Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Resistance to Corruption of Strategic Argumentation

Michael J. Maher
School of Engineering and Information Technology
UNSW, Canberra
E-mail: michael.maher@unsw.edu.au

Abstract

Strategic argumentation provides a simple model of disputation. We investigate it in the context of Dung's abstract argumentation. We show that strategic argumentation under the grounded semantics is resistant to corruption – specifically, collusion and espionage – in a sense similar to Bartholdi et al's notion of a voting scheme resistant to manipulation. Under the stable semantics, strategic argumentation is resistant to espionage, but its resistance to collusion varies according to the aims of the disputants. These results are extended to a variety of concrete languages for argumentation.

Introduction

Organizations have many mechanisms to discourage the risk of corruption of their processes by the individuals performing these processes: managerial oversight, transparency through audit trails, the presence of co-workers, random inspections, etc. (Bartholdi, Tovey, and Trick 1989) introduced a further way in which corruption is discouraged: the computational difficulty of determining what an individual must do to achieve a specific aim (in their problem, to alter the result of an election). In this paper, we adapt this approach for strategic argumentation, which provides a simple model of disputation and negotiation among agents.

A key source of intuition for this work is legal disputation. (Prakken and Sartor 1998) have argued persuasively that legal precedents can be represented by defeasible rules, and reasoned with using arguments constructed from the rules. Lawyers act as agents for their clients and, in particular, may present oral arguments to a judge in support of their client's position and to refute arguments by opposing lawyers. Such a process can be represented as a game where each player, in turn, asserts additional arguments that overcome their opponent's arguments and leave the game in a desired state. ((Prakken 2005) calls such dialogues *relevant*.) Eventually a winner emerges when one player is unable to refute the opponent's argument.

To model legal disputation requires a game of incomplete knowledge, since lawyers generally are not aware of all the arguments their opponent will employ. In theory, and implicit in the rules of such games, the knowledge of which arguments are available to a player is kept confidential, as a strategic advantage. However, in practice, a player may behave corruptly to violate this confidentiality. In particular, a player might learn of her opponent's arguments through espionage, or collude with her nominal opponent to achieve a pre-determined outcome. Such behaviour is against legal ethics, and in this paper we investigate whether strategic argumentation is resistant to this form of corruption, that is, whether it is computationally difficult to exploit the corrupt behaviour.

The games we consider in this paper are two-player adversarial games. In each move a player adds entire arguments, rather than single rules, to the dispute. These games also require a player to commit to the arguments she plays, in the sense that there is no ability to retract an argument. The only way an argument loses force is as a result of it being attacked by other arguments. We call playing such games *strategic argumentation*, following (Governatori et al. 2014b). These games are quite different – in purpose and in technical detail, as discussed above – from dialogue games that are used to provide an operational interpretation of defeasible rules and argumentation (for example, (Prakken and Sartor 1998; Prakken 2005; Thang, Dung, and Hung 2012), among many).

It might appear more natural to formulate strategic argumentation games in concrete terms, with precedents represented by defeasible rules, as argued by (Prakken and Sartor 1998). Indeed (Maher 2014) does exactly this, using the defeasible logic DL (Antoniou et al. 2001). However, we choose to first formulate them in terms of Dung's model of abstract argumentation (Dung 1995). Although this has the danger that important features are abstracted away, we will show that results obtained in the abstract setting can be extended to a wide range of concrete formalisms supporting defeasible rules.

We address two semantics for abstract argumentation: the grounded and stable semantics (Dung 1995). These semantics seem the most natural for adjudicating disputes. The grounded semantics accepts arguments that (iteratively) are only attacked by defeated arguments and hence are relatively uncontroversial. The stable semantics consists of those adjudications where each argument is accepted or rejected. Furthermore, most concrete languages expressing defeasible rules reflect one of these semantics.

In the next section we define strategic abstract argumentation, and the decision problems that arise from the play of an argumentation game, including those that arise when exploiting corrupt behaviour. The next two sections then establish the complexity of these problems under, respectively, the grounded and the stable semantics. These allow us to reach conclusions on the resistance to corruption of strategic abstract argumentation. These results are then extended to many concrete formalisms. For lack of space, proofs are either sketched or omitted.

Background

Our work is based on abstract argumentation in the sense of (Dung 1995), which addresses the evaluation of a static set of arguments. An *argumentation framework* $\mathcal{A} = (S, \gg)$ consists of a set of arguments S and a binary relation $\gg$ over S, called the attack relation. The semantics of an argumentation framework is given in terms of *extensions*, which are subsets of S.

