# A Game-Theoretic Perspective on the Notion of Argument Strength in Abstract Argumentation

Paul-Amaury MATT and Francesca TONI
*Imperial College London, Department of Computing*

**Abstract.** This paper is concerned with the problem of quantifying the strength of arguments in controversial debates, which we model as abstract argumentation frameworks [Dung, 1995]. Standard approaches to abstract argumentation provide only a qualitative account of the status of arguments, whereas numerical measures of argument strength might provide a more precise evaluation of their individual status. Intuitively, the strength of an argument in a debate essentially depends on how a proponent of that argument would defend himself against the criticisms of someone opposed to the argument. Since there can be many ways of defending and attacking an opinion, we essentially conceive argument strength as an equilibrium resulting from the interactions taking place between the opinions that a proponent and an opponent of the argument could a priori embrace. More formally, we define argument strength in terms of the value of a repeated two-person zero-sum strategic game with imperfect information. Then, using the game-theoretic properties of such games and notably the von Neumann minimax theorem [Neumann, 1928], we establish and illustrate the main properties of this new argument strength measure.

**Keywords.** Abstract argumentation, acceptability, game theory, minimax theorem.

## 1. Introduction

Controversial debate can essentially be defined as the exchange of arguments and counter-arguments, and deliberation, as the careful consideration of all sides of a debate before making a decision. Abstract argumentation [Dung, 1995] is an elegant paradigm that allows to identify which arguments – in a debate of interest – are rationally acceptable. By considering only the acceptable arguments and rejecting the others, one usually greatly simplifies the deliberation process. To illustrate this process of reasoning, let us consider the following controversial question:

*Can capital punishment be just ?*

in the light of the following arguments[1]

a) Death penalty is an adequate form of punishment as it is a proportionate punishment for murder.

[1] These arguments where found on debatepedia at http://wiki.idebate.org

b) Death penalty devalues the respect we place on human life.
c) With capital punishment, a court is unable to correct its past errors.
d) Capital punishment may cause the convict excessive pain.
e) The issue of pain is simply a matter of implementation and not a matter of the basic principles of justice.
f) Life imprisonment without parole is better than capital punishment because it is more compassionate and allows for a prisoner to develop remorse and repent.
g) Life imprisonment without parole is not more compassionate than capital punishment.

The central and starting argument in this debate is a, which supports the claim that capital punishment is just. This first argument is attacked by four arguments, *viz.* b, c, d, and f. The last two arguments d and f are in turn attacked by e and g respectively. It is not entirely clear whether argument $a$ should be completely rejected, because even though strong arguments are raised against it, some of these are in turn criticised. In abstract argumentation, one seeks to either accept or reject arguments, and in this debate, $a$ would be simply rejected[2]. By rejecting $a$, one would simply come to the conclusion that capital punishment is not just and decide therefore never to have recourse to it. This seems like a reasonable course of action.

Principles and methods for accepting or rejecting arguments, such as those offered by argumentation theory, constitute a simple and qualitative way of understanding debates and drawing decisions from them. This paper proposes to build upon the fundamental principles of rationality used in argumentation theory to provide additional quantitative insight on the individual status of arguments. We aim at assessing numerically the acceptability of arguments, on a scale ranging from zero to one, so as to produce a total ranking of arguments, identify which arguments are most strongly criticised and understand the influence that new arguments and attacks have on the current state of a debate.

The remainder of this paper is organised as follows. In the next section, we briefly recall some background on abstract argumentation. We then borrow some fundamental concepts from this domain and define the rules of a two-person zero-sum game confronting a proponent of some argument of interest, to an opponent of the argument. The superiority of the proponent, which can be measured by the proponent's expected payoff, fundamentally relates to the intrinsic strength of the argument he embraces. We will explain how to calculate this value and show that this new measure exhibits a number of intuitively appealing properties.

## 2. Abstract argumentation

The arguments and structure of controversial debates can be represented in an elegant manner using graphs, whereby arguments appear as nodes and attacks between arguments as directed edges. Such graphs correspond to *abstract argumentation frameworks* and constitute the basis of abstract argumentation theory [Dung, 1995]. Formally,

**Definition 1 (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 of attack between them.*

[2] Under semantics such as *e.g.* admissibility or stability.

For every $a, b \in Arg$, $(a, b) \in att$ is read '$a$ attacks $b$'. Arguments capture the knowledge available from a debate. The attack relationship however structures the existing conflicts in the debate. In the moral debate on capital punishment, the arguments involved are $Arg = \{a, b, c, d, e, f, g\}$ and the attack relationship is given by the ordered pairs $att = \{(b, a), (c, a), (d, a), (f, a), (e, d), (g, f)\}$. The argumentation framework thus corresponds to the directed graph shown in (Fig. 1). The opinions held by the par-

