

A Game-Theoretic Measure of Argument Strength for Abstract Argumentation

Paul-Amaury Matt and Francesca Toni
pmatt,f.toni@imperial.ac.uk

Department of Computing,
Imperial College London

Abstract. Abstract argumentation (Dung 1995) is a theory of dialectic that allows us to formalise and study various notions of argument acceptability. We depart from this standard approach and formalise a measure of argument strength by applying the concept of value of a game, as defined in Game Theory (von Neumann 1928). The measure thus obtained satisfies a number of intuitively appealing properties that can be derived mathematically from the minimax theorem.

1 Introduction

Dialectic corresponds informally to the art or practise of logical discussion as employed in investigating the truth of a theory or an opinion. In classical philosophy, dialectic is controversy and consists in the exchange of arguments and counter-arguments respectively advocating propositions (theses) and counter-propositions (antitheses). Abstract argumentation (Dung 1995) can be seen as a modern theory of dialectic that allows us to model conflict between arguments and formalise various notions of argument acceptability. The adoption of arguments thus deemed dialectically acceptable combined with the rejection of unacceptable arguments constitutes a natural approach to deliberation.

This type of deliberation is however somewhat simplistic, as it classifies arguments into two categories only, *viz.*, acceptable and unacceptable arguments. Several works, e.g. (Krause *et al.* 1995, Jakobovits and Vermeir 1999, Besnard and Hunter 2001, Cayrol and Lagasquie-Schiex 2005) have considered and explored the possibility of discriminating between arguments using a larger number of categories or continuous numerical scales. The implicit common objective of such approaches is to eventually elaborate a theory of “careful deliberation” rooted in dialectic. We aim at following these works by assessing the strength of arguments on a scale of values ranging from 0 to 1 so as to finely compare and rank arguments in decreasing order of acceptability, identify the weakest arguments and better understand the influences that arguments have on each other in disputes. This fits well with recent interest in quantitative measures for the analysis of persuasion dialogues (Amgoud and Dupin de Saint-Cyr 2008, Budzyńska *et al.* 2008).

The most fundamental ideas used to formalise argument strength in this paper are essentially the same as those found in abstract argumentation theory:

an argument may be called strong whenever the argument can be defended by one or several well-formed opinion(s) that properly withstand(s) external criticism. In order to assess the strength of a given argument in a dispute, we will essentially have to confront two fictitious persons, endorsing the roles of proponent and opponent of the argument. Situations of conflict between two persons such as this one can be rigorously analysed using the paradigm of Game Theory (von Neumann and Morgenstern 1944). We will thus introduce a special two-person game called *game of argumentation strategy* to confront the opinions of the proponent and opponent of an argument and assess its strength.

The remainder of the paper is organised as follows. In Section 2, we provide a short introduction to abstract argumentation theory. In Section 3, we set the exact rules of a game of argumentation strategy. In Section 4, we justify why the expected outcome of the game – also called game’s value – may be adopted as strength value. We dedicate Section 5 to the mathematical study of this game-theoretic argument strength measure. We finally summarise the contribution of the paper and discuss related works in Section 6.

2 Abstract argumentation

Arguments, opinions and the conflicts opposing opinions in a dispute can be represented in an elegant fashion using directed graphs whereby arguments appear as nodes and attacks between pairs of arguments appear as directed edges. Such graphs correspond to *abstract argumentation frameworks* and constitute the basis of abstract argumentation theory (Dung 1995). Formally,

Definition 1 (abstract argumentation framework). *An abstract argumentation framework is a pair (Arg, att) where Arg is a set of arguments and $att \subseteq Arg \times Arg$ is a binary relation between arguments.*

For example, a framework $F = (Arg, att)$ may consist of $Arg = \{a, b, c, d, e, f\}$ and $att = \{(a, b), (b, a), (b, c), (c, d), (e, c), (f, e)\}$. The corresponding directed graph is shown in Fig. 1.

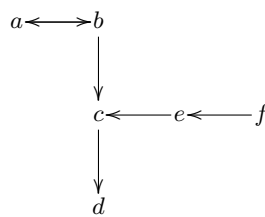


Fig. 1. A simple abstract argumentation framework

The opinions held by the participants of a dispute can be simply represented by the sets of arguments they embrace. Thus, opinions formally correspond to

sets of arguments $X \subseteq Arg$. Conflicts between opinions can be formalised in terms of attack between sets of arguments. We say that an opinion $X \subseteq Arg$ attacks the opinion $Y \subseteq Arg$ when there exists an attack $(x, y) \in X \times Y$ which originates from an argument $x \in X$ and is directed against an argument $y \in Y$. In the framework of Fig. 1, it holds for instance that $\{a, c, f\}$ attacks $\{b, e\}$, that $\{b, e\}$ attacks $\{c, d\}$ and that $\{c, d\}$ attacks itself.

The main purpose of argumentation theory is to identify which arguments and opinions are rationally "acceptable". To address this problem, several notions of *acceptability* have been put forward in the literature (Dung 1995, Bondarenko *et al.* 1997, Dung *et al.* 2006, 2007). In this paper, we will mostly deal with the notions of conflict-freeness, admissibility and stability.