Given an argumentation framework, an argument a is said to be *accepted* in an extension E if $a \in E$, and said to be *rejected* in E if some $b \in E$ attacks a. An extension E is *conflict-free* if the restriction of $\gg$ to E is empty. An argument a is *defended* by E if every argument that attacks a is attacked by some argument in E. An extension E of $\mathcal{A}$ is *complete* if it is conflict-free and, $a \in E$ iff a is defended by E. The complete extension under the containment ordering exists and is called the *grounded* extension. An extension E of $\mathcal{A}$ is *stable* if it is conflict-free and for every argument $a \in S \backslash E$ there is an argument in E that attacks a. Every stable extension is complete, and so is a superset of the grounded extension.

In this paper we consider only two semantics of argumentation frameworks: the grounded semantics, consisting of the grounded extension, and the stable semantics, consisting of the stable extensions. Complexity results for acceptance under these semantics (among others) are presented in (Dunne and Wooldridge 2009).

An argumentation framework is *well-founded* if there is no infinite sequence of arguments $a_1, a_2, \ldots, a_i, a_{i+1}, \ldots$ such that, for each i, a_{i+1} attacks a_i. Such argumentation frameworks have a single complete extension, which must be the grounded extension and the sole stable extension (Dung 1995).

Strategic Abstract Argumentation

Strategic argumentation provides a simple model of dynamic argumentation, where players take turns to add arguments to the argumentation framework. It is formalized as follows. We assume there are two players, a proponent P and her opponent O. A *split argumentation framework* $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ consists of three sets of arguments: $\mathcal{A}_{Com}$ the arguments that are common knowledge to P and O; $\mathcal{A}_P$ the arguments available to P, and $\mathcal{A}_O$ the arguments available to O; and an attack relation $\gg$ over $\mathcal{A}_{Com} \cup \mathcal{A}_P \cup \mathcal{A}_O$. $\mathcal{A}_P$ is assumed to be unknown to O, and $\mathcal{A}_O$ is unknown to P. Each player is aware of $\gg$ restricted to the arguments they know. Each player has a *desired outcome* or *aim*, and usually the desired outcomes of P and O conflict. Starting with P, the players take turns in adding sets of arguments to $\mathcal{A}_{Com}$ from their available arguments, ensuring that their desired outcome is a consequence of the resulting argumentation framework[1]. As play continues, the set of arguments that are common knowledge $\mathcal{A}_{Com}$ becomes larger. When a player is unable to achieve her aim, she loses[2].

We follow the convention that P's desired outcome is to have a distinguished argument a accepted, in some sense, while O's aim is to prevent this. The notion of desired outcome may vary, depending on the argumentation semantics and the attitude of the player. For example, under the grounded semantics, P might desire that a is accepted in the grounded extension, while O desires that a is not accepted, but O might aim for the stronger outcome that a is rejected.

Note that an argument played by P (say) may attack other arguments in $\mathcal{A}_P$, which might otherwise have been used later to attack arguments played by O. Thus the game truly is strategic in nature[3]. Furthermore, it is preferable, in general, to play as few arguments as possible, to retain as much strategic advantage as possible from the confidentiality of the arguments available. However, the model omits strategizing aspects of play, such as opponent modelling, to focus on the fundamental issues.

The key problem in strategic argumentation, which must be solved by each player at each move, is to choose a set of arguments I to play that will achieve her desired outcome. We refer to this as the Desired Outcome (or DO) Problem.

The Desired Outcome Problem for P

Instance A split argumentation framework $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ and a desired outcome for P.

Question Is there a set $I \subseteq \mathcal{A}_P$ such that P's desired outcome is achieved in the argumentation framework $(\mathcal{A}_{Com} \cup I, \gg)$?

A similar problem is called the strategic argumentation problem in (Governatori et al. 2014b), but we prefer a more specific name, since there are several problems associated with strategic argumentation. The Desired Outcome problem is essentially an abduction problem (Booth et al. 2014; Maher 2014). The operation of adding arguments to an argumentation framework has been studied in (Cayrol, de Saint-Cyr, and Lagasquie-Schiex 2010), which addresses structural properties, and (Baumann and Brewka 2010), which addresses enforcing a desired extension, rather than a single desired argument. More recently, (Bisquert et al. 2013) addresses a problem similar to the Desired Outcome problem and establishes a relationship to belief update.

Implicit in the treatment of argumentation games (and, indeed, in games generally) is that the rules of the game are

[1] Each player's move is a normal expansion, in the terminology of (Baumann and Brewka 2010).

[2] For brevity, we ignore the possibility that neither player can achieve her desired outcome.

