Sign, then Ratify : Negotiating under Threshold Constraints

Sylvie Thoron

GREQAM

Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes Etudes en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail n°2006-50

Sign, then Ratify: Negotiating under Threshold Constraints

Sylvie THORON

December 2006

Sign, then Ratify:
Negotiating under Threshold Constraints[1][2]

Sylvie Thoron[b]

December 2006

Abstract :

The procedure for implementing any international treaty necessarily involves two steps. The negotiation phase which culminates in the signature of the treaty is followed by a ratification phase. This last phase is governed by a rule which determines how far the ratification process has to advance before the treaty can come into effect. The purpose of this paper is to analyse, using a game theoretical approach, the possible consequences of this minimum participation rule for the ratification phase and for the negotiation phase. I consider the case of International Environmental Agreements in which, during the negotiation phase, the different parties have to decide on the level of a global target and on how to share the efforts necessary to reach it.

I use a cooperative approach to define what is called the threshold value (T-value). For a given coalition of parties, the T-value gives the expected outcome of the negotiation over sharing a global target, when the parties take into account the minimum participation rule.

Given this T-value, I use a non-cooperative approach to determine which coalition will sign the agreement and what will be its global target. The minimum participation constraint has in fact no impact on the ratification phase because it is always better to refuse to sign rather than to sign and then refuse to ratify. However, I show that the minimum participation constraint can modify the outcome of the negotiation phase. Indeed, it plays a role in a mechanism which can be used by a coalition to signal its leadership commitment. I analyse the conditions under which, at the equilibrium, the leading coalition can provoke an expansion of the signing coalition.

Keywords: International Environmental Agreements. Negotiation. Minimum participation rule. Shapley value. Ratification. Leadership commitment.

JEL Classification C71, C72, D74, K33, N5

b GREQAM, Université de Toulon, France, thoron@univmed.fr

[1] Comments and suggestions on a previous version of this paper by Eric Maskin and Hervé Moulin are gratefully acknowledged. I also benefited a lot from discussions with Alan Kirman. Finally, I would like to thanks Jana von Stein for her helpful comments.

[2] While this paper has been written, I was a visitor at the Institute for Advanced Study in Princeton and at the Department of Economics at Princeton University. I want to thank the Department for its hospitality and the Institute for its open mind atmosphere.

When Martina Navratilova announced her retirement from tennis
a journalist asked her:
« - When you retire, will you still be involved in tennis? »
« - I will not just be involved, I am committed to tennis. »
« - What is the difference? »
« - Think of ham and eggs.
The chicken is involved but the pig is committed. »

1 Introduction

Contrary to popular belief, international treaties are not negotiated, then signed and then enter into force. The situation is, in fact, more complicated, particularly in the case of multilateral agreements. In accordance with the 1969 Vienna Convention on treaty law, the procedure for implementing any international treaty must involve two steps. The first or negotiation phase culminates in the signature of the treaty by the different parties. In the particular case of the Kyoto protocol, for example, the parties had to agree on the amount by which total greenhouse gas emissions should be reduced and to decide on how to split the efforts necessary to achieve this amount. The second or ratification phase requires the ratification by the appropriate national bodies who thus give the official consent of each country to the agreement. When the agreement is multilateral, there may be special requirements specifying what stage in the ratification process has to be reached before the agreement enters into force. The ratification phase is then governed by a minimum participation constraint which plays the role of a threshold. Depending on the treaty in question, this threshold is defined in very different ways: as a set of countries, a number of countries or a percentage of the targeted polluting substances (cf. Barret (2003) pages 165-195). Sometimes, the rule is a combination of the three. In the particular case of the Kyoto protocol, two conditions had to be met before the agreement came into effect. The first rule was that at least 55 countries which originally signed had also to have ratified. The second rule required that the countries which had ratified must have accounted for at least 55% of the pollution produced in the reference year (1990).

The rule governing the implementation of an international agreement is therefore nothing other than the imposition of a threshold which must be met before the treaty comes into effect. The question I address in this paper concerns the consequences of this minimum participation constraint or threshold for the ratification phase *and* for the negotiation phase.

Such thresholds are often invoked in the literature on games of contribution to a public good, for example. It is argued there that they are incentive compatible (see for example the mechanisms proposed by Bagnoli and Lipman (1989), Palfrey and Rosenthal (1984), or the experimental evidence presented in a survey by Ledyard (1995)). What the negotiators are seeking when they decide on the rule which will govern the ratification phase is to induce the highest

number of countries to participate in the agreement. In a book which describes the negotiations leading to the Montreal Protocol on protection of the ozone Layer, Ambassador Richard Benedick (1998) explains that the U.S. started by proposing a very high threshold:

"There was concern that the United States could, in a situation analogous to its unilateral 1978 action, find itself bound to the obligations of an "international" protocol while its major competitors were not. As a legacy of the domestic debate, some U.S. agencies insisted on pushing for a proportion of consumption of 90 percent or higher as the trigger for entry into force and other actions."

But as a the result of the negotiation:

"An inevitable, and reasonable, compromise was struck in Montreal, providing that entry into force would require ratification by at least 11 parties, together constituting at least two-thirds of estimated global consumption of controlled substances as of 1986 (article 16)[...] Most observers believed that this would provide a sufficient critical mass to increase the pressure on any potential large holdouts to join the treaty."

Indeed, in the political science literature, different arguments can be found, to sustain the idea that the decision to ratify is, in part, driven by the actions of other states: if other governments are ratifying en masse, governments have an incentive to ratify as well (von Stein (2006)). An argument developed by Finnemore and Sikkink (1998) is that at a certain "tipping point" in a norm's evolution, a "norm cascade" takes place, and once ratification becomes the norm, states ratify in large numbers because of pressure from other states and non-state actors. However, these different aguments do not take into account the simple fact that the countries which have to ratify have already signed the agreement. To say that they have to take a decision again about their participation when they have to ratify amounts to saying that there is a strategic exploitation of this two step procedure. Some countries may have an incentive to sign, in order to encourage the others to do the same, whilst knowing that they themselves will not ratify in order to free ride. Will they always do this? In the model developed in this paper it will be the case that it is better not to sign rather than to sign and not ratify.

However, when Benedick speaks about a pressure on any potential large holdouts to join the treaty, this can be understood in a different way. This does not necessarily refer to the ratification process but can concern the paticipation in the first negotiation phase and signature. I will focus on the problem of participation, analyzing the consequences of the minimum participation rule during the negotiation. I will claim that, in the case of multilateral agreements, the different parties take the rule of the ratification phase into account while they are negotiating their contributions. The negotiators clearly understand

that they can use the minimum participation constraint strategically. With regard to the 90 percent proposal of the United States at Montreal, Benedick (1998) says:

"Many observers feared that such a requirement could hold the treaty hostage to Japan or the Soviet Union, which might then weaken the protocol by extracting other concessions as a price for adherence."

Barret's (2003) interpretation of this observation is that these countries could exploit a pivotal position to increase their weight in the negotiation. He says (page 319):

"So why was the two-thirds hurdle preferable to the 90 percent threshold? There are, I think, two reasons. The first is that the 90 percent threshold would have given *bargaining power* to the USSR and Japan. If the two-thirds hurdle were satisfied and the treaty entered into force, then it probably would have been in the interests of the USSR and Japan to join [...] However, if the 90 percent threshold were not satisfied, entry by the USSR and Japan would decide *whether* the treaty entered into force [...] Hence, [these countries] might have used the higher threshold to obtain concessions."

Therefore, he considers that the *bargaining power* of the different parties to the negotiation is measured by their *voting power,* if we interpret the ratification as a vote. However, if this were the case, the same argument should be applied to the big CFC-consumers. The European Union and the United States had the greatest bargaining power, in both respects. Did these two parties try to weaken the treaty nevertheless? In fact the opposite was true, the U.S. and the EU were considered as the leaders in the negotiation and their high final contributions to the treaty do not support this argument.

My claim is that the difference between the U.S. and the E.U. on one hand and the USSR and Japan on the other hand was not that the latter had more bargaining power than the former but that the U.S. and the E.U. were more committed. The way in which the threshold is defined can be interpreted as a signal about the degree to which the different countries are committed. Of course, this depends on the kind of rule considered. When the minimum participation rule is defined as a number of countries there is no way, for one country or the other, to play a specific role and the threshold cannot be used as a signal about commitment. At the other extreme, when the rule is a list of countries, it is clear that these countries are committed to the agreement. In the case of the rule used in the Montreal protocol or the Kyoto protocol there is a signal but one which is noisy.

Another signal about a country's commitment is its contribution to the agreement. Heike Schröder was an official note maker during the negotiations

of the Kyoto protocol. Commenting on the result of the negotiation about the reduction commitments she reports that:

"With 8 per cent of 1990 levels, the EU retains its leadership commitment, albeit on a very narrow margin. The US consented to a 7 per cent reduction target, which is a substantial commitment given its increase in emissions since 1990 of about 23 per cent." (Schröder (2001) p. 79 and 80)

But what is the link between these two sorts of signals? In this paper I give a game theoretical explanation of the effect the existence of a given threshold. I show that it can modify the result of a negotiation in terms of the differences between the countries' contributions and also in terms of the level of participation. The intuition is the following. Once the countries which have ratified satisfy the minimum participation constraint, the agreement is implemented by those countries. At this point, the countries which will ratify later will benefit from this implementation by the others. Indeed, the cost of implementing the agreement is much higher for the initiating countries[1]. If it is likely that one country will be among the first to ratify, it is also likely that this country will bear a higher cost in implementing the agreement. In other words, the *leadership commitment* mentioned by Schröder can be understood in a literal sense. As a consequence, such a country will enter into the negotiation with a generous attitude.