Definition 2 (acceptability). *A set $X \subseteq Arg$ of arguments is said to be*

- *conflict-free if and only if X does not attack itself*
- *admissible if and only if X is conflict-free and attacks every argument that attacks X*
- *stable if and only if X is conflict-free and attacks every argument that is not an element of X*

Intuitively, conflict-freeness conveys the idea that well-formed opinions should be internally consistent. Admissibility is a stronger notion of acceptability according to which opinions should not only be conflict-free but also incorporate the counter-arguments necessary to resist (external) criticism. Finally stability is an even stronger notion of admissibility which requires all arguments not embraced by the opinion to be attacked.

In the example of Fig. 1, $\{a, c, f\}$ is conflict-free, admissible and stable, $\{b\}$ is conflict-free and admissible but not stable, $\{c, f\}$ is conflict-free but not admissible and $\{a, b, c\}$ is neither stable nor admissible nor conflict-free.

3 Games of argumentation strategy

In classical abstract argumentation, arguments are either acceptable or unacceptable, given a chosen notion of acceptability. This gives a rather coarse way to compare arguments. So, for example, for the framework given in Fig. 1, b and f are both equally admissible arguments. However, intuitively f can be deemed to be "stronger" than b , as it is not "weakened" by any attacking argument (whereas b is "weakened" by a). In general, in order to assess the strength of an argument, we will essentially weigh the opinions embracing that argument (opinions pro) against the possible criticisms that can be raised against them (opinions con). We will define, in Section 4, a notion of argument strength matching this intuition. This notion will be defined in terms of the value of a game of strategy (Borel 1921, von Neumann 1928, von Neumann and Morgenstern 1944) confronting two fictive players endorsing the roles of *proponent* and *opponent* of some argument of interest. In this section we will define this game.

Let us assume given and fixed an abstract argumentation framework (F, x) representing a dispute and denote by $x \in Arg$ the argument whose strength is

to be measured. In the remainder of this section, we introduce the exact rules of a game that is specific to the argument x and based on the structure of F . This game will be referred to as (F, x) *game of argumentation strategy*, or for convenience simply as (F, x) game.

In Game Theory, the choices available to the players are referred to as pure strategies. In the (F, x) game, strategies are sets of arguments $X \subseteq \text{Arg}$ and are interpreted as opinions. The proponent of argument x is required to embrace it, so we impose that x belongs to the set of arguments (strategy) played by the proponent. The opponent is however free to select any set of arguments to play the game.

Definition 3 (pure strategies). *The sets of pure strategies for the proponent and opponent players are $\{P \mid P \subseteq \text{Arg}, x \in P\}$ and $\{O \mid O \subseteq \text{Arg}\}$ respectively.*

Let (P, O) be an arbitrary pair of strategies chosen by the proponent and opponent respectively. A degree of acceptability of P with respect to O can be defined on the basis of the attacks directed from P to O and from O to P within the abstract argumentation framework F . Let us then denote for every set of arguments $A, B \subseteq \text{Arg}$ in the framework $F = (\text{Arg}, \text{att})$

Notation 1 (set of attacks) $B_F^{\leftarrow A} = \{(a, b) \in A \times B \mid (a, b) \in \text{att}\}$

the set of attacks from A against B . According to this notation, $O_F^{\leftarrow P}$ represents the set of attacks from P against O and $P_F^{\leftarrow O}$ the set of attacks from O against P in F . In a dispute, it is better for the proponent of an argument to have more attacks against opponents to the argument and fewer attacks from them. To make sense dialectically, the degree of acceptability $\phi(P, O)$ of P with respect to O shall thus be as great as $O_F^{\leftarrow P}$ is big and as low as $P_F^{\leftarrow O}$ is small. The sets of attacks $O_F^{\leftarrow P}$ and $P_F^{\leftarrow O}$ may be arbitrarily large, so in order to construct a bounded acceptability scale, we transform their sizes $|O_F^{\leftarrow P}|$ and $|P_F^{\leftarrow O}|$ into values $x = f(|O_F^{\leftarrow P}|)$ and $y = f(|P_F^{\leftarrow O}|)$ using a monotonic increasing mapping $f : \mathbb{N} \rightarrow [0, 1[$ such that $f(0) = 0$ and $\lim_{n \rightarrow \infty} f(n) = 1$. The degree of acceptability may then be expressed as a function h of the variables x and y . Several choices are possible for such a function h , but the function $h(x, y) = \frac{1}{2}(1 + x - y)$ is remarkably the only one amongst those of the general form $h(x, y) = ax + by + c$ which fulfils simultaneously $h(x, y) = 1 - h(y, x)$ and $h(0, 1) = 0$. We thus adopt the following simple analytical expression

Definition 4 (degree of acceptability of P with respect to O).

$$\phi(P, O) = \frac{1}{2} [1 + f(|O_F^{\leftarrow P}|) - f(|P_F^{\leftarrow O}|)]$$

