightarrow \Theta$</td>
<td>$\Sigma_i$ is an argumentation strategy, $g : \Sigma \rightarrow 2^A$</td>
</tr>
<tr>
<td>Direct mechanism: $\Sigma_i = \Theta_i$</td>
<td>(every agent reveals a set of arguments)</td>
</tr>
</tbody>
</table>

Therefore, by definition of the focal argument: if $\hat{\alpha} \in \text{GE}(\langle S \cup A_{-i}, \rightarrow \rangle)$ then $\hat{\alpha} \in \text{GE}(\langle A_i \cup A_{-i}, \rightarrow \rangle)$. Then, by contraposition we have that:

$\hat{\alpha} \not\in \text{GE}(\langle A_i \cup A_{-i}, \rightarrow \rangle)$ implies $\hat{\alpha} \not\in \text{GE}(\langle S \cup A_{-i}, \rightarrow \rangle)$.

Suppose that given any $A_{-i}$, we have that $\forall i \in I, \forall S \subseteq A_i$, $\hat{\alpha} \not\in \text{GE}(\langle A_i \cup A_{-i}, \rightarrow \rangle)$ implies $\hat{\alpha} \not\in \text{GE}(\langle S \cup A_{-i}, \rightarrow \rangle)$.

We want to prove that:

$u_i(GE(\langle A_i \cup A_{-i}, \rightarrow \rangle), A_i) \geq u_i(GE(\langle S \cup A_{-i}, \rightarrow \rangle), A_i)$.

Suppose not. Then $\exists i$ and $\exists S' \subseteq A_i$ such that

$u_i(GE(\langle A_i \cup A_{-i}, \rightarrow \rangle), A_i) < u_i(GE(\langle S' \cup A_{-i}, \rightarrow \rangle), A_i)$.

But this means that $\hat{\alpha} \in \text{GE}(\langle S' \cup A_{-i}, \rightarrow \rangle)$ while $\hat{\alpha} \not\in \text{GE}(\langle A_i \cup A_{-i}, \rightarrow \rangle)$. Contradiction. Therefore, $i$ has no incentive to declare any arguments other than those of her type, and thus the mechanism is strategy-proof.

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

**Theorem 2.** Suppose every agent $i \in I$ has a focal argument $\hat{\alpha} \in A_i$. If each agent’s type contains no (in)direct defeat against $\hat{\alpha}$ (formally $\forall i \in I, \forall \alpha \in A_i$ such that $\alpha \rightarrow \hat{\alpha}$), then $\mathcal{M}_F^{\text{grnd}}$ is strategy-proof.

**Proof.** Let $A_{-i} = \langle A'_1, \ldots, A'_{i-1}, A'_{i+1}, \ldots, A'_n \rangle$ be arbitrary revelations from all agents not including $i$. We will show that agent $i$ is always best off revealing $A_i$. That is, it is never better for the mechanism designer to 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 $AF_1 = \langle A_1, \rightarrow_1 \rangle$ and $AF_2 = \langle A \cup \{\alpha\}, \rightarrow_2 \rangle$ such that $\rightarrow_1 \subseteq \rightarrow_2$ and $\rightarrow_2 \setminus \rightarrow_1 \subseteq \{\alpha\} \times A \cup \{A \times \{\alpha\}\}$. If $\alpha$ is in the grounded extension of $AF_1$ and $\alpha$ does not indirectly defeat $\alpha$, then $\alpha$ is also in the grounded extension of $AF_2$.

Note that in the theorem, $\rightarrow_1$ is over all arguments in $A$. 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 $A_x = \{\alpha_1, \alpha_4\}$ is conflict-free, when agents $y$ and $z$ presented their own arguments $\alpha_2$ and $\alpha_3$, 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 truth-telling as a dominant strategy despite the presence of indirect defeats among its own arguments.

\[
\begin{array}{c}
\alpha_1 \\
\rightarrow \\
\alpha_2 \\
\rightarrow \\
\alpha_3 \\
\rightarrow \\
\alpha_4 \\
\end{array}
\]

Figure 2: Strategy-proofness despite indirect self-defeat

**Example 2.** Consider the variant of Example 1 with the additional argument $\alpha_5$ and defeat $(\alpha_3, \alpha_5)$. Let the agent types be $A_x = \{\alpha_1, \alpha_4, \alpha_5\}, A_y = \{\alpha_2\}$ and $A_z = \{\alpha_3\}$ respectively. The 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 $\alpha_1 \rightarrow \alpha_4$ (note that here, $x$ gains nothing by hiding $\alpha_1$). This hinges on the presence of an argument (namely $\alpha_5$) 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 hide 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 $A$ instead of $A_i$).