If the interpretation I have given makes sense, we should find a relationship between a country's position with respect to the threshold rule (is it necessary or not for example) and its involvement in the negotiation or its willingness to contribute to the agreement. Unfortunately, the latter is very difficult to measure. However, the Kyoto protocol constitutes a case which is interesting and unique from this point of view. First, its minimum participation constraint is a rather sophisticated double rule. The first rule which specifies a number of countries is the most common rule applied in environmental treaties. The second rule which specifies a volume of the polluting substances targeted is much rarer, and it is even more unusual to have a combination of this rule with another rule. As far as I know, the Montreal protocol is the only other environmental agreement with a double rule of this sort (see Barret's list (2001) pages 165-195 and the site of the Environmental Treaties and Resource Indicators). Furthermore, the Kyoto protocol had another characteristic, which is that the countries have decided to fix not only a global target but also individual targets, called the *Assigned Amounts*. These assigned amounts and the way they differ from the global target can be considered as a proxy for the different countries' involvement in the negotiation. The 5.5% decrease of global greenhouse gas emissions is shared in a very unequal way (see Table 1 in Section 2). For example, in terms of its own emissions, the contribution of the EU (responsible for 24% of total CO_2 emissions) is -8%, the contribution of the US (36% of total CO_2

[1] See for example the paper by G. Heal (1993) in which he gives a cost side analysis.

emissions) is -7% while the contribution of Russia (17.5% of total CO_2 emissions) is 0% and Australia, which is responsible for just a little more than 2% of emissions can *increase* its pollution by 8 percent. Different justifications, historical, technological, and so forth, for these inequalities can be found in the environmental literature. However it is worth noting that the biggest polluters which are necessary to reach the threshold, are the biggest contributors.

How can we analyze this ex-ante effect of a threshold on the result of a negotiation? Here, I adopt a cooperative approach. Cooperative game theory proposes different tools, solution concepts, to share a worth or a cost based on different principles. For example, the Shapley value is based on the principle that each player is remunerated according to its incremental worth, that is, to its contribution to the contribution of a given coalition. Then, for a given player, her Shapley value is her expected incremental contribution, when it is assumed that the different orders in which the partners join the coalition are equally probable. The Shapley value has been proposed "to evaluate the players' prospects"- Hart and Kurz (1983) p1047. Let me also quote Shapley (1953 p. 307):

"At the foundation of the theory of games is the assumption that the players of a game can evaluate, in their utility scales, every "prospect" that might arise as a result of a play. In attempting to apply the theory to any field, one would normally expect to be permitted to include, in the class of "prospects," the prospect of having to play a game. The possibility of evaluating games is therefore of critical importance."

The idea that a value can be interpreted as the expected outcome of a negotiation has been explicitly developed by Hart and Kurz. Since then, there have been different attempts in the non cooperative literature to prove what was originally just an interpretation: that the Shapley value can be considered as the expected outcome of a negotiation (see for example Perez-Castrillo and Wettstein (2001) or Maskin (2004)). Indeed empirical studies had already corroborated this interpretation (see for example Littlechild and Thompson (1979)). In this paper, I will not try to propose a non-cooperative game to describe the negotiation. I will adopt the cooperative approach keeping in mind the interpretation that the value corresponds to an evaluation of players' prospects.

Another important aspect of the interpretation of the Shapley value concerns the *weights* of the different players in the negotiation. The value is defined for a given cooperative game (N, v) and $v(N)$ is then the amount the players have to share. In our specific framework $v(N)$ can be for example the cost of pollution control and, in general, is the total cost of implementing the agreement. The coalitional function v represents the contribution to this cost of every coalition. However, it cannot reflect all the elements which could play a role in the negotiation. These elements are incorporated in the *weights* that the different players have in the negotiation. Kalai and Samet (1988) say that: "The weights

should be determined by considering such factors as bargaining ability, patience rates, or past experiences". In the Shapley value, everything is considered to be incorporated in the coalitional function and these weights are therefore symmetrical. By contrast, a weighted Shapley value is defined for a vector of exogenous weights. Kalai and Samet (1985) have proposed an axiomatization of the family of weighted Shapley values. Hart and Kurz (1983) have proposed a coalition structure value which represents the Shapley value players can expect when they form coalitions. Intuitively, this is what the players can expect when they form coalitions to increase their weight in the negotiation: "Our view is that the reason coalitions form is not in order to get their worth, but to be in a better position when bargaining with the others on how to divide the maximal amount available." (p. 1052). In this paper, I will consider that the different countries have different weights because some of them want to be leaders in the negotiation. Whilst, in the coalition structure value, players form coalitions to increase their weights in the negotiation, here countries form a leading coalition to decrease their share of the gain or to increase their contribution to the cost. Why should they behave in this way? Because, by doing so, they can increase the level of participation, convincing additional partners to sign a more favorable agreement. This aspect cannot be understood in a cooperative framework. In the second part of this paper, I use a non-cooperative approach to describe the agreement formation.

In Section 2, I propose a simple environmental cooperative game with three countries, to illustrate how the minimum participation constraint can modify the expected outcome of the negotiation. The *threshold value (T-value)*, which is defined for any cooperative game and for any given threshold is then formally presented in Section 3. The T-value is then used to represent the expected outcome of an environmental agreement negotiation when a threshold govern the ratification phase. Section 4 presents three different non-cooperative models of agreement formation. In the simplest game the countries only have to decide whether to participate in the agreement. A ratification phase is introduced in the second game. It is shown that, under mild assumptions, the ratification phase is always completed. In the third game a coalition can use the minimum participation rule to signal its leadership commitment. Then, the T-value represent the situation in which the leading coalition is willing to contribute more in order to provoke an expansion of the agreement. Section 5 concludes.

2 How should the efforts to reach a global target be shared?

2.1 Who gets the best deal?

In economic theory in general and in cooperative game theory in particular, we are used to analyse sharing rules. In real environmental treaties the problem is more complex since the outcome of the negotiation is generaly a global target,

a time table and various rules which are common to every party. However the Kyoto protocol is particular from this point of view since it gives country-specific targets. Therefore, it is possible to measure the different parties' involvement in the protocol comparing the different specific targets. In Table 1 it can be seen that the biggest parties to the Kyoto protocol were assigned rather stringent targets, whereas smaller countries got more lenient targets. The size of the different parties is here measured by their volume of green house gas emmissions.

country	gg emissions	percent total	target
Australia	448,71	0,026293337	1,08
Canada	663,39	0,038873074	0,94
Iceland	3,17	0,000185754	1,1
Japan	1351,83	0,079214018	0,94
New Zealand	65,66	0,00384752	1
Norway	52,78	0,003092782	1,01
Russian Federation	1951,81	0,114371417	1
Switzerland	51,5	0,003017777	0,92
Ukraine	577,18	0,033821373	1
United States	6614,85	0,387614456	0,93
EU25	5261,47	0,308309611	0,923

Figure 1 plots Country-specific targets in the Kyoto protocol against the percentage of the parties' total green house emissions. The relationship appears to be decreasing.

This does not correspond to what the theory usually predicts nor to the normal intuition in this kind of situation. The usual idea is that biggest players get the best "deal" because their participation is essential if treaty is to enter into force. Why does this interpretation of bargaining power not correspond to what we observe in the case of the Kyoto protocol? In this Section, I will use a cooperative approach to show how the minimum participation constraint can modify the sharing rule. My claim is that, the definition of the threshold necessary for a treaty to enter into force can be interpreted as a signal about how committed different countries are to the treaty. In the case of the Kyoto protocol, does the fact that the threshold was defined as a volume of pollution produced mean that the biggest actors were more committed? We know that the European Union played a key role (see for example the book by Gupta and Grubb (2000) Climate Change and European Leadership*)*. The case of the USA is more controversial. The fact that they refused to ratify seems to show a posteriori that they were not committed. However, as Schröder (2001) explains in her report on the negotiations, the USA were rather active during the discussion about the assigned amounts. She says (page 79):

"On 8 December, US Vice President Al Gore made a 12-hour trip to the Kyoto Confrence where he instructed his delegation "to show increased negotiating flexibility if a comprehensive plan can be put into place". This message may have swayed US negotiators' willingness to accept higher targets".

Figure 1:

2.2 The Shapley value of the environmental game

To illustrate the ex-ante effect of the minimum participation rule on the result of the negotiation, I propose to consider an environmental agreement as a cost sharing game. Let the set of players be a set of three countries $N = \{1,2,3\}$, which want to decrease their total pollution: $TP = 610$ units by 10 percent. This is what will be called *the global target*. Assume that the marginal cost of decreasing pollution is decreasing, the cost of reducing pollution by R units is $C(R) = \sqrt{R}$. Therefore, the three countries have to share a cost $TC = 7.81$. At the outset, each country i generates a percentage of the total pollution TP. Let us assume that each county's pollution is, respectively, $P_1 = 360 = 59\%TP$, $P_2 = 160 = 26.23\%TP$, $P_3 = 90 = 14.75\%TP$. Given the 10 percent target, the characteristic function of the sharing cost game is given as follows:

$$\forall S \subset N, c(S) = \sqrt{10\% \sum_{i \in S} P_i}$$

Therefore, $c(S)$ represents the cost borne by coalition S when it decreases its pollution by 10% on its own. Obviously, this cost sharing game is sub-additive. For each country, the cost of reducing its pollution by 10 percent is smaller if the other countries do the same thing. This simple characteristic function captures the idea that cooperation of countries increases their efficiency in decreasing pollution. The Shapley value for this game is, for each country, its expected

incremental cost (expected IC), given that the orders in which the countries join the coalition are equally probable. An order r is defined on the set of countries N. The set of all possible orders is denoted by $R(N)$. For a given order $r \in R(N)$, each country i's ranking is denoted by r_i. The Shapley value can then be calculated with the following Table:

	Orders							
first	1	1	2	2	3	3		
second	2	3	1	3	1	2		
third	3	2	3	1	2	1		
							Sum	Shap. val.
Country 1's IC	6	6	3.21	2.81	3.71	2.81	24.54	$\frac{24.54}{6}$
Country 2's IC	1.21	1.1	4	4	1.1	2	13.41	$\frac{13.41}{6}$
Country 3's IC	0.6	0.71	0.6	1	3	3	8.91	$\frac{8.91}{6}$

$$\begin{aligned} \varphi_1(v) &= 4.03 = 51.6\%\ TC; \\ \varphi_2(v) &= 2.23 = 28.55\%\ TC; \\ \varphi_3(v) &= 1.48 = 18.95\%\ TC \end{aligned}$$

Note that the contribution increases with the country's volume of pollution but less than proportionally. Now, suppose that to each country is allocated a reduction in pollution in proportion to its contribution to the cost. This gives:

Country 1's pollution reduction: $\frac{4.03}{7.8}360 = 31.93 = 8.86\%$.
Country 2's pollution reduction: $\frac{2.23}{7.8}160 = 17.46 = 10.9\%$.
Country 3's pollution reduction: $\frac{1.48}{7.8}90 = 11.6 = 12.86\%$.