[3] For a similar argumentation game, results of (Rahwan, Larson, and Tohmé 2009) suggest games are strategy-proof only under very constraining conditions.

adhered to by the players. Specifically, in games of incomplete knowledge, the privacy of the knowledge of a player is assumed to be respected, and this privacy is a strategic advantage. However, as pointed out by (Maher 2014) in a more concrete context, it is possible for a player or players to subvert the game by violating the assumed confidential nature of $\mathcal{A}_P$ and $\mathcal{A}_O$, through either espionage or collusion. Such subversion corresponds to a violation of legal ethics, and so it is in the interests of the subverting player(s) that the play of the game has the *appearance* of normal play. This gives rise to the problem of how to achieve this appearance while producing the desired outcome of the subverter(s). We consider the corresponding decision problems.

In the case of collusion between P and O to ensure that (say) P wins, the players must arrange a sequence of moves that satisfy the rules of the game and leads to P winning.

The Winning Sequence Problem for P
Instance A split argumentation framework $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ and a desired outcome for P.
Question Is there a sequence of moves such that P wins?

Recall that each player must achieve her desired outcome at the end of each move and O must have no more moves at the end of the sequence, so deciding whether such a sequence exists is non-trivial.

In the case of espionage, one player, say P, knows her opponents arguments $\mathcal{A}_O$ and desires a strategy that will ensure P wins, no matter what moves O makes. A *strategy* for P in a split argumentation framework $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ is a function from a set of common arguments and a set of playable arguments to the set of rules to be played in the next move. A sequence of moves $S_1, T_1, S_2, T_2, \ldots$ resulting in common arguments $\mathcal{A}^{P,1}_{Com}, \mathcal{A}^{O,1}_{Com}, \mathcal{A}^{P,2}_{Com}, \mathcal{A}^{O,2}_{Com}, \ldots$ is *consistent with* a strategy s for P if, for every j, $S_{j+1} = s(\mathcal{A}^{O,j}_{Com}, \mathcal{A}_P)$. A strategy for P is *winning* if every valid sequence of moves consistent with the strategy is won by P.

The Winning Strategy Problem for P
Instance A split argumentation framework $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ and a desired outcome for P.
Question Is there a winning strategy for P?

We say that strategic argumentation is *resistant to collusion* (respectively, *espionage*) if the complexity of the Winning Sequence problem (Winning Strategy problem) is greater than the complexity of the Desired Outcome problem, under the commonly believed complexity-theoretic assumption that the polynomial hierarchy does not collapse. We say that strategic argumentation is *resistant to corruption* if it is resistant to both espionage and collusion.

Thus, argumentation is resistant to corruption if the computational cost of exploiting the corruption is greater (under the complexity-theoretic assumption) than the cost of simply playing the game. This computational cost is a potential barrier to corruption.

Example 1 *Consider a split argumentation framework $(\mathcal{A}_{Com}, \mathcal{A}_P, \mathcal{A}_O, \gg)$ with arguments A, B, C, and D where $\mathcal{A}_{Com} = \emptyset$, $\mathcal{A}_P = \{A, D\}$, $\mathcal{A}_O = \{B, C\}$, and D attacks B, B attacks A and C, C attacks A and B. P's aim is to have A accepted. Note that, under the grounded (and stable) semantics, A and B are rejected while C and D are accepted.*

With espionage, P (say) knows all of O's arguments, but cannot control how O plays them. Thus, after P plays A, O might play C and O would win.

With collusion, P and O know all the arguments and can control how they are played. However, they are constrained by the necessity to play them to give the appearance of normal play, so that their collusion is not exposed. In collusive play to have A accepted, P plays A, O plays B, P plays D and wins.

Grounded Semantics

We assume that the desired outcome for P is to have a distinguished argument a accepted in the grounded extension, while the desired outcome for O is that a is not accepted.

We first show that, even without strategizing, performing a single move in strategic abstract argumentation is difficult, in the worst case. Specifically, it is NP-complete. The proof is by reduction of SAT to the Desired Outcome problem.

Theorem 2 *The Desired Outcome problem for P under the grounded semantics is NP-complete*

The aim of the opponent O is to ensure that the distinguished argument is not accepted in the grounded extension.

Corollary 3 *The Desired Outcome problem for O under the grounded semantics is NP-complete*

Alternatively, O might want the distinguished argument to be rejected, rather than simply not accepted.

Corollary 4 *The Desired Outcome problem for O under the grounded semantics, where O's aim is to reject the distinguished argument, is NP-complete*

Thus the difficulty of playing a strategic argumentation game under the grounded semantics is the same for P and O.