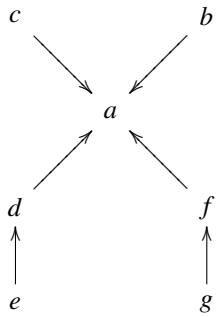


**Figure 1.** Abstract framework structuring the moral debate on capital punishment.

ticipants of a debate can be very simply represented by the set of arguments they respectively embrace. Moreover, an attack or conflict from opinion $X$ against an opinion $Y$ corresponds to an edge $(x, y) \in X \times Y$. For example, a proponent of capital punishment would embrace argument $a$ and may adopt opinion $\{a, e\}$. A participant of the debate that opposes to capital punishment for moral reasons could embrace arguments $\{b, f\}$ and thus criticise the proponent's opinion with the attacks $(b, a)$ and $(f, a)$. The main purpose of argumentation theory is to identify the most rational opinions that can be formed in debates. To address this problem, several candidate characterisations for sets of *acceptable* arguments have been proposed the literature [Dung, 1995; Bondarenko *et al.*, 1997; Dung *et al.*, 2006; Caminada, 2006; Dung *et al.*, 2007; Matt and Toni, 2008]. In this paper, we will only consider the three following ones.

**Definition 2 (notions of acceptability)** *A set X 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 it, and*
- *stable if and only if X is conflict-free and attacks every argument it does not contain.*

Conflict-freeness is certainly the most basic notion of acceptability available in argumentation. It states that an opinion is irrational whenever it contradicts itself. Admissibility is a more advanced notion of acceptability according to which acceptable opinions are those that do not just self-contradict but also resist any external criticism. Finally, the notion of stability characterises as acceptable the opinions which are free of self-contradiction and that ruin all the arguments that are not part of it. In fact, it can be shown that every stable sets of arguments is always also an admissible one.

In the capital punishment debate, the set of arguments $\{c, b, e, g\}$ is conflict-free, admissible and stable. The opinion formed by this set of arguments is acceptable under three different interpretations of acceptability and is therefore according to argumentation theory quite "strong". This opinion attacks the central argument $a$ in favour of capital punishment. Note that the only acceptable sets containing $a$ are $\{a\}$, $\{a, e\}$, $\{a, g\}$ and $\{a, e, g\}$ and are conflict-free but neither admissible nor stable. Consequently, we find strong theoretical reasons – in the context of this very debate – for rejecting the argument in favour of capital punishment and arrive at the conclusion that capital punishment is not just.

## 3. Games of argumentation strategy

Abstract argumentation in its current state does not allow to appreciate quantitatively the degree of acceptability or "strength" of arguments. In order to get an idea on how strong an argument might be, we may look at opinions embracing the argument and the possible criticisms that can be raised against such opinions. Therefore propose to consider a game of strategy [Borel, 1921; Neumann, 1928; von Neumann and Morgenstern, 1944] confronting two players, endorsing the roles of *proponent* and *opponent* of the argument. The proponent shall form an opinion embracing the argument $x$ and the opponent attack this opinion. By weighting the acceptability of the proponent's opinion against the opponent's one, we expect to obtain a value representative of the intrinsic strength of the argument $x$. This game of argumentation played will be fully determined by the argument $x$ and framework $F$ and throughout the paper will be referred to as $(F, x)$ *game of argumentation strategy*. In the remainder of this section, we present the exact rules of this game.

In game theory, the elementary choices available to the players are referred to as pure strategies. In the $(F, x)$ game, the player strategies correspond to opinions in the framework $F$. In other words, if $P$ and $O$ denote proponent and opponent strategies respectively, then $P$ and $O$ correspond to subsets of $Arg$, where $Arg$ is the set of arguments in the argumentation framework $F$. The proponent shall embrace the argument $x$, therefore we must impose the constraint that $x \in P$. In a nutshell,

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

To defend his argument properly, the proponent should avoid to contradict himself, i.e. his opinions should always correspond to sets of arguments that are at least conflict-free. Also, since the opponent's role in the game is to criticise the proponent, the opponent should get a maximal penalty whenever his opinion fails to attack the proponent's one. Finally, the game should provide an incentive for the proponent to attack the opponent's opinion with as many attacks as possible and at the same time force him to avoid the opponent's attacks. These three principles actually reflect the intuition behind the notions of conflict-freeness, admissibility and stability used in classical abstract argumentation. To implement these principles, it is necessary to choose a reward function which reflects the relative degree of acceptability of the players opinions.

To achieve this, we need to consider the interaction between opinions and take into account the attacks existing between them. Let us then adopt the following notation for pairs of sets of arguments $X$ and $Y \subseteq Arg$:

**Notation 1 (attacks from $X$ against $Y$)** $Y_F^{\leftarrow X} = \{(x, y) \in X \times Y \mid (x, y) \in att\}$

and denote by $Y_F^{\leftarrow X}$ the set of attacks from $X$ against $Y$. With 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 $P$. The acceptability of $X$ with respect to $Y$ should monotonically increases with the size of $O_F^{\leftarrow P}$ and decreases with the size of $P_F^{\leftarrow O}$. We propose to use a simple anaylitcal expression such as

**Notation 2 (degree of acceptability of $P$ with respect to $O$)**

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

where $f$ can be any monotonic increasing mapping $f : \mathbb{N} \rightarrow [0, 1[$ such that $f(0) = 0$ and $\lim_{n\rightarrow\infty} f(n) = 1$. In the remainder of the paper, we will consider that $\forall n \in \mathbb{N}$,

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

In an $(F, x)$ game, the rewards are set in the following way.