Concretely, for illustration purposes (and for the first part of proposition 5), we will use the mapping f defined $\forall n \in \mathbb{N}$ as

$$f(n) = 1 - \frac{1}{n+1} = \frac{n}{n+1}$$

For abstract the argumentation framework F shown in Fig. 1 and considering the opinions $P = \{a, c, f\}$ and $O = \{b, d, e\}$, the sets of attacks from P to O and from O to P are respectively $O_F^{-P} = \{(a, b), (c, d), (f, e)\}$, $P_F^{-O} = \{(b, a), (e, c)\}$ and the degree of acceptability of P with respect to O is $\phi(P, O) = \frac{1}{2}[1 + f(3) - f(2)] = \frac{1}{2}[1 + \frac{3}{4} - \frac{2}{3}] = \frac{13}{24}$.

The notion of degree of acceptability can be used to define the notion of reward for the players in the given game of argumentation strategy. To properly defend the argument x , the proponent should naturally avoid self-contradiction. In other words, P should if possible be chosen so as to be conflict-free. Furthermore, to really play his role in the game, the opponent should contradict the proponent, therefore he should be maximally penalised whenever his opinion O fails to attack P . Finally, each player should seek to maximise the degree of acceptability of his opinion with respect to the one of his adversary. Rewards can be assigned to the players of the game of argumentation strategy in such a way as to give them a (material) incentive to follow these three fundamental principles of rationality.

Definition 5 (players' reward). *If P is not conflict-free, then the opponent should pay the proponent the sum $r_F(P, O) = 0$. If P is conflict-free and O does not attack P , then the opponent should pay the proponent the sum $r_F(P, O) = 1$. Otherwise, the opponent should pay the proponent the sum $r_F(P, O) = \phi(P, O)$.*

The following properties, which will only be used later (see Section 5), can be proved straightforwardly.

Proposition 1.

- 1) $0 \leq r_F(P, O) \leq 1$
- 2.a) $r_F(P, O) = 0$ if and only if P is not conflict-free
- 2.b) $r_F(P, O) = 1$ if and only if P is conflict-free and O does not attack P
- 3) if P is admissible (or stable), then $r_F(P, O) \geq \frac{1}{2}[1 + f(|O|) - f(k|O|)]$, where k is the maximal out-degree (number of outgoing attacks) of the arguments contained in O
- 4) if there exist k attacks from O against P in F , then $r_F(P, O) < 1 - \frac{1}{2} f(k)$

According to definition 5, the proponent's reward is always equal to the opponent's loss. Games of argumentation strategy thus fall into the category of zero-sum games¹. Note also that if the opponent fails to attack the proponent, then he is penalised with a maximal loss of 1. To reduce his loss, the opponent must then seek to minimise the number $|O_F^{-P}|$ of attacks against his opinion O and maximise the number $|P_F^{-O}|$ of attacks against the proponent's opinion P .

Finally, we impose that the players choose their strategy without prior knowledge of the strategy their adversary intends to play with. Games of argumentation strategy therefore also fall within the category of games with *imperfect information*. Since the outcome of a single round of an (F, x) game is random, in the next section we will be exclusively interested in the game's outcome in the long run / after a large number of rounds – as is customary in Game Theory (Dresher 1981) when considering two-person zero-sum games with imperfect information.

¹ Such games have been extensively studied in the literature on Game Theory and used for analysing conflict situations in non-cooperative domains.

4 Strength of arguments

In this section we define the proponent's long run expected payoff (the game's value) as a value of strength for the argument he embraces. Below, we explain how this value is mathematically defined and actually computed.

Intuitively, the proponent wants his reward $r_F(P, O)$ to be as large as possible, but he controls only the choice of P . The opponent wants to make his loss $r_F(P, O)$ as small as possible, but he only controls the choice of his strategy O . What are the guiding principles which should determine the player's choices and what is the expected outcome of such a game ?

As indicated at the end of section 3, each game of strategy needs to be repeated a large number of times. If a player were always to choose the same strategy, then his adversary could adapt his own strategy to it and get a better payoff. Therefore, it is important for players engaged in a repeated game with imperfect information to randomise their strategies over time. We therefore consider that each time the game is played, the proponent and opponent choose their strategies according to some probability distributions $X = (x_i)$ and $Y = (y_j)$. Thus, the probability of the proponent choosing his i th strategy P_i corresponds to x_i and the probability of the opponent choosing his j th strategy O_j corresponds to y_j . The probability distributions X and Y are called *mixed strategies*. If we denote by m and n the number of strategies available to the proponent and opponent respectively, then, to be valid distributions, X and Y must obviously be positive ($x_i, y_j \geq 0$) and sum up to 1 ($\sum_{i=1}^m x_i = \sum_{j=1}^n y_j = 1$).