We now turn to the two problems that arise when exploiting information obtained by corrupt behaviour. To exploit knowledge of the opponent's arguments requires a winning strategy, determining the existence of which is PSPACE-complete, as would be expected from the theory of complete games.

Theorem 5 *Under the grounded semantics the Winning Strategy problem is PSPACE-complete.*

The proof that the problem is PSPACE-hard involves modelling quantified Boolean formulas, where each player corresponds to a quantifier, and each block of variables corresponds to arguments that a player may play, extending the construction of Theorem 2. The main technical difficulty is in ensuring that playing of arguments adheres to the order in which variables are quantified, since argumentation games permit the playing of any argument in any move.

Exploiting collusion might appear to have similar requirements to exploiting espionage but, because the players are coordinating, it is easier.

Theorem 6 *Under the grounded semantics, the Winning Sequence problem is Σ_2^p-complete.*

The proof of hardness is essentially a simpler form of the proof of the previous theorem.

For these two problems we see that strategic argumentation under the grounded semantics is resistant to corruption, since the complexity of exploiting the information obtained corruptly is greater than the complexity of playing the game.

Stable Semantics

The stable semantics generates, in general, multiple extensions of an argumentation framework $\mathcal{A}$. Each stable extension represents a coherent classification of the arguments as accepted or rejected. These extensions are exactly the possible results of rational adjudication of the arguments in $\mathcal{A}$, under the assumption that the judge rules on all arguments.

The multiplicity of extensions makes for a greater range of aims available to a player than under the grounded semantics. A player may aim to have her distinguished argument a accepted in *all* stable extensions. This corresponds to the aim of the grounded semantics: to unequivocally establish a.

Alternatively, a player might aim only to have her distinguished argument accepted in *at least one* stable extension. In a situation where summary judgement must be avoided, the existence of one extension supporting the player's contention a can be seen as sufficient reason for continuing legal proceedings. Achieving this weaker aim will, in general, expose fewer of the player's arguments.

A player might aim only to have a accepted in *more than half* of the stable extensions. This aim might be sufficient in situations where the judgement criterion is the preponderance of evidence, or the balance of probabilities. Alternatively, a player might aim to have a better than 2:1 ratio of stable extensions accepting a to those that do not. This represents a more decisive advantage over the opponent than simply having more stable extensions accepting a.

For later reference we enumerate these aims.

1. **Universal:** a is accepted in all stable extensions
2. **Existential:** a is accepted in at least one stable extension
3. **Majority:** a is accepted in more than half of the stable extensions
4. **Supermajority:** the number of stable extensions where a is accepted is more than twice the number of stable extensions where it is not accepted

Clearly the universal aim is closely related to sceptical acceptance, while the existential aim is closely related to credulous acceptance of arguments. The difference is in the strategic choice of additional arguments to achieve this acceptance.

In addition to these aims, a player might wish to prevent her opponent from achieving such aims. Thus, for each of the above aims there is a "spoiler" aim that is the complement of the original aim. For example, a spoiler for the universal aim intends that there is at least one stable extension in which a is not accepted, while a spoiler for the majority aim intends that fewer than (or exactly) half of the stable extensions accept a. For convenience, we will refer to the spoiler aims as the Desired Outcome problem for the player O, while the original aims will be referred to as the Desired Outcome problem for player P.

We now turn to describing the complexity of the Desired Outcome problem under the stable semantics with the various aims.

Theorem 7 *The Desired Outcome problem under the stable semantics with the universal aim is Σ_2^p-complete.*

Proof Consider the following algorithm.

> Nondeterministically choose a subset I of $\mathcal{A}_P$ and consider the argumentation framework with arguments $\mathcal{A}_{Com} \cup I$ and the restriction of $\gg$ to these arguments. Check that the distinguished argument a is in all stable extensions of this argumentation framework.

The problem of checking that a is accepted in all stable extensions is in co-NP so this algorithm shows that this problem is in Σ_2^p.

To show that the DO problem is Σ_2^p-hard we reduce the satisfiability problem of $\exists\forall$ quantified Boolean formulas to the DO problem. Let X and Y be disjoint sets of Boolean variables, and let ψ be a Boolean formula over those variables in disjunctive normal form. Let $D_1, \ldots, D_n$ be the disjuncts in ψ. Consider the formula $\exists X \forall Y\ \psi$. We construct a split argumentation framework as follows.