**Definition 4 (rewards of the game)** *If $P$ is not conflict-free, then the opponent should pay to 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 him the sum $r_F(P, O) = 1$. Otherwise, the opponent should pay the proponent a sum equal to $r_F(P, O) = \phi(P, O)$.*

By definition, the proponent's reward is equal to the opponent's loss. In the terminology of game theory, games of argumentation strategy belong to the category of *zero-sum* games. These games are essential for analysing non-cooperative domains. Observe that if the opponent fails to attack the proponent, then the opponent is penalised with a maximal loss of 1. To reduce his losses, the opponent must seek to minimise the number $|O_F^{\leftarrow P}|$ of attacks against his opinion $O$ and maximise the number $|P_F^{\leftarrow O}|$ of attacks against the proponent's opinion $P$. Besides, it is straightforward to establish that for every proponent and opponent strategies $P$ and $O$,

**Proposition 1** *The rewards are such that*
*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* $r_F(P, O) \geq \frac{1}{2}$
*4) if there exist k attacks of $O$ against $P$ then* $r_F(P, O) < 1 - \frac{1}{2} f(k)$

The strategies and rewards of the $(F, x)$ game have been defined. The only thing that has not been defined yet is the knowledge available to the players during the game. Basically, each player is informed about the argument $x$ to defend/attack and is given the full structure of the argumentation framework $F$, but no other piece of information is provided. This means that the players are asked to choose their strategies without knowledge of their adversary's strategy. Games of argumentation strategy therefore also fall within the category of games with *imperfect information*. Since the outcome of one

round of an $(F, x)$ game is random, one is only interested the game's outcome on the long run (i.e. after a large number of repetitions) as always done in game theory for two-person zero-sum games with imperfect information [Dresher, 1981].

## 4. Strength of arguments

Our intention is to use the proponent's long term expected payoff (the game's value) as a measure of the *strength* of the argument he embraces. In the next section, we will study the properties of such a measure, but in the present section, we first shall 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 its 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 ? Recall that games of argumentation strategy are by definition repeated a large number of times. The rationale for reapting such games is that the player have the choice in each round between multiple strategies and an objective measure of argument strength should at least have some statistical significance.

If a player was choosing always 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 of imperfect information to randomise their strategies over time. We therefore consider that each time the game is played, the proponent and the opponent choose their strategies according to some probability distributions $X$ and $Y$. So, the probability of the proponent choosing his $i$th strategy, which we may denote by convenience $P_i$, is equal to $x_i$. Similarly, the probability of the opponent choosing his $j$th strategy $O_j$ is $y_j$. 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 be such that all $x_i$ and $y_j$ are positive and sum up to one. With these notations, the proponent's expected payoff [3] is given by [Dresher, 1981]

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

and 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

[3] $X^T$ denotes the transpose of vector $X$ and $R$ the matrix $((r_{i,j}))_{m \times n}$ where $r_{i,j} = r_F(P_i, O_j)$.

$$\min_{Y} \max_{X} X^T R Y$$

by playing with some strategy $Y^*$. The *minimax theorem* of [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*. It is both the expected payoff that is guaranteed to the proponent and the maximal expected loss of the opponent. Consequently, we adopt the following

**Definition 5 (strength $s_F(x)$ of argument $x$)** *The strength of argument $x$ in the abstract argumentation framework $F$ is noted $s_F(x)$ and defined as the value of the $(F, x)$ game of argumentation strategy.*

General books of 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]. $v$ is the solution of the problem of maximising $x_{m+1}$, subject to the $n + m + 2$ linear inequality constraints

$$\forall j \in \{1, ..., n\} \quad \sum_{i=1}^{m} r_{i,j} x_i - x_{m+1} \geq 0$$

$$\sum_{i=1}^{m} x_i = 1$$

$$x_1, ..., x_m, x_{m+1} \geq 0$$

The higher the value of $v$, the better off is the proponent of the argument. The value of $v$, which only depends on $x$ and $F$, can thus be interpreted as the strength of $x$ in the context of the debate modelled by the abstract argumentation framework $F$. For instance, the strength of the arguments composing the moral debate on capital punishment appears in Fig. 2.

As one can notice, the strength of the central argument $a$ in favour of capital punishment is quite low compared to the one of other arguments, as suggested our initial qualitative analysis. Moreover, we observe that the arguments composing the opinion $\{b, c, e, g\}$ (which has been shown to be strong in the qualitative sense) against capital punishment all have maximal strength of one. These numerical results thus validate the conclusions of our preliminary analysis in this specific case. The reader may familiarise herself/himself with the strength measure by examining Fig. 4 which displays all the possible configurations of frameworks with one, two or three arguments that do not attack themselves (arguments attacking themselves can be considered as rare in practise) and that correspond to connected graphs.