By denoting the payoff matrix $R = ((r_{i,j}))_{m \times n}$ where $r_{i,j} = r_F(P_i, O_j)$ and by X^T the transpose of the m -dimensional vector $X = (x_i)$, the proponent's expected payoff is given by

$$E = X^T R Y = \sum_{j=1}^n \sum_{i=1}^m r_{i,j} x_i y_j$$

The proponent can therefore expect to get at least $\min_Y X^T R Y$, where the minimum is taken over all mixed strategies available to the opponent. Since the proponent has the choice of X , he will select X so that this minimum is as large as possible. Hence the proponent can pick a mixed strategy, denoted X^* , which will guarantee him an expectation of at least $\max_X \min_Y X^T R Y$ irrespective of what the opponent does. Similarly, the opponent can make the proponent's expected payoff at most equal to $\min_Y \max_X X^T R Y$ by playing with some strategy Y^* . The *minimax theorem* (von Neumann 1928) states that these two quantities always have a common value v

$$\max_X \min_Y X^T R Y = \min_Y \max_X X^T R Y = v$$

which is called the *value of the game*. This value is both the expected payoff that is guaranteed to the proponent and the maximal expected loss of the opponent. The strength measure we are after can be consequently defined as follows.

Definition 6 (argument strength). The strength $s_F(x)$ of the argument x in the framework F is the value of the (F, x) game of argumentation strategy.

Textbooks on Operations Research (Hillier and Lieberman 1995) explain how to compute v – when the game’s value can be shown to be *a priori* positive – by solving a linear program with the simplex algorithm (Dantzig *et al.* 1955). It can be shown that v corresponds to the solution of the problem that consists in maximising the variable x_{m+1} , subject to the following $(n + m + 2)$ linear inequality constraints

$$\begin{aligned} \forall j \in \{1, \dots, n\} : \sum_{i=1}^m r_{i,j} x_i - x_{m+1} &\geq 0 \\ \sum_{i=1}^m x_i &= 1 \\ x_1, \dots, x_m, x_{m+1} &\geq 0 \end{aligned}$$

Several examples of argument strength in elementary argumentation frameworks are provided in Table 1. For each one of these frameworks $F = (Arg, att)$ and each single argument x in them, we have constructed the (F, x) game payoff matrix $R = ((r_{i,j}))$ where $r_{i,j} = r_F(P_i, O_j)$ and computed the game’s value using the simplex algorithm, as described above. The given ordering on the right-hand column provides a ranking over arguments, for each given framework. Fig. 2

Table 1. Strength (given in alphabetical order of the arguments) and ordering of arguments obtained in several abstract argumentation frameworks $F = (Arg, att)$.

Ref.	Arguments Arg	Attacks att	Strength values	Ordering
F_1	{a}	{}	1	a
F_2	{a, b}	{(a,b)}	1, 0.25	a > b
F_3	{a, b}	{(a,b), (b,a)}	0.5, 0.5	a = b
F_4	{a, ..., d}	{(a,b), (c,b), (d,c)}	1, 0.25, 0.25, 1	a = d > b = c
F_5	{a, ..., d}	{(a,b), (c,b), (d,b), (b,d)}	1, 0.167, 1, 0.625	a = c > d > b
F_6	{a, ..., e}	{(a,b), (c,b), (d,a), (e,c)}	0.25, 0.5, 0.25, 1, 1	d = e > b > a = c
F_7	{a, ..., f}	{(a,b), (b,c), (c,d), (d,e), (e,f)}	1, 0.25, 0.5, 0.386, 0.5, 0.425	a = c = e > f > d > b
F_8	{a, ..., f}	{(a,b), (b,a), (b,c), (c,d), (e,c), (f,e)}	0.5, 0.5, 0.417, 0.5, 0.25, 1	f > a = b = d > c > e

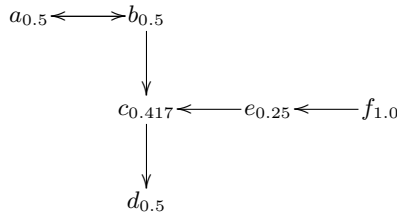


Fig. 2. Strength of arguments in F_8 in Table 1.

shows the results obtained with framework F_8 (this is the framework already

illustrated in Fig. 1). The proponent's optimal strategy ² is to play $\{a, f\}$ for argument a , $\{b\}$ for b , $\{a, c, f\}$ with probability $\frac{2}{3}$ and $\{a, c\}$ with probability $\frac{1}{3}$ for c , $\{b, d\}$ for d , $\{b, e\}$ for e and $\{f\}$ for f . The ranking obtained is in decreasing order of strength $f > a = b = d > c > e$. This ranking mismatches with the ranking $f > d > a = b > e > c$ obtained using the measures proposed in (Besnard and Hunter 2001) and (Cayrol and Lagasquie-Schiex 2005). We regard the latter ranking as unintuitive, as it ranks argument e higher than argument c despite the facts that c is contained in the stable set of arguments $\{a, c, f\}$ and e is not even contained in an admissible set of arguments.

5 Core properties of argument strength

In this section we conduct a thorough mathematical analysis of the properties of this game-theoretic measure of argument strength. Our first result shows that the argument strength scale is bounded.