For each literal q, there is an argument denoted A_q. There is an argument A_D for each disjunct D, and arguments A_ψ and $A_{\neg\psi}$. There are also, for each variable p, arguments N_p and B_p. $\mathcal{A}_{Com}$ consists of all these arguments. $\mathcal{A}_P$ consists of an argument I_q, for each literal q based on X. $\mathcal{A}_O$ is arbitrary, since it is not relevant to the DO problem.

The attack relation $\gg$ among these arguments is defined as follows[4].

- For each variable p in X,
  I_p attacks $A_{\neg p}$ and $I_{\neg p}$ attacks A_p
- For each variable p in X,
  I_p attacks N_p and $I_{\neg p}$ attacks N_p
  A_p attacks B_p and $A_{\neg p}$ attacks B_p
- For each variable p in X,
  N_p attacks A_ψ and B_p attacks A_ψ
- For each variable p in Y,
  A_p attacks $A_{\neg p}$ and $A_{\neg p}$ attacks A_p
- For each conjunction D, and each literal q in D,
  $A_{\sim q}$ attacks A_D
- For each conjunction D,
  A_D attacks $A_{\neg\psi}$
- $A_{\neg\psi}$ attacks A_ψ

The intuition of the construction is that arguments of the form I_q played by P represent an assignment for X, arguments N_p (respectively B_p) express that neither I_p nor $I_{\neg p}$ (respectively, both I_p and $I_{\neg p}$) are present in I (the set of arguments played by P), and arguments A_α represent that

[4]For any literal q, $\sim q$ is the complement of q. That is, if q has the form $\neg p$ then $\sim q$ is p, while if q is a variable p then $\sim q$ is $\neg p$.

the formula α is true under an assignment derived from the stable extension.

The remainder of the proof is a verification that finding $I \subseteq \mathcal{A}_P$ such that A_ψ is in all stable extensions of $\mathcal{A}_I = (\mathcal{A}_{Com} \cup I, \gg)$ is equivalent to finding an assignment for X that satisfies $\forall Y\ \psi$. □

Given that the credulous acceptance problem (is a in some stable extension of $\mathcal{A}$?) is NP-complete, it is straightforward to show that the corresponding Desired Outcome problem is also NP-complete.

Theorem 8 *The Desired Outcome problem under the stable semantics with the existential aim is NP-complete.*

To characterize the complexity of the remaining aims we need a relatively obscure complexity class. The complexity class PP was originally formulated in probabilistic terms (Gill 1977), but to align better with its use in this paper we use an equivalent formulation using non-deterministic computation (Fortnow 1997). A language L is in PP iff there is a nondeterministic polynomial-time Turing machine M such that x is in L if and only if $M(x)$ has more accepting computation paths than rejecting paths. PP is somewhat similar to the complexity class #P, which counts the number of accepting paths, but involves decision problems, rather than function problems. P^{PP} contains the entire polynomial hierarchy, but is contained in PSPACE.

Several problems related to probabilistic planning are PP-complete or $\mathrm{NP}^{\mathrm{PP}}$-complete (Littman, Goldsmith, and Mundhenk 1998). Some problems related to manipulation in weighted voting games are PP-complete(Faliszewski and Hemaspaandra 2009); some others are $\mathrm{NP}^{\mathrm{PP}}$-complete (Rey and Rothe 2014).

Theorem 9 *The Desired Outcome problem under the stable semantics with the majority aim is NP^{PP}-complete.*

In fact, the use of any (rational) ratio between accepting and rejecting computation paths (other than 0:1 and 1:0) leads to the same complexity class, $\mathrm{NP}^{\mathrm{PP}}$. Consequently, any supermajority aim, including the ratio 2:1, is also $\mathrm{NP}^{\mathrm{PP}}$-complete.

Corollary 10 *The Desired Outcome problem under the stable semantics with the supermajority aim is NP^{PP}-complete.*

We might wish to give weights to each stable extension and aim to have the sum of weights of stable extensions accepting the distinguished argument be greater than the sum of the weights of remaining extensions. Again, the Desired Outcome problem with this aim is $\mathrm{NP}^{\mathrm{PP}}$-complete.

Unlike the grounded semantics, the difficulty of playing a strategic argumentation game under the stable semantics is different, in general, for player P and player O.

Proposition 11 *The complexity of the Desired Outcome problem under the stable semantics for player O is:*

- *NP-complete when P has the universal aim*
- *Σ_2^p-complete when P has the existential aim*
- *NP^{PP}-complete when P has the majority or supermajority aim*

Finally, an idiosyncrasy of the stable semantics is that some argumentation frameworks have no stable extension. Thus a player might want a move to create such an argumentation framework, particularly in cases where she appears likely to lose. This aim is as hard as the universal aim.