The reason for showing only connected graphs will be explained in the next section. Basically, we will prove that the strength of arguments in disconnected parts of a graph can be assessed totally independently of each other. Remark that this property may sometimes allow to simplify (computationally speaking) the analysis of complex situations

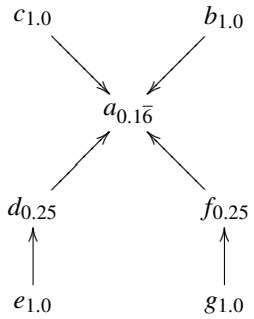


**Figure 2.** Strength of arguments in the moral debate on capital punishment.

where many arguments are grouped within argument "clusters" or sub-graphs that do not conflict with one another. In expert systems for instance, such clusters may correspond to arguments emanating from independent empirical theories.

Fig. 3 shows a small complementary collection of situations involving self-attacking arguments and a slightly larger number of arguments[4].

So as to understand how the argument strength measure behaves in general, we now provide a thorough mathematical study of its properties. This is the objective of the next section.

## 5. Properties of argument strength

We are going to examine three groups of properties. Properties in the first group (cf. propositions 1, 2 and 3) relate to the the boundedness of the measure and the characterisation of the conditions under which an argument's strength attains extreme values. Properties of the second group (cf. proposition 4) concerns arguments of medium to high strength and gives and links the notions of admissibility and stability to this domain of the strength spectrum. Finally, the third group of properties (cf. propositions 6, 7, 8 and 9) explain the impact of adding new attacks or arguments to an argumentation framework. This last group of properties thus allows to understand the evolution of the individual status of arguments in debates that are dynamically constructed.

Let us start with the first group of properties. By construction of the reward function in games of argumentation strategy, the strength of an argument is a real number which we can show to be bounded between 0 and 1. This is an almost direct consequence of proposition 1.1 and von Neumann's minimax theorem. In the next two propositions, we will see that these bounds are in fact tight and can be attained.

**Proposition 2 (bounds of argument strength)** *The strength of an argument is always comprised between zero and one.*

[4] As the size of the players strategy spaces grows exponentially fast with the total number of arguments in the framework considered, the optimisation technique we use at the moment to measure argument strength does not scale up to much more than a dozen of arguments