Proposition 2 (bounds of argument strength). *The strength $s_F(x)$ of an argument x is such that $0 \leq s_F(x) \leq 1$.*

Proof. According to proposition 1, item 1), $\forall(i, j), r_{i,j} \in [0, 1]$. For every mixed strategies X and Y , we also have $X^T R Y \in [0, 1]$, which implies $0 \leq \min_Y X^T R Y$ and $\max_X X^T R Y \leq 1$. Therefore, $0 \leq \max_X \min_Y X^T R Y$ and $\min_Y \max_X X^T R Y \leq 1$. By the minimax theorem, $0 \leq v \leq 1$, and thus $v = s_F(x) \in [0, 1]$.

The next two propositions show that the bounds found are both tight.

Proposition 3 (self-contradiction must be avoided). *The strength $s_F(x)$ of an argument x is 0 if and only if x attacks itself.*

Proof. \Rightarrow : $s_F(x) = v = \min_Y \max_X X^T R Y = 0$ implies the existence of Y^* such that $\forall X, X^T R Y^* \leq 0$. This holds notably for any $X = e_i$ (the vector whose components are all equal to 0 except the i th one which is equal to 1), hence, $\forall i, \sum_j r_{i,j} y_j^* \leq 0$. Since $r_{i,j} y_j^* \geq 0$, it is clear that $\forall(i, j), r_{i,j} y_j^* = 0$. Y^* is a probability distribution, so there exists k such that $y_{j_k}^* > 0$. It is then necessary that $\forall i, r_{i,j_k} = 0$. According to proposition 1, item 2.a), $\forall i, P_i$ attacks itself. In particular, $P_i = \{x\}$ attacks itself, *i.e.* argument x attacks itself.

\Leftarrow : If x attacks itself, then all proponent strategies in the (F, x) game are non-conflict-free sets of arguments. By proposition 1, item 2.a), $R = ((0))$ and $v = 0$.

Proposition 4 (unattacked arguments are the strongest). *The strength $s_F(x)$ of argument x is 1 if and only if there is no argument attacking x in F .*

Proof. \Rightarrow : If $s_F(x) = v = 1$, then we have $\max_X \min_Y X^T R Y = 1$. Y ranges over the set of all real-valued probability distributions which is larger than the set S of all zero-one valued probability distributions. Thus, $\forall X, \min_{Y \in S} X^T R Y \geq \min_Y X^T R Y$. Therefore, $\max_X \min_{Y \in S} X^T R Y \geq \max_X \min_Y X^T R Y = 1$. This can be rewritten as $\max_X \min_j \sum_i r_{i,j} x_i \geq 1$. $\exists X^*$ s.t. $\min_j \sum_i r_{i,j} x_i^* \geq 1$, *i.e.* $\forall j, \sum_i r_{i,j} x_i^* \geq 1$. Since

² An optimal strategy is a (mixed) strategy with maximal expected payoff.

$\forall(i, j), r_{i,j} \leq 1$ and X^* is a probability distribution, $\forall j, \sum_i r_{i,j} x_i^* \leq 1$, so that in fact $\forall j, \sum_i r_{i,j} x_i^* = 1$. This may only hold if $\forall(i, j), r_{i,j} < 1 \Rightarrow x_i^* = 0$. X^* is a probability distribution, so there exists k such that $x_k^* > 0$. By contraposition of the previous implications, $\forall j, \neg(r_{k,j} < 1)$, i.e. $r_{k,j} \geq 1$. By proposition 1, item 1), $\forall j, r_{k,j} = 1$. By proposition 1, item 2.b), $\forall j, P_k$ is conflict-free and O_j does not attack P_k . $x \in P_k$ so there is no opponent strategy or argument that attacks x .

\Leftarrow : By selecting strategy $\{x\}$ with probability 1, the proponent has a guaranteed payoff of 1 irrespective of what the opponent does. Therefore, $v \geq 1$. In fact, v is bounded up by 1 (by proposition 2) and $s_F(x) = 1$.

We can also show that admissible and stable arguments occupy the band of medium to high strength values³ but that attacks against arguments reduce their strength below 1. We assume in part a) of the next proposition that f is defined, $\forall n \in \mathbb{N}$, by $f(n) = \frac{n}{n+1}$ as suggested in Section 3.

Proposition 5 (acceptable arguments have medium to high strength).

- a) If there exists an admissible (or stable) set of arguments containing x , then $s_F(x) \geq \frac{1}{2}[1 + \frac{1}{2} - \frac{k}{k+1}]$ where k is the maximal out-degree of arguments in F .
b) If there exist n attacks against x , then $s_F(x) < 1 - \frac{1}{2}f(n)$.

Proof. a) If P is admissible, then by proposition 1, item 3), $\forall O, r_F(P, O) \geq \frac{1}{2}[1 + f(|O|) - f(k|O|)] \geq \frac{1}{2}[1 + f(1) - f(k)]$ (when $f(n) = \frac{n}{n+1}$). By playing P with probability 1 the proponent of x can secure a payoff of at least $\frac{1}{2}[1 + \frac{1}{2} - \frac{k}{k+1}]$. If P is stable, then P is also admissible and the same inequality holds. b) If there exist n attacks against x , then there exists an opponent strategy O with n attacks against x . For this strategy, and whatever the proponent strategy P , there must also exist at least n attacks from O against P and $r_F(P, O) < 1 - \frac{1}{2}f(n)$ by proposition 1, item 4). By playing O with a probability of 1, the opponent can secure a strict maximum loss of $1 - \frac{1}{2}f(n)$.