Theorem 12 *The Desired Outcome problem under the stable semantics, where the aim is to have no stable extensions, is Σ_2^p-complete.*

We now turn to the cost of exploiting corruption under the stable semantics.

Theorem 13 *Under the stable semantics, with any of the aims, the Winning Strategy problem is PSPACE-complete.*

This result is essentially a corollary to Theorem 5, based on the fact that the construction in that theorem creates well-founded argumentation frameworks.

Similarly, the Winning Sequence problem, no matter which aim, is Σ_2^p-hard, from Theorem 6. For the majority/supermajority aim, a stronger lower bound is obtained by reduction of the EA-MAJSAT problem.

Theorem 14 *Under the stable semantics, the Winning Sequence problem for P is;*

- *Σ_2^p-complete for the universal and existential aims,*
- *$NP^{NP^{PP}}$-complete for the majority and supermajority aims.*

Thus we see that, under the stable semantics, strategic argumentation is resistant to espionage. However, it is not resistant to collusion under the existential or universal aims, because the Winning Sequence problem has the same complexity as the Desired Outcome problem for one of the players. Surprisingly, then, strategic argumentation *is* resistant to collusion under the majority/supermajority aims.

Concrete Systems

There are numerous systems for defeasible reasoning, any of which might be used as the basis of strategic argumentation at the concrete level. Among systems for defeasible reasoning that reflect the grounded semantics are: the defeasible logics NDL and ADL (Maier and Nute 2010), the defeasible logics in the **DL** and **WFDL** (Billington et al. 2010; Maher 2013) frameworks, the extended defeasible logics of Billington (for example (Billington 2011)), courteous logic programs (Grosof 1999) and its more recent incarnations $LPDA$[5] (Wan et al. 2009) and Rulelog (Grosof and Kifer 2013), Ordered Logic (Laenens and Vermeir 1990), logic programming without negation as failure (LPwNF) (Dimopoulos and Kakas 1995), and Defeasible Logic Programming (DeLP) (García and Simari 2004). Similarly, structured argumentation systems ASPIC (Amgoud et al. 2006) and its derivatives (Prakken 2010; Wu and Podlaszewski 2015) and assumption-based argumentation (ABA) (Bondarenko et al. 1997) support grounded semantics.

[5]We assume that LPDA theories have the overriding property (Wan et al. 2009).

Strategic argumentation in these concrete systems is more precise than in abstract argumentation in the sense that players can re-use individual rules – rather than entire arguments – that are played by their opponent. Nevertheless, hardness results are relatively easily extended from abstract arguments to concrete systems. For example:

Theorem 15 *The Desired Outcome problem is NP-hard for the defeasible logics NDL and ADL, the logics in the frameworks* **DL** *and* **WFDL***, Billington's extended defeasible logics, for the formalisms Ordered Logic, LPwNF, courteous logic programs, LPDA, Rulelog, and DeLP, and for the argumentation systems ABA,* ASPIC *and its derivatives under the grounded semantics.*

The proof uses results in (Maher 2015) showing that these concrete argumentation languages can imitate abstract argumentation frameworks.

Parts of this theorem have already been proved. Specifically, (Governatori et al. 2014b) addresses one defeasible logic, (Maher 2014) addresses many of the defeasible logics, and (Governatori et al. 2014a) addresses ASPIC under the grounded semantics. However, all use more complicated technical machinery than is necessary when starting from the abstract level. It would take substantial work to use that machinery to obtain this result.

Some concrete systems have features that suggest that such lower bounds are not tight. Nevertheless, for many of these formalisms our results at the abstract level extend to the concrete level. Hence

Theorem 16 *Consider the following formalisms: defeasible logics in the* **DL** *and* **WFDL** *frameworks, ADL and NDL (assuming minimal conflict sets), courteous logic programs, LPDA (assuming the overriding property), Ordered Logic, LPwNF, and flat ABA,* ASPIC *and its derivatives under the grounded semantics.*

Strategic argumentation in any of these formalisms is resistant to corruption.

Several concrete languages reflect the stable semantics, rather than the grounded semantics including: defeasible logics under stable model semantics (Antoniou et al. 2000; Maier 2013), DefLog (Verheij 2003), $ASPDA$[6] (Wan, Kifer, and Grosof 2015), as well as ABA, ASPIC and its derivatives under the stable semantics.