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

**Proof.** The proof is essentially the same as the proof of Theorem 1, such that $S$ ranges over $A$ instead of $A_i$.

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 $\hat{\alpha}^i \in A_i$. If the following conditions hold:

- (A) no agent type contains (in)direct defeat against its focal argument (formally $\forall i \in I, \beta \in A_i$ such that $\beta \rightarrow \hat{\alpha}^i$);
- (B) no argument outside any agent’s type (in)directly defends its focal argument (formally $\forall i \in I, \beta \in A_i\setminus A_i$ such that $\beta \nrightarrow \hat{\alpha}^i$);

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

**Proof.** What we want to prove is that $\forall i, \forall A_i - i$ and $\forall S \neq A_i$, if (A) and (B) hold, then $u_i(GE(\langle A_i \cup A_i - i, \rightarrow \rangle), A_i) \geq u_i(GE(\langle S \cup A_i - i, \rightarrow \rangle), A_i)$.

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

$\hat{\alpha}^i \notin GE(\langle A_i \cup A_i - i, \rightarrow \rangle)$ implies $\hat{\alpha}^i \notin GE(\langle S \cup A_i - i, \rightarrow \rangle)$.

Suppose $\hat{\alpha}^i \notin GE(\langle A_i \cup A_i - i, \rightarrow \rangle)$, and let us show that $\hat{\alpha}^i \notin GE(\langle S \cup A_i - i, \rightarrow \rangle)$.

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

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

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

We will now show that $\exists z \in GE(\langle S \cup A_i - i, \rightarrow \rangle)$ such that $z \rightarrow \beta^1$.

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

Then:

$z^1 \notin GE(\langle A_i \cup A_i - i, \rightarrow \rangle)$ implies $z^1 \notin GE(\langle S \cup A_i - i, \rightarrow \rangle)$ (*).
Suppose now that $\hat{\alpha}^i \notin GE((S \cup A_{-i}, \rightarrow))$. This would mean that for every $\beta \in \hat{\alpha}^i \rightarrow \hat{\alpha}^i$, $z \notin E((S \cup A_{-i}, \rightarrow))$ such that $z \rightarrow \beta$, but then, for $\beta$ there should exist a $z^1$ satisfying this condition. But by (4), no $z^1 \in GE((S \cup A_{-i}, \rightarrow))$. Therefore, we have from (4) that:

$\hat{\alpha}^i \notin GE((A_i \cup A_{-i}, \rightarrow))$ implies $\hat{\alpha}^i \notin GE((S \cup A_{-i}, \rightarrow))$.

**Recessive Step:** Let $z^k \in A_i \cup A_{-i}$ be an arbitrary argument such that $z^k \rightarrow \hat{\alpha}^i$ for some $\hat{\beta}^k \rightarrow \hat{\alpha}^i$. Assume that $z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$. We will show that $z^k \notin GE((S \cup A_{-i}, \rightarrow))$.

From $z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$, it follows that $\exists \beta \in A_i \cup A_{-i}$ such that $\beta \rightarrow z^k$ while $\beta z^{k+1} \in GE((A_i \cup A_{-i}, \rightarrow))$. Let $\beta z^{k+1} = \beta$ be such a defeater.

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

We will now show that $\beta z \in GE((S \cup A_{-i}, \rightarrow))$ such that $z \rightarrow \beta z^{k+1}$

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

$$z^{k+1} \notin GE((A_i \cup A_{-i}, \rightarrow)) \quad (**)$$

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

$z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$ implies $z^k \notin GE((S \cup A_{-i}, \rightarrow))$.

**Recursion Termination:** The recursion must eventually reach some $\beta^k \in IN(A_i \cup A_{-i}, \rightarrow))$ with $\beta^k \rightarrow \hat{\alpha}^i$. Let $z^k \in S \cup A_{-i}$ such that $z^k \rightarrow \beta^k$ be an arbitrary defeater of $\beta^k$ when agent $i$ lies. By assumption (B), we conclude that $z^k \notin A_i \cup A_{-i}$ (since otherwise, we would have $z^k \rightarrow \beta^k \rightarrow \hat{\alpha}^i$ and therefore $z \equiv \hat{\alpha}^i$ for some $z \in A_i \setminus A_i$). But this in turn implies that $z^k \in A_i$. But this contradicts with the fact that $\beta^k \notin IN(A_i \cup A_{-i}, \rightarrow))$. Hence, no such $z^k$ exists and therefore $z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$.

Furthermore, we conclude that $\beta^k \notin IN((S \cup A_{-i}, \rightarrow))$ and thus $\beta^k \notin GE((S \cup A_{-i}, \rightarrow))$. Therefore, $z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$ such that $z^k \rightarrow \beta^k$. In summary:

$z^k \notin GE((A_i \cup A_{-i}, \rightarrow))$ implies $z^k \notin GE((S \cup A_{-i}, \rightarrow))$.

By the above recursion, we conclude that, since $\hat{\alpha}^i \notin GE((A_i \cup A_{-i}, \rightarrow))$ it follows that $\hat{\alpha}^i \notin GE((S \cup A_{-i}, \rightarrow))$.

Therefore, $M_{AP}^{AP}$ 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