Here again, note that the biggest country reduces its pollution by the smallest percentage. This is a direct consequence of the cost fonction concavity. The small country benefits much more from the cooperation of the big country than the opposite. As a consequence, the big country has a higher bargaining power and gets the better deal in the partnership.

2.3 Ratification Process and Thresholds

Now, consider that the game is modified by the introduction of a ratification process. Indeed, the decisions taken by the different countries' delegates during the international negotiation have to be approved by the national institutions. Given this new process, different decisional mechanisms could be considered.

First, consider that the three countries sign an agreement in which it is specified that the signatories are committed to reach a target of a 10% decrease of their total pollution emissions and which gives the rule which will govern the ratification process. The rule says that the agreement will be considered to be

ratified if and only if the countries which have already ratified meet a minimum participation constraint. Furthermore, as soon as the agreement is ratified it will be implemented by those who have ratified it. What does this mean? Once the threshold is reached, the countries which have ratified and form a coalition S negotiate the sharing of the efforts needed to reach their own common target. I will call this *an ex-post negotiation*. The expected outcome of this ex-post negotiation can be considered to be the Shapley value $\varphi(c_S)$, but it depends on the threshold level specified in the ratification rule and on the order in which the three countries ratify.

An extreme case is the case in which the agreement can only be implemented if all the countries have ratified. Then, at the end of the ratification phase, the three countries are exactly in the same situation as before (cf. sub-section 2.1) and the outcome is the Shapley value $\varphi(c)$, whatever the order in which the countries ratify.

Now, consider another extreme case in which the only condition is that Country 1, the biggest country, has ratified. Then, if it is the first country to ratify, it will have to bear the biggest cost of $v(1) = 6$. If it is the second country to ratify, after country 2 it will bear a cost $\varphi_1(c_{12}) = 4.6$, after country 3, $\varphi_1(c_{13}) = 4.85$. If it is the third country to ratify, its cost share is, as in the previous case, its Shapley value.

When the agreement can be implemented as soon as the two bigest polluters have ratified, two scenarii can occur. First scenario: countries 1 and 2 are the first countries to ratify and the threshold is reached without country 3. Second scenario: country 3 is the first or the second country to ratify and the threshold is only reached when all the countries have ratified. We will denote by $T(\tau) = \{N, 12\}$ the set of coalitions of countries which can reach the threshold. Thus, to each random order of ratification corresponds one or the other of the two coalitions. In the first scenario countries 1 and 2 have to implement the agreement without country 3. In order to do so, they decide to share the common cost of decreasing their pollution by 10%: $c(12) = \sqrt{10\%(P_1 + P_2)} = 7.2$ using the Shapley value for their two person game. Their contributions are then:

$$\begin{aligned} T &= \{N, 12\}, \text{ first scenario} \\ \varphi_1(v_{12}) &= \frac{6 + 3.2}{2} = 4.6 \\ \varphi_2(v_{12}) &= \frac{4 + 1.2}{2} = 2.6 \end{aligned}$$

When it is country 3's turn to ratify, then, its incremental cost is just $c(123) - c(12) = 7.81 - 7.2 = 0.61$. However note that, in the second scenario, in spite of the threshold, the sharing rule will be $\varphi(c)$.

Now, consider the case in which the minimum participation constraint is not defined as a coalition of designated countries but *a percentage of pollution*. For example, if this threshold is $\tau, 60\% < \tau \leq 65\%$ both the coalition of countries 1 and 2 and the coalition of countries 1 and 3 can reach the threshold and $T(\tau) = \{N, 12, 13\}$.

Then, there are three different scenari. In the orders $r = (r_1 = 1^{rst}; r_2 = 2^{nd}; r_3 = 3^{rd})$ or $r' = (r'_1 = 2^{nd}; r'_2 = 1^{rst}; r'_3 = 3^{rd})$ countries 1 and 2 ratify first and implement the agreement. Their contributions are:

$$\varphi_1(c_{12}) = 4.6 \text{ and } \varphi_2(c_{12}) = 2.6.$$

In the orders $\widetilde{r} = (r_1 = 1; r_2 = 3; r_3 = 2)$ and $\widetilde{r}' = (r_1 = 3; r_2 = 1; r_3 = 2)$, countries 1 and 3 ratify first and implement the agreement. Their target is $C(13) = 6.7$ and their contributions:

$$\varphi_1(c_{13}) = 4.85 \text{ and } \varphi_3(c_{13}) = 1.85.$$

Lastly, in the orders $\widehat{r} = (r_1 = 3; r_2 = 2; r_3 = 1)$; $\widehat{r}' = (r_1 = 2; r_2 = 3; r_3 = 1)$ the agreement is only implemented when the three countries have ratified and the contributions correspond to the Shapley value.

However, the sequentiality of decisions described here is not consistent with what happens in the case of a real environmental agreement.

2.4 Putting the Cart before the Horse

In the case of a real environmental agreement, the sharing of the efforts necessary to reach the global target is decided prior to the ratification process, before the countries know which of them will be active first. Therefore, let us modify the previous game as follows. The objective of the international negotiation is to decide how to share the efforts needed to reach the global target, given that it will be followed by a ratification phase. I call this *an ex-ante negotiation*. I make two assumptions about the ratification process and the countries' conjectures about how it will go.

First, I assume that ratification always occurs. In the long term, all the countries will have ratified. However, in our framework there is no explicit representation of timing. The only aspect which matters is the order in which the different countries ratify. This matters because a country is not in the same position when it has to start the implementation of the agreement or when it can join a group of countries which have already started. Of course this is only true if there is a delay between the time the agreement enters into force and the time at which the additional country ratifies. Therefore, the assumption is that this delay is enough for the first countries to bear a higher cost of implementing the agreement and for the additional country to benefit from the decrease in marginal cost.

The second assumption is with regard to the conjectures of the countries' delegates who participate in the negotiation. I assume that the participants in the international negotiation have no idea about the order in which the different countries will ratify. This is not a strong assumption since the delegates who participate in the international negotiation are not the same individuals as those who participate in the national process of ratification. The delay between the two phases is another justification for this assumption. Therefore, all the countries will ratify but the delegates do not know in which order and they

attribute the same probability to each one of these orders. We can find different arguments along these lines in the literature on political science.

However the negotiating parties know the rule which will govern the ratification process. First note that, without thresholds the situation is similar to that we described in subsection 2.1, where we presented the game without a ratification process. The sharing rule decided during the international negotiation is just postponed until the end of the ratification process. The same thing is true if the threshold is so high that it can only be reached when the three countries have ratified. Here, this means that the threshold τ is such that $\tau > 75\%TP$. Therefore, in this situation the expected outcome of the negotiation is the Shapley value $\varphi(c)$.

The question is now to know what happens when a binding threshold is introduced. Then, as we saw in 2.2, the outcome of an ex-post negotiation depends on the order of ratification. Now, let us consider the expectation of the outcome of the ex-post negotiation, given that the different possible orders of countries ratifying are equally probable. My claim is that this is the outcome of the ex-ante negotiation and this define a value for the cooperative game (N, c) and for the given threshold. I will denote by $\phi(c, T)$ what I will call the *threshold value*. This value can be interpreted as follows: It is the delegates' "prospects" in the ex-ante negotiation when the outcomes of what would be ex-post negotiations are common knowledge. It does not mean that these ex-post negotiations will occur. It means that *the countries use what would be the outcome of an ex-post negotiation as an argument in the ex-ante negotiation.* I will define the T-value precisely in the following sections but let us see what happens in our example.

The first case I will consider is the case in which the threshold is high. That is when:

$$T = \{N, 12\} \text{ or } 65\% < \tau \leq 75\%,$$

In that case we saw in the previous sub-section that in the first scenario, that is in two orders out of six, countries 1 and 2 contribute more than their Shapley value, respectively, $\varphi_1(c_{12}) = 4.6 > 4.08$ and $\varphi_2(c_{12}) = 2.6 > 2.23$. On the other hand, country 3 contributes much less than its Shapley value $c(123) - c(12) = 0.61 < 1.48$. In the other scenario the three countries contribute their Shapley value. Therefore, the threshold value can be calculated as shown in the following table:

	Orders							
first	**1**	**1**	**2**	**2**	**3**	**3**		
second	**2**	**3**	**1**	**3**	**1**	**2**		
third	3	**2**	3	**1**	**2**	**1**		
							Sum	T-value
1's S-value[2]	4.6	4.03	4.6	4.03	4.03	4.03	24.54	$\frac{24.54}{6}$
2's S-value	2.61	1.1	2.61	3	1.1	3	13.41	$\frac{13.41}{6}$
3's S-value	0.6	1.85	0.6	2	1.85	2	8.91	$\frac{8.91}{6}$

[2] The Shapley value is calculated, for Country 1 in each given order.

$$\phi_1(c,\tau) = \frac{4*4.08+2*4.6}{6} = 4.25$$
$$\phi_2(c,\tau) = \frac{4*2.23+2*2.6}{6} = 2.35$$
$$\phi_3(c,\tau) = \frac{4*1.48+2*0.61}{6} = 0.99$$

Note that, in each given ratification order the global target is reached. However, the contribution of country 3 which may be able to ratify after the implementation of the agreement is less than its Shapley value. Country 3 gains at the expense of countries 1 and 2.