Theorem 17 *Consider the following formalisms: ambiguity blocking defeasible logics in the* **DL** *framework under the stable models semantics; NDL (assuming minimal conflict sets) under the β-stable sets semantics; ASPDA (assuming the overriding property); and flat ABA,* ASPIC *and its derivatives under the stable semantics.*

Strategic argumentation in any of these formalisms is resistant to espionage, and is resistant to collusion for the majority and supermajority aims.

[6]We assume that ASPDA theories have the overriding property (Wan et al. 2009).

Related Work

(Maher 2014) introduced the idea of resistance to corruption for the defeasible logic DL (Antoniou et al. 2001). He showed the equivalent of Theorems 2, 5, and 6 for this logic. This paper goes further in establishing similar results for abstract argumentation and for a wider range of concrete languages that reflect the grounded semantics. In addition, this paper addresses abstract argumentation and concrete languages that reflect the *stable* semantics. The stable semantics presents significantly more complications, and the results on resistance to corruption are more nuanced. (Governatori et al. 2014b; 2014a) each only addressed the Desired Outcome problem for a single formalism. (Booth et al. 2014) view the Desired Outcome problem as an abduction problem where a game of hypothetical moves refines the set of abducibles that are needed to explain the acceptance of an observation. Theorem 2 identifies the complexity of playing their game under the grounded semantics.

Conclusion

We have seen that playing strategic argumentation games is difficult, but the degree of difficulty depends on the argumentation semantics and the aim of the player. Under the grounded semantics, the difficulty of playing the game is the same for P and O. In contrast, under the stable semantics, if P has a universal aim then she has a harder task than her opponent.

We have also seen that, in general, argumentation games are resistant to corruption. The games are resistant to espionage under both grounded and stable semantics. They are also resistant to collusion under the grounded semantics and under the stable semantics when the player's aim is a majority or supermajority. However, when P's aim is universal or existential, the complexity of colluding (that is, the complexity of finding a winning sequence) is the same as the complexity of finding a next move for one of the players. Thus, from a worst-case viewpoint, the computational difficulty of exploiting collusion is not a barrier or disincentive to collusive behaviour. These results were proved for abstract argumentation, and then extended to a wide variety of concrete languages for expressing argumentation.

This work suggests several avenues for further research: How can the results here be adapted to negotiation, where P and O do not have inconsistent aims, and to multiple players? There are many other semantics for abstract argumentation; are games resistant to collusion under these semantics? Do similar results apply to other notions of abstract argumentation?

References

Amgoud, L.; Bodenstaff, L.; Caminada, M.; McBurney, P.; Parsons, S.; Prakken, H.; van Veenen, J.; and Vreeswijk, G. 2006. Final review and report on formal argumentation system. Technical report.

Antoniou, G.; Billington, D.; Governatori, G.; and Maher, M. J. 2000. A flexible framework for defeasible logics. In *AAAI/IAAI*, 405–410. AAAI Press / The MIT Press.

Antoniou, G.; Billington, D.; Governatori, G.; and Maher, M. J. 2001. Representation results for defeasible logic. *ACM Trans. Comput. Log.* 2(2):255–287.

Bartholdi, J. J.; Tovey, C. A.; and Trick, M. A. 1989. The computational difficulty of manipulating an election. *Social Choice and Welfare* 6(3):227–241.

Baumann, R., and Brewka, G. 2010. Expanding argumentation frameworks: Enforcing and monotonicity results. In *COMMA*, 75–86.

Billington, D.; Antoniou, G.; Governatori, G.; and Maher, M. J. 2010. An inclusion theorem for defeasible logics. *ACM Trans. Comput. Log.* 12(1):6.

Billington, D. 2011. A defeasible logic for clauses. In *AI 2011: Advances in Artificial Intelligence*, volume 7106 of *Lecture Notes in Computer Science*, 472–480. Springer.

Bisquert, P.; Cayrol, C.; de Saint-Cyr, F. D.; and Lagasquie-Schiex, M. 2013. Enforcement in argumentation is a kind of update. In *Proc. Scalable Uncertainty Management*, 30–43.

Bondarenko, A.; Dung, P. M.; Kowalski, R. A.; and Toni, F. 1997. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.* 93:63–101.

Booth, R.; Gabbay, D. M.; Kaci, S.; Rienstra, T.; and van der Torre, L. W. N. 2014. Abduction and dialogical proof in argumentation and logic programming. In *ECAI 2014 - 21st European Conference on Artificial Intelligence*, 117–122.

Cayrol, C.; de Saint-Cyr, F. D.; and Lagasquie-Schiex, M.-C. 2010. Change in abstract argumentation frameworks: Adding an argument. *J. Artif. Intell. Res. (JAIR)* 38:49–84.