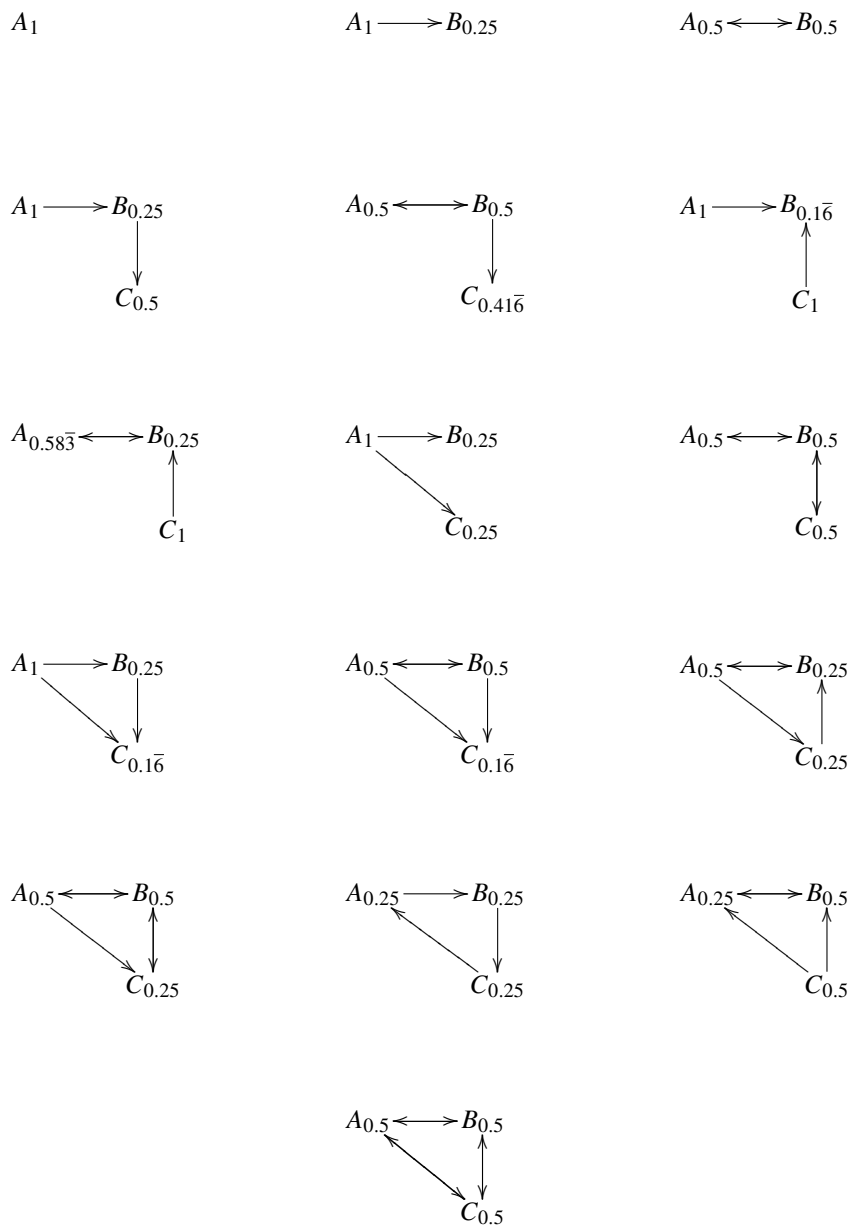


**Figure 3.** Catalogue of some basic configurations.

**Proof 1** *According to proposition 1.1,* $\forall(i, j), r_{i,j} \in 0, 1$. *For every mixed strategies* $X$ *and* $Y$, *we also have* $X^T RY \in 0, 1$, *which implies* $0 \leq \min_Y X^T RY$ *and* $\max_X X^T RY \leq 1$. *Therefore,* $0 \leq \max_X \min_Y X^T RY$ *and* $\min_Y \max_X X^T RY \leq 1$. *By the minimax theorem,* $0 \leq v \leq 1$, *and thus* $v = s_F(x) \in [0, 1]$.

The lowest possible strength of an argument is zero and this situation only occurs when

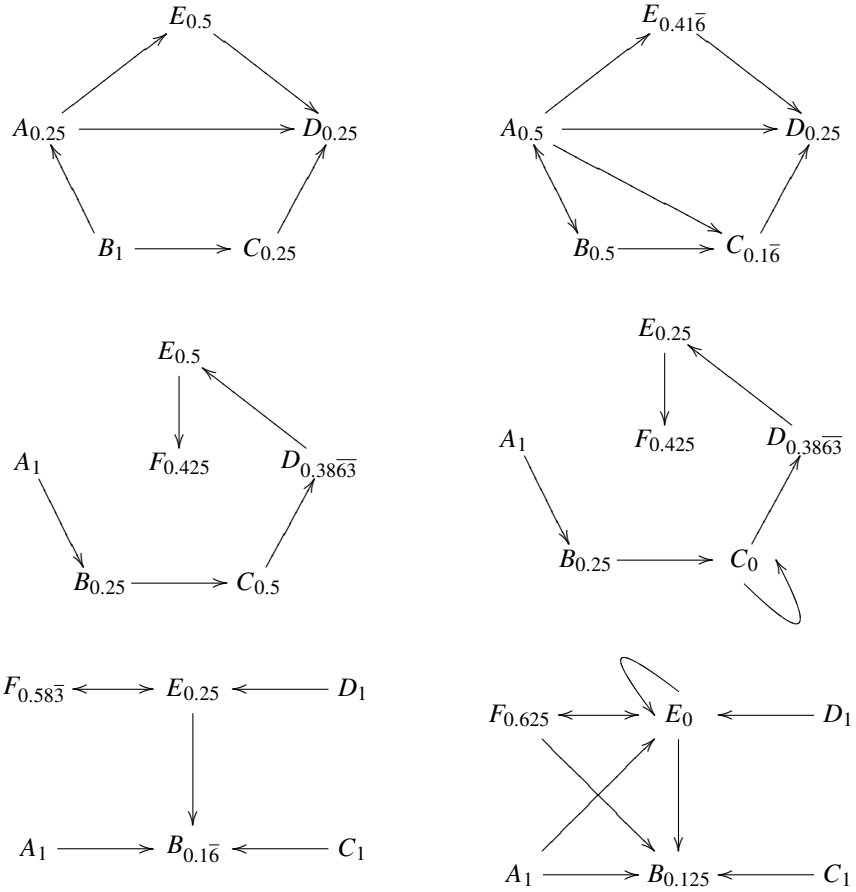


**Figure 4.** Sample of various and more complex debate configurations.

the argument attacks itself. This situation is however rare in practise. Remark also that whenever an argument attacks itself, its proponent is forced to play with strategies that all correspond to sets of arguments that attack themselves and that yield a null payoff irrespective of what the opponent does.

**Proposition 3 (self-contradiction must be avoided)** *The strength of an argument is null if and only if the argument attacks itself.*

**Proof 2** $\Rightarrow$*:* $s_F(x) = v = \min_Y \max_X X^T RY = 0$ *implies the existence of* $Y^*$ *such that* $\forall X$, $X^T RY^* \leq 0$. *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 1.2,* $\forall i$, $P_i$ *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 1.2),* $R = ((0))$ *and* $v = s_F(x) = 0$.

The highest possible strength of an argument is one and this situation only occurs when the argument is not attacked by any other argument. This situation is quite common. Indeed, the participants in a debate usually focus their attacks against one or a few arguments amongst the most important arguments of their opponents, leaving thus many arguments unattacked.