The second case is the case in which the threshold is at an intermediate level. That is when:

$$60\% < \tau \leq 65\% \text{ or } T(\tau) = \{N, 12, 13\}$$

Then, the threshold value is:

$$\phi_1(c,\tau) = \frac{2*4.6+2*4.85+2*4.08}{6} = 4.51$$
$$\phi_2(c,\tau) = \frac{2*2.6+2*2.23+2*1.11}{6} = 1.98$$
$$\phi_3(c,\tau) = \frac{2*1.48+2*0.61+2*1.85}{6} = 1.31$$

In that case, countries 2 and 3 gain at the expense of country 1. Country 1, which always participates in the first implementation of the agreement, is the biggest contributor.

In the last case Country 1's participation is the only necessary condition for the agreement to enter into force. The T-value is decribed in the following table:

	Orders							
first	**1**	**1**	**2**	**2**	**3**	**3**		
second	2	3	**1**	**3**	**1**	**2**		
third	3	2	3	**1**	2	**1**		
							Sum	T-value
1's order value	6	6	4.6	4.03	4.86	4.03	29.52	$\frac{29.52}{6}$
2's order value	1.15	1.15	2.61	2.23	1.1	2.23	8.24	$\frac{8.24}{6}$
3's order value	0.65	0.65	0.6	1.48	1.85	1.48	6.71	$\frac{6.71}{6}$

$$\begin{aligned}\phi_1(c,1) &= 4.92 = 63\%\ TC;\\ \phi_2(c,1) &= 1.37 = 17.58\%\ TC;\\ \phi_3(c,1) &= 1.12 = 14.31\%\ TC\end{aligned}$$

3 The T-value

3.1 Notation and definitions

In this sub-section, well known and new concepts and results which will be used in the following Section will be presented. Some of them are new, others are standard. Consider a set of players U. A coalition of players is a subset $S \subset U$. A coalitional function game (U, v) is defined for a given coalitional function v, which associates *a worth* $v(S)$ to each coalition $S \subset U$. I will only consider coalitional function games which are *superadditive*, that is, for two disjoint coalitions S and $T \subset U$, $v(S) + v(T) \leq v(S \cup T)$. The incremental worth of player i to coalition $S, i \in S$ is $v(S) - v(S \backslash i)$. A *null player* is a player i whose incremental worth to each coalition is zero: $v(S \cup i) - v(S) = 0, \forall S \subset U$. A sub-set $N \subset U$ is called *a carrier* of game (U, v) if $\forall i \in U \backslash N$, i is a null player. The property of superadditivity captures an important characteristic of environmental agreements. The interpretation of superadditivity is straighforward and this concept is often used in the literature on agreement formation. Indeed, it is a common idea that cooperation, generating synergies or externalities, can be represented by a superadditive game. I will also use further below the following properties of the coalitional function:

Definition 1 *The coalitional function game (U, v) is concave (convex) if, for each player $i \in U$,*

$$v(T) - v(T \backslash i) \leq (\geq) v(S) - v(S \backslash i), \forall S \subset T, i \in S$$

A coalition game is concave if, for each player, her incremental worth decreases when additional players are added to the coalition she joins. *A value* of the coalition game (U, v) is a function which associates with each player i in U a real number. For example, the Shapley value can be interpreted as a mutual acceptable sharing of $v(N)$ when the players agree in the following principles:

1- Each player must be remunerated at the level of her incremental worth.

2- This incremental worth depends on the group she joins.

3- An average incremental worth must be calculated taking into account the order in which the different players join the group. The different orders have the same weight (or are equiprobable).

Formally, for every game in coalitional function form (U, v), the Shapley value can be characterized by three very simple axioms (cf Shapley (1953)).

Axiom 1 *Efficiency and null player. If $N \subset U$ is a carrier of game v, then:*

$$\sum_{i \in N} \varphi_i(v) = v(N)$$

Axiom 2 *Additivity. If (U, w) is defined such that $w(S) = v(S) + u(S), \forall S \subset U$ then:*

$$\varphi(w) = \varphi(v) + \varphi(u)$$

Axiom 3 *Anonymity. If* π *is a permutation of* U *and* πv *defined such that* $\pi v\left(\pi S\right) = v(S)$, $\forall S \subset U$ *then:*

$$\varphi_{\pi i}\left(\pi v\right) = \varphi_i\left(v\right)$$

Definition 2 $\forall K \subseteq U, \forall \alpha \in \mathbb{R}$, *a unanimity game* αU_K *can be defined as follows:*

$$\forall S \subset U, v(S) = \begin{cases} \alpha \text{ if } K \subset S \\ 0 \text{ otherwise} \end{cases}$$

In an axiomatization proposed by Aumann (2005), Anonymity and Null player are replaced by the following Axiom:

Axiom 4 $\forall K \subseteq U, \forall \alpha \in \mathbb{R}, \varphi_i\left(\alpha U_K\right) = \begin{cases} \frac{\alpha}{|K|}, \text{ if } i \in K; \\ 0 \text{ otherwise} \end{cases}$

If N is a finite carrier of game (U, v), the Shapley value for player i can be calculated as her expected incremental value over all possible orders defined on N, under the assumption that each order appears with the same probability:

$$\begin{aligned} \forall i &\in N, \\ \varphi_i\left(v\right) &= \frac{\sum_{S \subset N, i \in S}\left(s-1\right)!\left(n-s\right)!\left(v\left(S\right) - v\left(S \backslash i\right)\right)}{n!} \end{aligned}$$

For each coalition $M \subset N$, define a coalitional function v_M such that:

$$v_M\left(S\right) = v\left(S \cap M\right)$$

Then:

$$\begin{aligned} \forall i &\in M \subset N, \\ \varphi_i\left(v_M\right) &= \frac{\sum_{S \subset M, i \in S}\left(s-1\right)!\left(m-s\right)!\left(v_M\left(S\right) - v_M\left(S \backslash i\right)\right)}{m!} \end{aligned}$$

In this last case, $\varphi_i\left(N, v_M\right)$ is the value obtained by the members of coalition $M \subset N$ when they share $v\left(M\right)$. Now, for each given coalition $M \subset N$, define a game (N, v_M^*), called the *incremental game*:

$$\forall S \subset N, v_M^*\left(S\right) = v(S) - v_M\left(S\right)$$

Note that, if $S \subset M$, $v_M^*\left(S\right) = 0$ and since the game is superadditive, $v^*\left(S\right) \geq v(S)$. The Shapley value for each game (N, v_M^*) can be defined as:

$$\begin{aligned} \forall i &\in N \backslash M, \\ \varphi_i\left(v_M^*\right) &= \frac{\sum_{S \subset N, i \in S}\left(s-1\right)!\left(m-s\right)!\left(v_M^*\left(S\right) - v_M^*\left(S \backslash i\right)\right)}{(n-m)!} \\ &= \frac{\sum_{S \subset M, i \in S}\left(s-1\right)!\left(m-s\right)!\left(v\left(S \cup N \backslash M\right) - v\left(S \backslash i \cup N \backslash M\right)\right)}{m!} \end{aligned}$$

This is the value obtained by players in $N \backslash M$, when they share what can be called, by extension, *their incremental value* $v(N) - v(M)$. In other words, this is the value which is calculated for each $N \backslash M$-member, assuming that a coalition M got its value $v(M)$ already and taking into account all possible orders of players in $N \backslash M$ when they join the coalition M.

Example 1 *An obvious case is that in which* M *is a singleton* $\{i\}$. *Then,* $\varphi_i\left(v_{\{i\}}\right) = v(\{i\})$ *and* $\varphi_i\left(v^*_{N\backslash i}\right) = v(N) - v(N \backslash \{i\})$ *is player* i*'s incremental value.*

Example 2 *Consider another case in which* $N = \{1, 2, 3, 4\}$ *and* $M = 12$. *Then,*

$$\varphi_1(v_{12}) = \frac{v(\{1\}) + v(12) - v(\{2\})}{2!}$$

and

$$\varphi_3(v^*_{34}) = \frac{v(N) - v(N \backslash 3) + v(123) - v(12)}{2!}$$

The following Proposition proved in a companion paper (Thoron 2006) shows that there is a general relationship between the Shapley values defined for a game v; for the games v_M, and for the games v^*_M. Consider C_m the class of coalitions $M \subset N$, of the same size m:

$$C_m = \{M \subset N : \#M = m\}$$

For each class C_m, the following relationship can be proved:

Proposition 1 *Let* N *any finite carrier of game* v. *For each given class* $C_m, m \leq n$, *the following relationship exists between the Shapley values defined for three categories of games: the original game* v, *the games* v_M *defined for each* $m-$*size sub-coalition* $M \subset N$, *and the associated incremental games* v^*_M*:*

$$\varphi_i(v) = \frac{m!(n-m)!}{n!}\left[\sum_{\substack{M \in C_m \\ i \in M}} \varphi_i(v_M) + \sum_{\substack{M \in C_m \\ i \notin M}} \varphi_i(v^*_M)\right]$$

3.2 Heuristic approach

Following the usual heuristic description of the Shapley value, players have to meet in a bargaining room to share the value of the grand coalition. They arrive sequentially and the order in which they do so is determined by chance, with all arrival orders equally probable. Each player, when she enters the room, demands and is promised the amount which her participation contributes to the value of the grand coalition.