Dimopoulos, Y., and Kakas, A. C. 1995. Logic programming without negation as failure. In *Proceedings of the 1995 International Symposium on Logic Programming*, 369–384.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* 77(2):321–358.

Dunne, P. E., and Wooldridge, M. 2009. Complexity of abstract argumentation. In Rahwan, I., and Simari, G., eds., *Argumentation in Artificial Intelligence*. Springer.

Faliszewski, P., and Hemaspaandra, L. A. 2009. The complexity of power-index comparison. *Theor. Comput. Sci.* 410(1):101–107.

Fortnow, L. 1997. Counting complexity. In *Complexity Theory Retrospective II*. Springer.

García, A. J., and Simari, G. R. 2004. Defeasible logic programming: An argumentative approach. *TPLP* 4(1-2):95–138.

Gill, J. 1977. Computational complexity of probabilistic turing machines. *SIAM J. Comput.* 6(4):675–695.

Governatori, G.; Maher, M. J.; Olivieri, F.; Scannapieco, S.; and Rotolo, A. 2014a. The complexity of strategic argumentation under grounded semantics. In *Proc. European Conf. on Multi-Agent Systems*, 379–387.

Governatori, G.; Olivieri, F.; Scannapieco, S.; Rotolo, A.; and Cristani, M. 2014b. Strategic argumentation is NP-complete. In *Proc. European Conf. on Artificial Intelligence*, 399–404.

Grosof, B., and Kifer, M. 2013. Rulelog: Syntax and semantics. http://ruleml.org/rif/rulelog/spec/Rulelog.html.

Grosof, B. N. 1999. Compiling prioritized default rules into ordinary logic programs. Technical report, IBM.

Laenens, E., and Vermeir, D. 1990. A fixpoint semantics for ordered logic. *J. Log. Comput.* 1(2):159–185.

Littman, M. L.; Goldsmith, J.; and Mundhenk, M. 1998. The computational complexity of probabilistic planning. *J. Artif. Intell. Res. (JAIR)* 9:1–36.

Maher, M. J. 2013. Relative expressiveness of well-founded defeasible logics. In *Proc. Australasian Joint Conf. on Artificial Intelligence*, 338–349.

Maher, M. J. 2014. Complexity of exploiting privacy violations in strategic argumentation. In *Proc. Pacific Rim International Conf. on Artificial Intelligence*, 523–535.

Maher, M. J. 2015. Relating concrete argumentation formalisms and abstract argumentation. In *Proc. Tech. Comm. Int. Conf. Logic Programming*.

Maier, F., and Nute, D. 2010. Well-founded semantics for defeasible logic. *Synthese* 176(2):243–274.

Maier, F. 2013. Interdefinability of defeasible logic and logic programming under the well-founded semantics. *TPLP* 13:107–142.

Prakken, H., and Sartor, G. 1998. Modelling reasoning with precedents in a formal dialogue game. *Artif. Intell. Law* 6(2-4):231–287.

Prakken, H. 2005. Coherence and flexibility in dialogue games for argumentation. *J. Log. Comput.* 15(6):1009–1040.

Prakken, H. 2010. An abstract framework for argumentation with structured arguments. *Argument and Computation* 1:93–124.

Rahwan, I.; Larson, K.; and Tohmé, F. A. 2009. A characterisation of strategy-proofness for grounded argumentation semantics. In Boutilier, C., ed., *IJCAI*, 251–256.

Rey, A., and Rothe, J. 2014. False-name manipulation in weighted voting games is hard for probabilistic polynomial time. *J. Artif. Intell. Res. (JAIR)* 50:573–601.

Thang, P. M.; Dung, P. M.; and Hung, N. D. 2012. Towards argument-based foundation for sceptical and credulous dialogue games. In *COMMA*, 398–409.

Verheij, B. 2003. DefLog: on the logical interpretation of prima facie justified assumptions. *J. Log. Comput.* 13(3):319–346.

Wan, H.; Grosof, B. N.; Kifer, M.; Fodor, P.; and Liang, S. 2009. Logic programming with defaults and argumentation theories. In Hill, P. M., and Warren, D. S., eds., *ICLP*, volume 5649 of *Lecture Notes in Computer Science*, 432–448. Springer.

Wan, H.; Kifer, M.; and Grosof, B. N. 2015. Defeasibility in answer set programs with defaults and argumentation rules. *Semantic Web* 6(1):81–98.

Wu, Y., and Podlaszewski, M. 2015. Implementing crash-resistance and non-interference in logic-based argumentation. *J. Log. Comput.* 25(2):303–333.