Note that when the maximal out-degree in F is $k = 1$, the strength of acceptable arguments is greater than $\frac{1}{2}$, hence the use of the term “medium”.

We now study how the strength of arguments varies as argumentation frameworks are expanded. This should allow us to understand quantitatively the impact of adding new arguments and attacks to a dispute. Then, suppose first that we add an attack (a, b) to the framework $F = (Arg, att)$, where $(a, b) \notin att$ and $a, b \in Arg$. For convenience, in this case we adopt

Notation 2 $F_{+(a,b)} = (Arg, att \cup \{(a, b)\})$

As intuitively expected, adding an attack against an argument reduces its strength:

Proposition 6 (criticism reduces argument strength). $s_{F_{+(a,b)}}(b) \leq s_F(b)$.

Proof. The sets of strategies available to the proponent and opponent are the same in the (F, b) and $(F_{+(a,b)}, b)$ games. Let P and O be proponent and opponent strategies. Note that $P_F^{\leftarrow O} \subseteq P_{F_{+(a,b)}}^{\leftarrow O}$ and either $O_F^{\leftarrow P} = O_{F_{+(a,b)}}^{\leftarrow P}$ (if $a \notin P$) or P attacks itself in $F_{+(a,b)}$ (if $a \in P$). By monotonicity of f , $\phi_{F_{+(a,b)}}(P, O) \leq \phi_F(P, O)$. In any case ($a \in P$ or $a \notin P$), $r_{F_{+(a,b)}}(P, O) \leq r_F(P, O)$. It follows that $s_{F_{+(a,b)}}(b) \leq s_F(b)$.

³ This property can be generalised to any notion of acceptability “stronger” than admissibility, such as e.g. the preferred, complete, grounded and ideal semantics (Dung 1995, Bondarenko *et al.* 1997, Dung *et al.* 2006, 2007).

Adding an attack from argument a against b gives an advantage to the proponent of a as long as b is not useful in the defence of a . Otherwise, this new attack constitutes a handicap for the proponent. To distinguish between these two possible cases, we say that

Definition 7 (superfluous argument). *Argument b is superfluous with respect to a if forbidding the proponent of a to play with strategies containing b does not decrease the proponent's payoff in the (F, a) game (the game's value).*

Proposition 7 (cautious extra-aggressiveness increases strength). *By adding an attack (a, b) one increases a 's strength ($s_{F_{+(a,b)}}(a) \geq s_F(a)$) if b is superfluous with respect to a and diminishes it ($s_{F_{+(a,b)}}(a) \leq s_F(a)$) otherwise.*

Proof. If b is superfluous with respect to a then there exists an optimal mixed strategy X^* for the (F, a) game such that $\forall i, x_i^* > 0 \Rightarrow b \notin P_i$. Let then P be an active strategy, i.e. $P = P_i$ and $x_i^* > 0$. Then, $\forall O$, we have $O_{F_{+(a,b)}}^- \subseteq O_{F_{+(a,b)}}^-$, $P_{F_{+(a,b)}}^- = P_{F_{+(a,b)}}^-$ (if it is not the case that $a \in O$ and $b \in P$) or P attacks itself in $F_{+(a,b)}$ (if $a \in O$ and $b \in P$). The last case does not occur ($b \notin P$) since b is assumed to be superfluous with respect to a . By monotonicity of f , $\phi_F(P, O) \leq \phi_{F_{+(a,b)}}(P, O)$. Since $b \notin P$, P is conflict-free in F iff P is conflict-free in $F_{+(a,b)}$ and O attacks P in F iff O attacks P in $F_{+(a,b)}$. Therefore, for every active strategy P under X^* we have $r_F(P, O) \leq r_{F_{+(a,b)}}(P, O)$. By playing with X^* in the $(F_{+(a,b)}, a)$ game, the proponent can secure a payoff of at least $s_F(a)$. Hence, $s_{F_{+(a,b)}}(a) \geq s_F(a)$. If b is not superfluous with respect to a , then the proponent of a is forced (otherwise his payoff is null) to play strategies containing a but not b , and thus his payoff is reduced.

Moreover, the strength of an argument x may be partially restored by adding an attack (a, b) against one of its attackers b .