Here, I modify this description by introducing two rooms: the waiting room and the bargaining room. The participants arrive sequentially and in a random order in the waiting room. But there is nothing to share before the coalition formed by the players in the waiting room has reached the threshold. When the last player necessary to reach this threshold enters in the waiting room, the door is closed from outside and the usual process is followed by the present players. Players arrive sequentially and in a random order in the bargaining room. Each player from the waiting room, when she enters the bargaining room, demands and is promised the amount which her participation contributes to the value of the coalition. For each coalition $S \subset N$, the value $\varphi_i(v_S)$ represents what player $i \in S$ can obtain in this negotiation.

When the waiting room is empty again, the door of the waiting room is reopened and the remaining players arrive sequentially and go straight to the bargaining room in order to demand the amount which their adherence contributes to the value of the grand coalition. For each coalition S, the value $\varphi_i(v_S^*)$ represents what player $i \in N \backslash S$ can obtain in this negotiation.

3.3 Characterization

Denote by $r(N)$ an order defined on the set of players. For a given order $r(N)$, each player i is associated with a ranking r_i. For a given leading coalition $L \subset N$ and a given order $r(N)$, consider the coalitions $S \subset N$, which satisfy the three following conditions:

(i) $L \subset S$

(ii) All the players who belong to coalition S arrive successively in $r(N)$: given $i \in S$, such that $r_i = Min_{k \in S} r_k$ and $j \in S$ such that $r_j = Max_{k \in S} r_k$, $\forall k \in S$, $r_j \geq r_k \geq r_i$. In other words, think about S as a block in the order $r(N)$.

(iii) One of the S-members arrives first in the order. In other words $r_i = Min_{k \in S} r_k = 1$.

For a given leading coalition $L \subset N$, to each order $r(N)$ corresponds a unique coalition S, which satisfies the three previous conditions and has the smallest number of members. But each coalition S may correspond to several orders. This defines an injective but not surjective application from the set of orders to the set of coalitions $\overline{L} = \{S : L \subset S\}$. Denote by α_S the number of orders associated with coalition S, this is the number of orders in which the S-members arrive in first position but among them the last to arrive is a L-member. In other words, in the order $r(N)$, the threshold is reached when the last L-member has arrived and all the players who have a smaller ranking constitute coalition S, associated with this order. Note that as a consequence of the injective application, we have: $\sum_{S \in \overline{L}} \alpha_S = n!$

Example 3 *For each coalition $S \in T(\tau)$, $\alpha_S = (n-s)!s!$. This is the number of orders in which the S-members arrive in the first positions.*

Example 4 *If $T(\tau) = C_s$, again $\alpha_S = (n-s)!s!$ but in that case, the coalitions in $\overline{T}(\tau)$ of larger size do not correspond to any order.*

Given this definition of weights $\alpha_S, S \in \overline{L}$, we can define the threshold value as follows:

Definition 3 *For any finite carrier N of game v and a leading coalition $L \subset N$, the T-value for each player $i \in N$ is given by:*

$$\phi_i(v, L) = \frac{\sum_{S \in \overline{L}, i \in S} \alpha_S \varphi_i(v_S) + \sum_{S \in \overline{L}, i \notin S} \alpha_S \varphi_i(v_S^*)}{n!}$$

In a companion paper (Thoron (2006)), I proved that the T-value is characterized by the following Axioms:

Axiom 5 *Efficiency and null player. If N is a carrier of game v, then $\sum_{i \in N} \phi_i(v, L) = v(N)$*

Axiom 6 *Additivity. If w is defined such that $w(S) = v(S) + u(S)$ then:*

$$\phi_i(w, L) = \phi_i(v, L) + \phi_i(u, L)$$

Axiom 7 *Anonymity. If π is a permutation of U and πv defined such that $\pi v(\pi S) = v(S)$, then,*

$$\phi_{\pi i}(\pi v, \pi L) = \phi_i(v, L)$$

Axiom 8 *Transfer. For any given a unanimity game αU_K. For any given $L \subseteq U$:*

$$\text{If } K \subseteq L, \phi_i(\alpha U_K, L) = \begin{cases} \frac{\alpha}{|K|}, \text{if } i \in K; \\ 0 \text{ otherwise} \end{cases}$$

$$\text{If } L \subseteq K, \begin{cases} \phi_i(\alpha U_K, L) = 0, \forall i \notin K, \\ \phi_j(\alpha U_K, L) = \phi_i(\alpha U_K, L) + \phi_j(\alpha U_K, K), \text{ if } i \in L \text{ and } j \in K \backslash L \end{cases}$$

This last Axiom is equivalent to the following Axiom:

Axiom 9 *For any given $K \subseteq U$, and $\alpha \in \mathbb{R}$,*

$$\text{If } K \subseteq L, \phi_i(\alpha U_K, L) = \begin{cases} \frac{\alpha}{|K|}, \text{if } i \in K; \\ 0 \text{ otherwise} \end{cases}$$

$$\text{If } L \subseteq K, \phi_i(\alpha U_K, L) = \begin{cases} \frac{\alpha}{|K|} \frac{|L|}{|K|}, \text{if } i \in L; \\ \frac{\alpha}{|K|} \left(1 + \frac{|L|}{|K|}\right), \text{if } i \in K \backslash L; \\ 0 \text{ otherwise} \end{cases}$$

What is the meaning of these two last axioms? Note that, for a given unanimity game αU_K, $\phi_i(\alpha U_K, K) = \varphi_i(\alpha U_K)$. When the coalition which constitutes the threshold is $L = K$, the T-value coincides with the Shapley value. Given this, the Transfer Axiom shows how a smaller threshold modifies the Shapley value. In a unanimity game, all the players are identical and they are all necessary to realize a worth. In that case, when there is no threshold to bias the

sharing, the value is expected to be the equal sharing of $v(K)$ between the K-members (see Axiom 4 of the Shapley value). Then, when a smaller threshold is introduced, it may attribute different positions to previously symmetric players. In that case, the player who is *above* the threshold has a higher value than the player who is *below* the threshold. Furthermore, the difference between the two values is just the value without threshold, that is the Shapley value. Therefore $\phi_i(\alpha U_K, L)$, which is the share received by the members of the threshold coalition L is also what they transfer to the non-member players. For example, if $2 * |L| = |K|$, the members of the threshold coalition transfer half of their share to the non members.

A natural idea would be that the players who are below the threshold have a kind of veto power since the sharing cannot be done without them. This is misleading. In a unanimity game αU_K the K-members all have this veto power. The worth α cannot be realized without each of the K-members. In fact, the threshold introduces an additional distinction. Remember that, because the game is superadditive, for any coalition S it is always better to get its incremental worth rather than its worth: $v(K) - v(K \backslash S) \geq v(S)$. Therefore, even if all players in K participate in the sharing of $v(K)$, there is a difference among them depending on their position with respect to the threshold. The players who are below the threshold will not have the possibility to share an incremental value $v(K) - v(K \backslash S)$ instead of a value $v(S)$. One can say that the players who are below the threshold are more committed to the sharing. They will initiate the sharing. The last axiom measure the price they pay for this.

4 Non-cooperative models of agreement formation

4.1 Incentives

The purpose of the previous Section was to explain how a country can signal its leadership commitment through the minimum participation constraint. The question remains as to why a country should behave like this. Consider the situation in which countries have the possibility to participate or not in an environmental agreement. Consider that they anticipate that the expected outcome of the negotiation will be the Shapley value. They also anticipate that the negotiation will be followed by a ratification phase. In this framework, a sub-set of countries may have the right incentives to sign and ratify the agreement. They simply benefit more from participating in the agreement than from the externalities from which they could benefit by staying independent. We say in this case that the coalition of countries is stable in the sense defined by D'Aspremont et al. (1983). This is also equivalent to say that there exists a Nash equilibrium of a simple coalition formation game in which the set of strategies is just binary (see Thoron (1998)). As it is well known from the literature on the theory of stable cartels, the size of this coalition depends, of course, on the payoff function,

but it is often rather limited. In other words, the stable agreement is "small". However, as Carraro and Siniscalco (1993) have shown, an expansion of this coalition can be induced through transfer and commitment. Transfer without commitment is not enough since, if the transfer is efficient and provokes the coalition's expansion, then the countries which belonged to the initial stable coalition may now have an incentive to leave and free ride to benefit from the larger externalities and to avoid bearing a higher cost. Therefore, Carraro and Siniscalco propose the following mechanism: The countries belonging to the initial stable coalition commit to cooperate. Once this is done, they choose the transfer in order to maximize the number of signatories:

1) transfers are self-financed, i.e. the total transfer must be lower than the gain that the committed countries obtain from expanding the coalition.

2) the move to a larger coalition must be Pareto-improving, i.e. all countries must be better off than in the situation preceding the coalition expansion, and better off than in the case of non-cooperation. (cf. Carraro and Siniscalco (1993) p 316).

The purpose of this Section is to give a formal framework for a similar mechanism. However here, it is assumed that the transfers are not direct monetary transfers but endogenously emerge from the negotiation as a modification of the sharing rule. Indeed, this is the simplest way to organize these transfers. In other words, the leading coalition is more "flexible" in the negotiation. I assume that the T-value is the new sharing rule when a leading coalition constitutes a threshold. It represents this leading coalition's willingness to organize a transfer in order to make an expansion possible. Given this new sharing rule, the question is to know which leading coalition will form and which expansion will be possible.

There are two degrees of commitment: the strongest degree of commitment is to send the signal that the country will be a leading country. The second degree of commitment is, for a country, to say that it will participate but needs to be subsidized. Some countries are willing to be strongly committed if and only if other countries participate. They only sign the agreement and assume the high commitment if the others sign also. The last possibility is just to refuse to participate.