**Proposition 4 (unattacked arguments are the strongest)** *The strength of an argument is one if and only if it there is no argument attacking it.*

**Proof 3** $\Rightarrow$*: If* $s_F(x) = v = 1$*, then we have* $\max_X \min_Y X^T RY = 1$*. 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,* $\forall X$, $\min_{Y \in S} X^T RY \geq \min_Y X^T RY$*. Therefore,* $\max_X \min_{Y \in S} X^T RY \geq \max_X \min_Y X^T RY = 1$*. 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* $\forall (i, j)$, $r_{i,j} \leq 1$ *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 1.1),* $\forall j$, $r_{k,j} = 1$*. By proposition 1.2),* $\forall j$*,* $P_k$ *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* $s_F(x) = 1$*.*

Apart from these two extreme cases, all remaining arguments have by elimination a strength value that is strictly comprised between zero and one. Amongst them, those which are still admissible or stable (i.e. contained in an admissible or stable set of arguments) have a strength that can be shown to be always above average. This important property is the guarantee that our quantitative strengh measure does not contradict the qualitative analysis of arguments that can be conducted using the standard notions of acceptability for abstract argumentation. However, argument strength offers the possiblity to compare acceptable arguments amongst themselves. The second part of the following proposition shows indeed that the upper bound of the strength of an acceptable argument monotonically decreases with the number of attacks existing against it.

**Proposition 5 (acceptable arguments have medium to high strength)** *If an argument is admissible or stable*[5] *then its strength is greater or equal to* $\frac{1}{2}$*. However, the strength of an argument that has k attacks is always strictly inferior to* $1 - \frac{1}{2} f(k)$*.*

**Proof 4** *If there exist k attacks against x, then there exists a strategy O with k attacks against x. For this strategy, and whatever the proponent strategy P, there must be also at least k attacks from O against P and* $r_F(P, O) < 1 - \frac{1}{2} f(k)$ *by proposition 1.4. By playing O with a probability of* 1*, the opponent can strcitly secure a maximum loss of* $1 - \frac{1}{2} f(k)$*, i.e.* $s_F(x) < 1 - \frac{1}{2} f(k)$*. If P is admissible, then by proposition 1.3)* $\forall O$*,* $r_F(P, O) \geq \frac{1}{2}$ *so by playing P with probability* 1 *the proponent of x can secure a payoff of at least* $\frac{1}{2}$*. If P is stable, then it is also admissible and the same result holds.*

[5] This property can actually be generalised to any semantics of argumentation that is stronger than the notion of admissibility, such as e.g. the preferred, complete, grounded and ideal semantics.

We now study how the strength of arguments varies as argumentation frameworks evolve, from an initial argument and no attack to a larger set of arguments with more attacks. This is important as it allows to understand quantitatively the impact of adding new arguments and attacks to a controversial debate and may be used strategically by participants of a debate to spot weaknesses in their adversaries opinions and influence the deliberation process. Propositions 6, 7 and 8 concern the addition of attacks to a framework. Proposition 9 deals with the addition of arguments, and more generally of groups of arguments.

Suppose first that an attack $(a, b)$ is added to the framework $F = (Arg, att)$, where $(a, b) \notin att$ and $a, b \in Arg$. By convenience, we will use

**Notation 3** $F_{+(a,b)} = (Arg, att \cup \{(a, b)\})$

When an attack is added against an argument we intuitively expects its strength to be reduced or, in the best case, to be maintained to the same value. This is a very intuitive and desirable property since people raise attacks against arguments in debates especially to reduce their "strength" and impact on the deliberation process.

**Proposition 6 (criticism reduces argument strength)** *Adding an attack against an argument reduces its strength or maintains it in the best case.*

**Proof 5** *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. Remark 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 $s_{F_{+(a,b)}}(b) \leq s_F(b)$.*

When adding an attack from $a$ against $b$, we also increase the degree of "aggressiveness" of opinions embracing argument $a$ towards these opinions embracing $b$. If the proponent does not point this aggressiveness against himself (i.e. $b$ is not part of the optimal strategies played by the proponent of $a$), then the proponent should be better off because his optimal strategies become more stable (in the dialectical sense). In such cases, we expect the strength of $a$ to increase. On the opposite, if the extra aggressiveness is not well targeted, the strength of the argument may be reduced. To distinguish between these two cases, we say that

**Definition 6 (superfluous argument)** *Argument $b$ is superfluous w.r.t. argument $a$ if by forbidding the proponent of $a$ to play with strategies containing $b$ one does not decrease the value of the $(F, a)$ game of argumentation strategy.*

**Proposition 7 (cautious extra-aggressiveness pays-off)** *Adding an attack from $a$ against $b$ increases (or preserves) the strength of $a$ when $b$ is superfluous with respect to $a$ and may reduce it otherwise.*

**Proof 6** *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 $O_F^{\leftarrow P} \subseteq O_{F_{+(a,b)}}^{\leftarrow P}$, $P_F^{\leftarrow O} = P_{F_{+(a,b)}}^{\leftarrow O}$ (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 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, $s_{F_{+(a,b)}}(a) \geq s_F(a)$.*

Since adding an attack against argument $b$ weakens that argument, one expects to see an increase in the strength of the "enemies" of $b$, i.e. the arguments $c$ which are attacked by $b$. We can prove that this additional property holds in general.