Proposition 8 (indirect counter-attack brings support). *If b attacks x , adding an attack (a, b) to F increases x 's strength ($s_{F_{+(a,b)}}(x) \geq s_F(x)$).*

Proof. The sets of strategies of the players are the same in the (F, x) and $(F_{+(a,b)}, x)$ games. We have $O_{F_{+(a,b)}}^- \subseteq O_{F_{+(a,b)}}^-$ (if $a \in P$ and $b \in O$) or $O_{F_{+(a,b)}}^- = O_{F_{+(a,b)}}^-$ otherwise. We also have $P_{F_{+(a,b)}}^- \subseteq P_{F_{+(a,b)}}^-$ (if $b \in P$ and $a \in O$) and $P_{F_{+(a,b)}}^- = P_{F_{+(a,b)}}^-$ otherwise. Note that if $b \in P$ then P attacks itself in both F and $F_{+(a,b)}$. So, $r_F(P, O) \leq r_{F_{+(a,b)}}(P, O)$ and $s_F(x) \leq s_{F_{+(a,b)}}(x)$.

So far we have considered adding attacks between existing arguments in a given dispute (argumentation framework). We finally consider adding new arguments, and show that, as intuitively expected, the status of arguments in this dispute is left unchanged if the newly added arguments are “disconnected” from the original ones.

Proposition 9 (insensitivity to irrelevant information). *If $F' = (Arg', att')$ is such that $Arg \cap Arg' = \emptyset$, then $s_{F+F'}(x) = s_F(x)$ where $F + F' = (Arg \cup Arg', att \cup att')$.*

Proof. Let us consider the $(F + F', x)$ game where $x \in Arg$. Since no argument in Arg' attacks x (the two frameworks are disconnected), the proponent of x is at least as well off in this new game as in the (F, x) game if he restricts himself to his old set of strategies build only from Arg . Therefore, $s_{F+F'}(x) \geq s_F(x)$. The same proposition also holds for the opponent of x , which means that $-s_{F+F'} \geq -s_F(x)$ or equivalently $s_{F+F'}(x) \leq s_F(x)$. In conclusion, $s_{F+F'}(x) = s_F(x)$.

6 Conclusion

Arguments, opinions and conflicts between opinions in disputes can conveniently be modelled using abstract argumentation frameworks (Dung 1995). In order to assess the strength of an argument in a dispute, we defined a repeated game of argumentation strategy whereby two players, *viz.* the proponent and opponent of the argument, simultaneously exchange sets of arguments representing respectively opinions for and against it. We defined a degree of acceptability and reward function for a single round based on the intuition that it is better to have more attacks on and fewer attacks from adversarial opinions within the framework representing the dispute considered. Then, players choose their opinions randomly in each round with a certain probability so as to maximise their expected reward in the long run. The strength of the argument of interest is finally determined by the value of the proposed game (von Neumann 1928) as defined in Game Theory for two-person zero-sum games with imperfect information and computed using the simplex algorithm (Dantzig *et al.* 1955).

We have shown that such a measure of strength is bounded (between 0 and 1), that these bounds are attained for arguments that respectively attacks themselves and that are not attacked, and also that arguments contained into admissible or stable extensions (Dung 1995, Bondarenko *et al.* 1997, Dung 2006, 2007) always have medium to high strength values, but that attacks against such acceptable arguments reduce their strength value below 1. This result also holds for preferred, complete, grounded or ideal extensions, but we have omitted its proof here for lack of space. We have examined the sensitivity of strength values with respect to changes operated on the underlying argumentation framework. Notably, we have seen that adding a new attack against an argument reduces its strength, that adding a new attack against another argument either increases or reduces its strength, depending on the usefulness of the target in the defence of that argument, and that the addition of indirect counter-attacks could restore the strength of an argument. Finally, we have proved that the addition of irrelevant groups of arguments to a dispute does not influence the status of its original arguments.

Several notions of argument strength have already been proposed in the literature on argumentation. One shall distinguish between the so-called "intrinsic" and "interaction-based" measures. The term *intrinsic* is used to refer to approaches whereby the strength of an argument is independent of its interaction with other arguments (Pollock 1992, Krause *et al.* 1995, Ambler 1996, Parsons 1997, Prakken and Sartor 1997, Amgoud and Cayrol 1998, Kohlas *et al.* 2000, Pollock 2001). On the other hand, the term *interaction-based* refers to

measures whereby the strength of an argument depends on the arguments attacking it (attackers), the attackers of its attackers (the defenders), *etc.* Amongst interaction-based measures of argument strength, one may again distinguish between qualitative (Dung 1995, Jakobovits and Vermeir 1999) and quantitative (Besnard and Hunter 2001, Cayrol and Lagasquie-Schiex 2005) measures.

Our interaction-based and quantitative measure is thus closest to the approaches by (Besnard and Hunter 2001) and (Cayrol and Lagasquie-Schiex 2005). However, we have observed (see last paragraph of Section 4), that these measures may produce a different ranking of arguments than ours, and have argued that in general they do not convey the intuition according to which the dialectical properties of admissibility and stability should confer more strength to an argument in a dispute. These measures thus somewhat transgress the principles of dialectic originally proposed by (Dung 1995).