4.2 Coalitional function

First, I will consider a simple and aggregated game called game G which will be used in the subsequent Sections as a benchmark. Consider a set of n countries $N = \{1, ..., n\}$, each one characterized by an amount of pollution $y_i, i \in N$, in a status quo situation. The implementation of an agreement signed by a subset of countries T means that they choose to reduce their pollution by a percentage $x \in [0, 1]$ which maximizes their joint net benefit. The net benefit is the increase of utility produced by the improvement of the environment less the cost of decreasing pollution. Denote the joint net benefit of coalition T by $P_T\left(x\left(y_i\right)_{i\in T}\right)$ in which x is a scalar and $Y_T = \left(y_i\right)_{i\in T}$ the vector of the different

T-members' pollution. Consider:

$$\widehat{x}_T = Arg \max_x P_T \left(x \left(y_i \right)_{i \in T} \right), \forall T \subset N$$

It is assumed that $\widehat{x}_i = 0, \forall i \in N$. The following assumption A1 holds:

$$A1 : \widehat{x}_T \leq \widehat{x}_S \text{ if } S \subset T, |S| > 1$$

Example 5 *The following payoff function is often used in the litterature on environmental agreements (cf. for example Barret (1994), Ray and Vohra (2001), Courtois and Haeringer (2005)...). Consider the symmetric case, $y_i = y = 1u.$, $\forall i \in N$. A coalition of s countries sign an agreement to decrease their pollution by x per cent. The individual benefit of the aggregate abatment level $xsy = xs$ is:*

$$\begin{aligned} Q \left(x \left(y_i \right)_{i \in S} \right) &= Q \left(xsy \right) \\ &= axsy - \frac{1}{2} x^2 s^2 y^2 = asx - \frac{1}{2} x^2 s^2 \end{aligned}$$

The symmetrical cost of decreasing pollution is, for each one of the s countries:

$$C \left(xy \right) = C \left(x \right) = \frac{c}{2} x^2$$

Where c is a positive parameter smaller than 4, $c \leq 4$. The percentage x is chosen to maximize the aggregate payoff of the s countries:

$$\begin{aligned} P_S \left(x_S \left(y_i \right)_{i \in N} \right) &= P_S \left(xsy \right) \\ &= \left(asx - \frac{1}{2} x^2 s^2 - \frac{c}{2} x^2 \right) s \end{aligned}$$

The solution is $\widehat{x}_S = \frac{as}{s^2 + c}$. It can be easily checked that, when $c \leq 4$, that is when the marginal cost is not "too high", $\frac{\partial \widehat{x}_S}{\partial s} \leq 0$ as soon as $s \geq 2$ and Assuption A1 is verified.

Note that, A1 is consistent with another more traditional assumption which is that the pollution abatment increases when the coalition expands. This will be the property of superadditivity of the characteristic function V defined as follows:

$$v \left(S \right) = \widehat{x}_S \sum_{i \in S} y_i$$

The worth $v(S)$ is the pollution abatment a coalition S would choose if its members signed a binding agreement. Note that assumption $A1$ implies that characteristic function v is *concave:* the incremental value of each country is decreasing. This worth could have been defined as a partition function, depending on a coalition structure. However, it is assumed that only one coalition can be formed. Implicitly, countries in $N \backslash S$ remain independent and, as specified above, do not decide any abatment. I will come back to this point below. The

characteristic function V is assumed to be superadditive. Therefore, in particular:

$$A2 : \forall S \subset N, \forall T \subset S, v(S) \geq v(S \backslash T)$$

In a given coalition T, members share the joint benefits P_T using a sharing rule $\psi(v)$ to decide about each country's reduction. Therefore, $\psi_T(v) = (\psi_i(v))_{i \in T}$ is a vector in which $\psi_i(v)$ is the percentage applied to country i. The sharing rule is assumed to be *efficient*:

$$A3 : \psi_T(v) Y_T = \sum_{i \in T} \psi_i(v) y_i = \widehat{x}_T \sum_{i \in T} y_i = v(T)$$

As a consequence of superadditivity of v, we can write:

$$v(T) = \widehat{x}_T \sum_{i \in T} y_i \geq \widehat{x}_{T \backslash j} \sum_{i \in T \backslash j} y_i = v(T \backslash j)$$

But, by Assumption A1:

$$\widehat{x}_{T \backslash j} \sum_{i \in T \backslash j} y_i \geq \widehat{x}_T \sum_{i \in T \backslash j} y_i$$

Therefore:

$$\widehat{x}_T \sum_{i \in T} y_i \geq \widehat{x}_{T \backslash j} \sum_{i \in T \backslash j} y_i \geq \widehat{x}_T \sum_{i \in T \backslash j} y_i$$

When an additional country joins the coalition, the coalitional function is such that the total reduction of pollution increases but the reduction by the initial countries decreases. In what follows, we will see that, as a consequence, the initial members benefit from the new membership through two different ways.

The vector of status quo $(y_i)_{i \in N}$ is fixed. Then, the different countries' payoffs are denoted by ${\pi_i}^\circ, \forall i \in N$. From this situation, if one agreement is signed by a coalition T, the T-members choose a pollution abatment $v(T)$ and share their efforts to reach this target using as a sharing rule the Shapley value: $\varphi(v)$. Each T-member is then characterized by the benefit function drawn from the partnership:

$$\forall i \in T : P_i(T, \varphi_T(v))$$

Note that, in general, the maximization problem depends on what the countries which have not signed the agreement do. This could only been represented by a partition function. However here, this is not necessary since there is only one agreement. The characteristic function is equivalent to a partition function in which the countries in $N \backslash T$ are independent. The payoff of these countries need to be defined nevertheless. The pollution reduction is an externality they can benefit from but of course they do not bear the same cost since they do not implement the agreement. If country i does not participate in the agreement, its benefit from the implementation of the agreement by the T-members is denoted by $Q_i(T, \varphi_T(v))$:

$$\forall i \in N \backslash T : Q_i(T, \varphi_T(v))$$

Note that, this means that there is no strategic adaptation of pollution reduction by the non signatories (no leakage). The interpretation is the following: countries which remain independent do not take into account the quality of the environment. This is why they do not decide any reduction. On the other hand, when other countries decrease pollution, they do not increase or decrease pollution as a best response. Simply, they do not *participate in* the abatment game. However, this assumption is not determinant in what follows. It just makes the model easier to write and to understand and do not affect the results. Indeed, we could write explicitly the abatment game between the signatories and the non-signatories without changing the results. In fact here, a non-signatorie' s payoff is just a function of the pollution abatment induced by the signatories. Therefore, it can also be written $Q_i\left(S,\varphi_S\left(v\right)\right)=Q_i\left(\varphi_S\left(v\right)Y_S\right)$. The following assumptions hold:

The game is essential: $\exists S\subset N:P_i\left(S,\varphi_S\left(v\right)\right)\geq{\pi_i}^{\circ},\forall i\in S$.

A4 Q_i is an increasing function of the abatment:

$$\text{If } S\subset T,\forall i\in N\backslash T,Q_i\left(T,\varphi_T\left(v\right)\right)\geq Q_i\left(S,\varphi_S\left(v\right)\right)$$

A5 For a given agreement T, the non signatories which do not bear the cost of pollution reduction but can benefit from it, have a higher payoff than the signatories:

$$\text{If } i \text{ and } j \text{ are identical but } i\notin T \text{ and } j\in T, \text{ then } Q_i\left(T,\varphi_T\left(v\right)\right)\geq P_i\left(T,\varphi_T\left(v\right)\right)$$

4.3 Stable agreements

Consider the following agreement formation game called game G. Countries in N have to decide to sign or not to sign an agreement to reduce pollution. For each country $i\in N$, its strategy is then denoted by: $\sigma_i\in\{0,1\}$. Each strategy profile σ generates a coalition $K=\{i:\sigma_i=1\}$ of countries which sign and implement this agreement to reduce pollution. Payoffs are then determined as in the previous section. A Nash equilibrium of this simple game σ^* generates a φ-stable agreement signed by the members of coalition $K^*=\{i:\sigma_i^*=1\}$. This implies that coalition K^* satisfies the two well known conditions of internal and external stability proposed by d'Aspremont et al. (1983) and which must be written in this paper's framework:

Coalition T^* is internally stable if $\forall i\in K^*$:

$$P_i\left(K^*,\varphi_{K^*}\left(v\right)\right)\geq Q_i\left(K^*\backslash i,\varphi_{K^*\backslash i}\left(v\right)\right)$$

Coalition T^* is externally stable if $\forall i\in N\backslash K^*$:

$$Q_i\left(K^*,\varphi_{K^*}\left(v\right)\right)\geq P_i\left(K^*\cup i,\varphi_{K^*\cup i}\left(v\right)\right)$$

When the different countries are identical, that is when $\forall i\in N$, $y_i=y$, Q and P only depend on the number of countries which sign the agreement.

Figure 2:

Then, the Shapley is the equal sharing and $Q_i\left(K^*, \varphi_{K^*}\left(v\right)\right) = Q\left(k^*\right)$ and $P_i\left(K^*, \varphi_{K^*}\left(v\right)\right) = P\left(k^*\right)$ (see Figure 2).

Let us denote by $\mathbb{N}^{\psi}$ the set of coalitions generated by a Nash equilibrium of game G when the sharing rule is ψ. Then, the members of each one of these coalitions will not have any difficulty to sign the agreement. When the game is symmetric, Thoron (1998) shows that this equilibrium always exists. Furthermore, using the concept of Coalition Proof Nash Equilibrium (CPNE), it is shown that this equilibrium is unique. The concept of CPNE is a refinement of Nash equilibria proposed by Bernheim, Peleg and Wilson (1987), which takes into account coalition deviations and satifies a property of consistency.