**Proposition 8 (indirect counter-attack brings support)** *If $b$ attacks $c$, then adding an attack (from $a$) against $b$ increases the strength of $c$.*

**Proof 7** *The sets of strategies of the players are the same in the $(F, c)$ and $(F_{+(a,b)}, c)$ games. We have $O_F^{\leftarrow P} \subseteq O_{F_{+(a,b)}}^{\leftarrow P}$ (if $a \in P$ and $b \in O$) or $O_F^{\leftarrow P} = O_{F_{+(a,b)}}^{\leftarrow P}$ otherwise. We also have $P_F^{\leftarrow O} \subseteq P_{F_{+(a,b)}}^{\leftarrow O}$ (if $b \in P$ and $a \in O$) and $P_F^{\leftarrow O} = P_{F_{+(a,b)}}^{\leftarrow O}$ otherwise. Remark 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 $s_F(c) \leq s_{F_{+(a,b)}}(c)$.*

So far, we have looked at changes in the framework structure which only concerned the attack relationship between arguments, but we also need to study the impact of adding new arguments to a debate. Let us now assume that $Arg$ are the arguments of the current debate and $Arg'$ represent some new arguments, i.e. that $Arg' \cap Arg = \emptyset$. Obviously, both current and new arguments may be in conflict according to some distinct attack relations $att \subset Arg \times Arg$ and $att' \subseteq Arg' \times Arg'$. The argumentation framework resulting from the addition of the new arguments $Arg'$ is thus simply $F_{Arg+Arg'} = (Arg \cup Arg', att \cup att')$. We have

**Proposition 9 (insensitivity to irrelevant information)** *The strength of arguments in a debate $(Arg, att)$ is unchanged by the addition of new arguments from a debate $(Arg', att')$ that is irrelevant to it, i.e. that verifies $Arg' \cap Arg = \emptyset$.*

**Proof 8** *Let us consider the $(F_{Arg+Arg'}, x)$ game where $x \in Arg$. Since none of the arguments in $Arg'$ attack $x$ (the two frameworks are disconnected), the proponent of $x$ is at least as well off in this new game as in the $(F, x)$ if he restricts himself to his old set of strategies build upon $Arg$. Therefore, $s_{F_{Arg+Arg'}}(x) \geq s_F(x)$. The same proposition also holds for the opponent of $x$, which means that $-s_{F_{Arg+Arg'}} \geq -s_F(x)$ or equivalently $s_{F_{Arg+Arg'}} \leq s_F(x)$. In conclusion, $s_{F_{Arg+Arg'}}(x) = s_F(x)$.*

As intuitively expected, the status of arguments in a debate is left unchanged by adjunction of new arguments disconnected from the current debate. As mentionned earlier, irrelevant groups of arguments may be brought into the debate by experts relying on disconnected or independent empirical theories. This last result thus enables us to analyse these groups of arguments independently of each other. Thus, the formal mechanism for aggregating non mutually conflicting groups of arguments simply corresponds to the juxtaposition of weighted graphs.

## 6. Summary and discussion of related works

Abstract argumentation frameworks [Dung, 1995] constitute an elegant and simple way of representing knowledge and structuring conflicts in controversial debates. Existing notions of acceptability in abstract argumentation provide useful qualitative insight on the status of arguments within such frameworks. Very recently, the argumentation research community has manifested a sudden interest for the use of quantitative measures in the analysis of persuasion dialogues [Amgoud and de Saint-Cyr, 2008; Budzyńska *et al.*, 2008]. The idea of using games of strategy and their value – as defined in game theory – for constructing such measures brings a technical novelty with regards to such approaches. In this paper, the notion of argument strength has been put in relation with the class of games of strategy. However, it is important to note that other types of games can be useful for argumentation. For instance, it has been argued in [Riveret *et al.*, 2008] that argumentation dialogue games [Prakken, 2005] could most suitably been treated as extensive games, rather than strategic games.

To measure the strength of an argument, we have essentially suggested to confront a proponent and opponent of an argument via a repeated game of argumentation strategy. The game's payoffs have been chosen so as to reflect numerically the interaction between proponent and opponent strategies. The statistical equilibrium resulting from the long term interaction between their respective possible strategies can be modelled by the argumentation game's value, which we have used as measure of strength. In practise, argument strength may be most efficiently computed using the simplex algorithm of [Dantzig *et al.*, 1955]. It has been shown in order that argument strength ranges between zero and one, that these bounds are attained for arguments that attack themselves and that are not attacked respectively, that admissible and stable arguments have above average strength and that argument strength allows to make comparisons between acceptable arguments. We have also established that criticism reduces argument strength, that cautious extra-aggressiveness increases argument strength and indirect counter-attacks may restore an argument strength. Finally, we could prove that the addition of irrelevant groups of arguments to a debate does not have any impact on argument strength.