There is other work in the field of argumentation which directly relates to Game Theory. This typically involves “extensive” games, namely multi-stage games that can be represented as trees. Argumentation games in extensive form have been proved to be useful to test the acceptability of arguments under various semantics (Vreeswijk and Prakken 2000, Dunne and Bench-Capon 2003) and to determine optimal strategies (Riveret *et al.* 2008) in dialogues (Prakken 2005). To this date and to the best of our knowledge, the only other work in the domain of argumentation that has sought to exploit strategic games is (Rahwan and Larson 2008), but to model argumentation between self-interest agents.

Acknowledgements. This work was funded by the Sixth Framework IST programme of the EC, under the 035200 ArguGRID project. We wish to express our gratitude to the anonymous referees for their helpful comments and suggestions.

References

- Ambler, S.J.: A Categorical Approach to the Semantics of Argumentation. *Mathematical Structures in Computer Science* 6(2), 167–188 (1996)
- Amgoud, L., Cayrol, C.: On the acceptability of arguments in preference-based argumentation. *14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*, 1–7 (1998)
- Amgoud, L., Prade, H.: Using Arguments for Making Decisions: A Possibilistic Logic Approach. *20th Conference of Uncertainty in Artificial Intelligence (UAI-2004)*, 10–17 (2004)
- Amgoud, L., Dupin de Saint-Cyr, F.: Measures for Persuasion Dialogs: A Preliminary Investigation. *2nd International Conference on Computational Models of Argument (COMMA 2008)*, 13–24 (2008)
- Besnard P., Hunter A.: A logic-based theory of deductive arguments. *Artificial Intelligence* 128, 203–235 (2001)
- Bondarenko, A., Dung, P.M., Kowalski, R.A., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence* 93(1-2), 63–101 (1997)
- Borel, E.: La théorie du jeu et les équations intégrales à noyau symétrique gauche. *Comptes Rendus de l’Académie des Sciences* (1921)

- Budzyńska, K., Kacprzak, M., Rembelski, P.: Modelling Persuasiveness: Change of Uncertainty Through Agents' Interactions. 2nd International Conference on Computational Models of Argument (COMMA 2008), 85–96 (2008)
- Cayrol, C., Lagasque-Schiek M.-C.: Graduality in Argumentation. *Journal of Artificial Intelligence Research* 23, 245–297 (2005)
- Dantzig, G.B., Orden, A., Wolfe, P.: The generalised simplex method for minimizing a linear form under linear inequality constraints. *Pacific Journal of Mathematics* 5(2), 183–195 (1955)
- Dresher, M.: *The Mathematics of Games of Strategy*. Dover Publications (1981)
- Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence* 77(2), 321–257 (1995)
- Dung, P.M., Kowalski, R., Toni, F.: Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence* 170(2), 114–159 (2006)
- Dung, P.M., Mancarella, P., Toni, F.: Computing ideal sceptical argumentation. *Artificial Intelligence* 171, 642–674 (2007)
- Dunne, P.E., Bench-Capon, T.J.M.: Two party immediate response disputes: Properties and efficiency. *Artificial Intelligence* 149, 221–250 (2003)
- Hillier, F.S., Lieberman, G.J.: *Introduction to Operations Research - 6th Edition*. McGraw-Hill (1995)
- Jakobovits, H., Vermeir, D.: Robust semantics for argumentation frameworks. *Journal of logic and computation* 9(2), 215–261 (1999)
- Kohlas, J., Haenni, R., Berzati, D.: Probabilistic argumentation systems and abduction. 8th International Workshops on Non-Monotonic Reasoning, 391–398 (2000)
- Krause, P., Ambler, S., Elvang-Gøransson, M., Fox, J.: A logic of argumentation for reasoning under uncertainty. *Computational Intelligence* 11, 113–131 (1995)
- Parsons S.: Normative argumentation and qualitative probability. 1st International Joint Conference on Qualitative and Quantitative Practical Reasoning, 466–480 (1997)
- Pollock J.L.: How to reason defeasibly. *Artificial Intelligence* 57, 1–42 (1992)
- Pollock J.L.: Defeasible reasoning with variable degrees of justification. *Artificial Intelligence* 133, 233–282 (2001)
- Poole, D.: Probabilistic Horn Abduction and Bayesian Networks. *Artificial Intelligence* 64, 81–129 (1993)
- Prakken, H., Sartor, G.: Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non Classical Logics* 7, 25–75 (1997)
- Prakken, H.: Coherence and Flexibility in Dialogue Games for Argumentation. *Journal of Logic and Computation* 15:6, 1009–1040 (2005)
- Rahwan, I., Larson, K.: Mechanism Design for Abstract Argumentation. 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), 1031–1038 (2008)
- Riveret, R., Prakken H., Rotolo, A., Sartor, G.: Heuristics in Argumentation: A Game-Theoretical Investigation. 2nd International Conference on Computational Models of Argument (COMMA 2008), 324–335 (2008)
- Vreeswijk, G., Prakken H.: Credulous and sceptical argument games for preferred semantics. 7th European Workshop on Logics in Artificial Intelligence (JELIA 2000), 239–253 (2000)
- von Neumann, J.: Zur Theorie des Gesellschaftsspiele. *Mathematische Annalen* 100, 295–320 (1928)
- von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press (1944)