Theorem 1 *Game G has one and only one coalition-proof stable cartel, which is the greatest stable cartel.*

4.4 Stable agreements with ratification phase

The previous game G is one shot: the countries decide symultaneously to sign or not and the agreement is then implemented. Now assume that this one shot game, which will be called the negotiation phase, is followed by a ratification phase. In other words, the signatories have to confirm their participation. This two-step game will be denoted by G^r. Assume that each country is then represented by a team of delegates during the first negotiation phase and by domestic

institutions (now on DI) during the ratification phase. The outcome of the first phase is a coalition T. The second phase of ratification is played by the domestic institutions of the different signatory countries $i \in T$. Therefore, each coalition outcome of the first phase $T \subset N$ defines a sub-game G_T^r in which the set of players is T. Then, at each period $t \in [0, \infty[$, one country's DI $i \in T$ is drawn randomly which has to decide to ratify or not: $\sigma_i = 0$ or 1. Each country's DI try to maximize their country's payoff. Each period t is characterized by a history h_t which is just the coalition of countries which have ratified in the previous periods. The ratification process is said to be *over* at period t if:

1- either $h_t = T$, and *the ratification process is* said to be over and *completed.* Payoffs are then:

$$P_i\left(T, \varphi_T\left(v\right)\right), \forall i \in T$$

$$Q_i\left(T, \varphi_T\left(v\right)\right), \forall i \in N \backslash T$$

2- or h_t is such that no country in $T \backslash h_t$ wants to ratify. Payoffs are then:

$$P_i\left(T, \varphi_{h_t}\left(v\right)\right), \forall i \in h_t$$

$$Q_i\left(T, \varphi_{h_t}\left(v\right)\right), \forall i \in N \backslash h_t$$

In this second case countries in h_t have to implement the pollution abatment they had been assigned, although the result will not be what expected in the agreement since some signatories will not participate in the implementation. However, the countries are assumed to be farsighted and can anticipate the ratification progress. But the outcome of the ratification process depends on the order in which the countries in T are drawn. Therefore, even if the countries are farsighted, generally they cannot know with certainty what will be their payoff. However, the following result shows that it is nevertheless the case, in specific situations which are of particular interest. Countries cannot anticipate the order of ratification, which is random, but they know the rule which governs the ratification phase and determines when the agreement has to be implemented. I will consider three cases, given a coalition T of signatories:

Rule R^O: the agreement is implemented when the ratification process is over.

Rule R^U: the agreement is implemented if and only if all signatories have ratified.

Rule R^L: the agreement is completed if and only if, when the ratification process is over, coalition $L \subset T$ of signatories have ratified.

The ratification process is said to be *always completed* if it is completed in each order.

Proposition 2 *If T^* is φ-stable, the ratification process is always completed, whatever the rule of the ratification process is.*

Proof. Assume that T^* is φ-stable, then:

$$P_i\left(T^*, \varphi_{T^*}\left(v\right)\right) \geq Q_i\left(T^* \backslash i, \varphi_{T^* \backslash i}\left(v\right)\right) \tag{1}$$

The external stability condition is useless here. The ratification process must be solved by backward induction.

Rule R^O: The agreement is implemented when the ratification process is over. Assume that $h_t = T \backslash i$. Then, country i ratifies if:

$$P_i\left(T^*, \varphi_{T^*}\left(v\right)\right) \geq Q_i\left(T^*, \varphi_{T^* \backslash i}\left(v\right)\right)$$

By definition and by Assumptions 1, the Shapley value, as an average of incremental worths, is larger than the smallest incremental worth:

$$\varphi_i\left(v_{T^*}\right) y_i \geq \widehat{x}_{T^*}\left(\sum\nolimits_{i \in T^*} y_i\right) - \widehat{x}_{T^* \backslash i}\left(\sum\nolimits_{j \in T^* \backslash i} y_j\right)$$

and:

$$\widehat{x}_{T^* \backslash i}\left(\sum\nolimits_{j \in T^* \backslash i} y_j\right) \geq \widehat{x}_{T^*}\left(\sum\nolimits_{i \in T^*} y_i\right) - \varphi_i\left(v_{T^*}\right) y_i$$

This implies, by efficiency of the Shapley value:

$$\varphi_{T^* \backslash i}\left(v_{T^* \backslash i}\right)\left(\sum\nolimits_{j \in T^* \backslash i} y_j\right) \geq \varphi_{T^* \backslash i}\left(v_{T^*}\right)\left(\sum\nolimits_{j \in T^* \backslash i} y_j\right)$$

Finally, by Assumption A4:

$$Q_i\left(T^* \backslash i, \varphi_{T^* \backslash i}\left(v\right)\right) \geq Q_i\left(T^*, \varphi_{T^* \backslash i}\left(v\right)\right)$$

and since T^* is φ-stable and verifies (1), country i ratifies. If $h_t = T^* \backslash i, j$ country j can anticipate that if it ratifies, country i will also ratify. If it does not ratify, the best situation for it would be that country i ratifies. However, even if it is the case, for the same reasons as before, it is better for it to ratify. Going backward, we check that for each history h_t, each country $i \in T^* \backslash h_t$ has always an incentive to ratify.

Rule R^U: If $h_t = T \backslash i$ country ratifies if the coalition of signatories is profitable, that is if:

$$P_i\left(T^*, \varphi_{T^*}\left(v\right)\right) \geq {\pi_i}^\circ$$

which is always satisfied since T^* is stable and $Q_i\left(T^* \backslash i, \varphi_{T \backslash i}\left(v\right)\right) \geq {\pi_i}^\circ$.

Rule R^L: If, when $h_t = T \backslash i$, the last country $i \in L$, the conditions are identical to R^U. Otherwise, $i \notin L$ and conditions are identical to R^O. In both cases the ratification process is always completed. ■

The previous proposition says that *it is better to refuse to sign than to sign and not ratify*. Therefore, there is no strategic exploitation of the ratification process. This is a consequence of Assumption 1: the pollution reduction implemented by the other signatories and as a consequence the externality, is bigger when it is the optimal reduction for them than when they expect an additional reduction. Note that this proposition is a necessary but not sufficient condition. Indeed, under rule R^L, non stable coalitions can be followed by always completed ratification process. Consider the following definition:

Definition 4 *Coalition T is internally φ-stable with respect to sub-coalition $S \subset T$ if $\forall i \in S$:*

$$P_i\left(T, \varphi_T\left(v\right)\right) \geq Q_i\left(T \backslash i, \varphi_{T \backslash i}\left(v\right)\right)$$

Under rule R^L, any coalition $M = L \cup K$ which is not necessarily ψ-stable but is internally ψ-stable with respect to sub-coalition K would be ratified in every order for the same reasons explained in the proof of the previous proposition. An important consequence of Proposition 2 is that the ratification phase of game G^r does not change the equilibria of game G.

Proposition 3 *Whatever the rule of the ratification process is, the set of agreements generated at the equilibria of game G coincides with the set of agreements generated at the equilibria of game G^r.*

Proof. Assume that T^* is φ-stable in game G, then whatever the rule of the ratification process is, it will be always completed and payoffs will be:

$$P_i\left(T^*, \varphi_{T^*}\left(v\right)\right), \forall i \in T^* \text{ and } Q_i\left(T^*, \varphi_{T^*}\left(v\right)\right), \forall i \in N \backslash T^*$$

If one country $i \in T^*$ considers deviating, it does not know anymore if the ratification process will be completed. However, it knows that its payoff will be at most $Q_i\left(\varphi_{T \backslash i}\left(T \backslash i, v\right)\right)$. Therefore, it does not have any incentive to deviate if:

$$P_i\left(T^*, \varphi_{T^*}\left(v\right)\right) \geq Q_i\left(T^* \backslash i, \varphi_{T^* \backslash i}\left(v\right)\right)$$

which is verified since T^* is stable in game G. Then T^* can be generated by an equilibrium in game G^r.

If T is not φ-stable in game G, it cannot be generated by an equilibrium in game G^r. Payoffs when the ratification process is completed are the highest payoffs a country can anticipate when it considers a deviation. If this is not enough because the coalition is not φ-stable, the ratification process cannot improve the incentive to sign. Indeed, it is not possible that more countries ratify in coalition $T \backslash i$ than in coalition T. ■

Therefore, the introduction of a ratification phase does not have any impact on the equilibria of the game. Indeed, under Assumption A1, the minimum participation rule which governs this ratification phase has no impact on the outcome. The ratification phase is always completed. What happens when Assumption A1 is not verified? If the coalitional function V is not concave but convex, indeed countries have an incentive to sign in order to increase their partners' contributions and then to refuse to ratify in order to free ride. In this case the only rule which can guarantee completeness of the ratification process is the unanimity rule R^U. However if this was the case, the outcome would be the grand coalition. It is much more plausible that, if the coalitional function is convex when the number of participants is not too big, it becomes concave at one point. When the coalitional fonction has a S-shape of this sort, a minimum

participation constraint can indeed eliminate incentives to free ride when the benefits of the partnership are below the level at which they become concave.

In what follows, consequences of the minimum participation rule on the previous phase of negotiation are analysed. A new mechanism is introduced in which, during the negotiation, the countries may decide to have different degrees of commitment. Again, in the second ratification phase, they have to sequentially confirm their participation by ratifying the agreement.

4.5 Endogenous leadership

The following game will be called game G^l. The negotiation phase of game G^l involves two steps. In a first step, the different countries have the possibility to form a coalition which will be the *leading coalition* in the second negotiation step. Being a leading coalition has two implications. First, in the second step, this coalition L proposes an expansion to another coalition K using the T-value as a sharing rule: $\phi(L \cup K, v, L)$. Second, the rule for the ratification phase which will follow is that the agreement will only be implemented if the L-members have ratified.