The authors of [Krause *et al.*, 1995] distinguish three types of argument strength measures, namely simple weights in a alphabet of symbols or in [0, 1], strength as probability of provability and qualitative measures based on purely symbolic form of dialectical argumentation. The measure here-exposed belongs to the first type, although derived from a purely symbolic representation of debates. [Ambler, 1996] rigorously discusses manipulations of the internal structure of arguments which allow to evaluate the strength of arguments. Our approach, based on Dung's abstract view of argumentation on the opposite cold-shoulders the internal structure of arguments and deliberately overlooks such information. Nevertheless, the results presented in the previous section show that the internal structure of arguments is not essential to obtain intuitively appealing results.

[Poole, 1993] has developed a rigorous notion of strength as probability of provability and [Amgoud and Prade, 2004] has developed an approach for assessing the strength of arguments rooted in possibility theory. In these works, a probability or possibility distribution is assumed to be known in advance and is used to model our confidence in elementary pieces of knowledge upon which arguments are constructed. The strength of arguments is intended to reflect an overall level of confidence into the propositions they support. Such paradigms may sometimes be impractical, as they require a subject

– often endowed with only qualitative and conflicting pieces of information – to fully specify probability or possibility distributions. The approach exposed here only requires the specification of a directed graph.

Finally, acceptability notions in abstract [Dung, 1995] or assumption-based argumentation [Bondarenko *et al.*, 1997; Dung *et al.*, 2006] constitute perhaps the most significant examples of qualitative measures of argument strength based on dialectical argumentation. We have exploited these notions to define numerical degrees of acceptability and rewards of games of argumentation strategy. We have observed in the moral debate on capital punishment and also justified with formal proofs, that the constructed measure matches with the intuition offered by standard acceptability notions, but also enables to get a slightly more detailed account of the status of arguments in controversial debates.

## References


[Ambler, 1996] Simon Ambler. A categorial approach to the semantics of argumentation. *Mathematical Structures in Computer Science*, 6(2):167–188, 1996.

[Amgoud and de Saint-Cyr, 2008] L. Amgoud and F. Dupin de Saint-Cyr. Measures for persuasion dialogs: A preliminary investigation. In *Second International Conference on Computational Models of Argument*, 2008.

[Amgoud and Prade, 2004] L. Amgoud and H. Prade. Using arguments for making decisions: A possibilistic logic approach. In *Proceedings 20th Conference of Uncertainty in AI*, pages 10–17, 2004.

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

[Borel, 1921] E. Borel. La théorie du jeu et les équations intégrales à noyau symétrique gauche. In *Comptes Rendus de l'Académie des Sciences*, 1921.

[Budzyńska *et al.*, 2008] K. Budzyńska, M. Kacprzak, and P. Rembelski. Modelling persuasiveness: Change of uncertainty through agents' interactions. In *Second International Conference on Computational Models of Argument*, 2008.

[Caminada, 2006] M. Caminada. Semi-stable semantics. In *Proceedings of 1st International Conference on Computational Models of Argument*, 2006.

[Dantzig *et al.*, 1955] G. B. Dantzig, A. Orden, and P. Wolfe. The generalized simplex method for minimizing a linear form under linear inequality constraints. *Pacific J. Math.*, 5(2):183–195, 1955.

[Dresher, 1981] M. Dresher. *The Mathematics of Games of Strategy*. Dover Publications, 1981.

[Dung *et al.*, 2006] P. M. Dung, R. A. Kowalski, and F. Toni. Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence*, 170(2):114–159, 2006.

[Dung *et al.*, 2007] P.M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence, Special Issue on Argumentation in Artificial Intelligence*, 171(10–15):642–674, 2007.

[Dung, 1995] P. M. Dung. 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.

[Hillier and Lieberman, 1995] F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research - 6th Edition*. McGraw-Hill, 1995.

[Krause *et al.*, 1995] P. Krause, S. Ambler, M. Elvang-Gøransson, and J. Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11:113–131, 1995.

[Matt and Toni, 2008] P.-A. Matt and F. Toni. Basic influence diagrams and the liberal stable semantics. In *Second International Conference on Computational Models of Argument*, 2008.

[Neumann, 1928] John Von Neumann. Zur theorie der gesellshaftsspiele. *Mathematical Annals*, 100:295–320, 1928.

[Poole, 1993] D. Poole. Probabilistic horn abduction and bayesian networks. *Artificial Intelligence*, 64(1):81–129, 1993.

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

[Riveret *et al.*, 2008] R. Riveret, H. Prakken, A. Rotolo, and G. Sartor. Heuristics in argumentation: A game-theoretical investigation. In *Second International Conference on Computational Models of Argument*, 2008.

[von Neumann and Morgenstern, 1944] J. von Neumann and O. Morgenstern. *Theory of games and economic behaviour*. Princeton UP, 1944.