4.5.1 Negotiation phase

The players are the n countries, $N = \{1, ..., n\}$ (their delegates). In the first step, each country i has to choose a leading coalition to which it wants to belong, its set of strategies is:

$$\sum\nolimits_i^1 = \{L : L \subset N \cup \varnothing; i \notin L \text{ if and only if } L = \varnothing\}$$

Given a strategy profile σ^1, a coalition L is feasible if $\forall i \in L, \sigma_i^1 = L$. The leading coalition generated by the strategy profile σ^1 is then the feasible coalition with the largest worth $V(L)$. I changed the strategy sets in this first step. The idea is that the formation of a leading coalition is not an open membership game but an exclusive membership game. The decision to become a leading coalition, which is willing to over-contribute is clearly a collective action. Then, in the second step, the remaining countries in $N \backslash L$ decide to participate or not in the agreement. Strategies are then defined like in the previous simple game G. For each country $i \in N \backslash L$, its strategy is then denoted by: $\sigma_i^L \in \{0, 1\}$. To summarize, for each country $i \in N$ and for the whole negotiation phase, its strategy is denoted by $\left(\sigma_i^1; \sigma_i^2\right)$, in which $\sigma_i^1 \in \sum_i^1$ and $\sigma_i^2 = \left(\sigma_i^L\right)_{L \subset N, i \notin L}, \forall L \subset N, i \notin L, \sigma_i^L \in \{0,1\}$. The outcome of the negotiation phase is a pair (L, K_L) :

$$\begin{aligned} L &= \arg\max \left\{v(L) : \forall i \in L, \sigma_i^1 = L\right\} \\ K_L &= \left\{i \in N : \sigma_i^L = 1\right\} \end{aligned}$$

4.5.2 Ratification phase

The ratification phase is described as in game G^r. Given that the outcome of the first phase is a pair (L, K_L) with $M = L \cup K_L$, the rule which governs

the ratification phase is R^L. Payoffs are determined by the threshold value, given that the leading coalition L constitutes the threshold. Therefore, if the ratification phase is always completed, payoffs are now:

$$P_i\left(M, \phi_M\left(v, L\right)\right), \forall i \in M$$

and

$$Q_i\left(M, \phi_M\left(v, L\right)\right), \forall i \notin M$$

From Proposition and the following discussion, we know that, given a pair $\left(L, K_L\right)$, if $M = L \cup K_L$ is ϕ-stable, the ratification process is always completed, whatever the rule of the ratification process is. Furthermore, under rule R^L a sufficient condition is that M is internally ϕ-stable for sub-coalition K_L.

4.5.3 Equilibria

In this subsection, the countries will be considered to be symmetrical: $y_i = y$, $\forall i \in N$. In order to find the sub-game perfect Nash equilibria, game G^l must be solved by backward induction. The previous result tells us that, if $M = L \cup K_L$ is internally ϕ-stable for sub-coalition K_L, then the ratification process by countries in M is always completed under rule R^L. Then, if the ratification phase is always completed, and given that the game is now symmetric, payoffs will be denoted by:

$$P_{K_L}\left(l + k, \phi\left(v, l\right)\right), \forall i \in K_L$$

$$P_L\left(l + k, \phi\left(v, l\right)\right), \forall i \in L$$

and

$$Q\left(l + k, \phi\left(v, l\right)\right), \forall i \notin M$$

Note that assumptions A1 to A5 are still verified. Compared with $P\left(K, \varphi_K\left(v\right)\right)$ the sharing rule $\phi\left(v_{L \cup K}, L\right)$ shifts $P_K\left(L \cup K, \phi\left(v, L\right)\right)$ up and $P_L\left(L \cup K, \phi\left(v, L\right)\right)$ down (cf. Figure 2). The curve $Q\left(K, \varphi\left(v\right)\right)$ is unchanged since it only depends on the total abatment and because sharing rule ϕ is also efficient, the total abatment is unchanged: $\phi_M\left(v_M, L\right) Y_M = \varphi_M\left(v_M\right) Y_M = v\left(M\right)$. Therefore, $Q\left(M, \phi_M\left(v, L\right)\right) = Q\left(M, \varphi_M\left(v\right)\right)$, $\forall i \notin M$.

For each leading coalition $L \subset N$, outcome of the first step of the negotiation phase, we can determine a sub-game: G_L^l. In each sub-game G_L^l, there is a coalition-proof Nash equilibrium of this sub-game and a unique coalition-proof stable agreement, denoted by K_L^*. Here, the only difference is that the stable coalition can be the empty set. Therefore, for each sub-game G_L^l, the equilibrium is a coalition $M = L \cup K_L^*$ stable for the sub-coalition K_L^*.

Consider the following strategy profile $\left(\sigma^1, \sigma^2\right)$:

$$\begin{aligned} \forall i &\in N, \\ \sigma_i^1 &= \begin{cases} L, \forall i \in L \\ \varnothing, \forall i \notin L \end{cases} \end{aligned}$$

Figure 3:

$$\sigma_i^2 = \left(\sigma_i^{L'}\right)_{L' \subset N, i \notin L'} \text{ such that } \sigma_i^{L'} = \begin{cases} 1, \forall i \in L' \cup K_{L'}^* \\ 0, \forall i \notin L' \cup K_{L'}^* \end{cases}$$

Now, consider the first step of the negotiation phase. If no leading coalition is formed, the outcome will be, in sub-game $G_{\varnothing}^{l}$, the stable coalition L^*. Therefore, countries whose strategy is to sign in the following step, may have an incentive to become leaders in the first step if they can, doing so, increase partnership and, as a consequence, their payoffs. The following Proposition explains when this expansion is possible. Depending on the characterization of the payoff functions, there are three types of equilibria in game G^l.

Proposition 4 *Let L^* be the agreement formed at the equilibrium of game G^r. At the equilibrium of game G^l:*

1- No expansion is possible and the agreement formed is L^* when:
$K_{L^*}^* = \varnothing$, *or* $K_{L^*}^* \neq \varnothing$, *and* $P_{L^*}\left(L^* \cup K_{L^*}^*, \phi\left(v, L^*\right)\right) < P_{L^*}\left(L^*, \varphi\left(v\right)\right)$.

2 - There is an expansion from the leading coalition L^* when:
$K_{L^*}^* \neq \varnothing$, *and* $P_{L^*}\left(L^* \cup K_{L^*}^*, \phi\left(v, L^*\right)\right) \geq P_{L^*}\left(L^*, \varphi\left(v\right)\right)$.

3 - The grand coalition is formed when:
$Q\left(L^*, \varphi\left(v\right)\right) < P\left(N, \varphi\left(v\right)\right)$.

Proof. Consider that $L^*$is the agreement generated by the CPNE in game G (under sharing rule $\varphi\left(v\right)$). If $K_{L^*}^* = \varnothing$ no expansion is possible from L^*, which cannot be a leading coalition. Furthermore, the following inequalities are

verified:

$$Q\left(L^{*},\varphi\left(v\right)\right)>P_{K}\left(L^{*}\cup K,\phi\left(v,L^{*}\right)\right),\forall K\subset N\backslash L^{*} \tag{2}$$

Can another coalition be a leading coalition in this case? First, consider a leading coalition $L, L\subset L^{*}$. Since no expansion is possible from L^{*}, even if an expansion is possible from L, $M=L\cup K_{L}^{*}$ cannot be larger than L^{*}. If a leading member deviates from L, L^{*} will be formed and the deviator will get at least $P\left(L^{*},\varphi\left(v\right)\right)$, which is larger than $P\left(M,\phi\left(v,L\right)\right)$. Therefore, $L, L\subset L^{*}$ cannot be a leading coalition at the equilibrium. Now, consider a leading coalition $L, L^{*}\subset L$. Assume that an expansion K_{L}^{*} is possible from L. Given $M=L\cup K_{L}^{*}$:

$$P_{L^{*}}\left(M,\phi\left(v,L^{*}\right)\right)<P_{L}\left(M,\phi\left(v,L\right)\right)$$

However, there is always a $K\subset N\backslash L^{*}$ such that:

$$P_{L^{*}}\left(L^{*}\cup K,\phi\left(v,L^{*}\right)\right)\geq P_{L}\left(L\cup K_{L}^{*},\phi\left(v,L\right)\right)$$

Therefore, from inequalities (1) and given that:

$$P_{K}\left(L^{*}\cup K,\phi\left(v,L^{*}\right)\right)>P_{L^{*}}\left(L^{*}\cup K,\phi\left(v,L^{*}\right)\right)$$

coalition L cannot be a leading coalition since each member has an incentive to deviate:

$$Q\left(L^{*},\varphi\left(v\right)\right)>P\left(L\cup K_{L}^{*},\phi\left(v,L\right)\right)$$

If $K_{L^{*}}^{*}\neq\varnothing$ and L^{*} is a leading coalition. Then, if one leading country deviates, choosing $\sigma_{i}^{1\prime}=\varnothing$, the leading coalition is dismanteled. In the subgame $G_{\varnothing}^{l}$ the deviator becomes a member when L^{*} is formed. Therefore, a condition for strategy profile $\left(\sigma^{1},\sigma^{2}\right)$ to be an equilibrium is, in this case, that the expansion of the stable agreement is beneficial for the leading countries:

$$P_{L^{*}}\left(L^{*}\cup K_{L^{*}}^{*},\phi\left(v,L^{*}\right)\right)\geq P_{L^{*}}\left(L^{*},\varphi\left(v\right)\right)$$

If $K_{L^{*}}^{*}$ satisfies this condition, the pair $(L^{*},K_{L^{*}}^{*})$ is generated by the CPNE of the game G^{l}. If $K_{L^{*}}^{*}$ does not satisfy this condition, L^{*} cannot be a leading coalition. There is no leading coalition and the agreement is signed by the stable coalition L^{*} formed at the second step.

Is there an equilibrium in which the leading coalition L is not initially stable in game G (under sharing rule φ)? There are two cases to consider.

First case: $Q\left(L^{*},\varphi\left(v\right)\right)<P\left(N,\varphi\left(v\right)\right)$. Then $\sigma_{i}^{1}=N,\forall i\in N$ is an equilibrium of the first step and $L=N$. Indeed, any deviation by a country eliminates the leading coalition and the stable agreement L^{*} forms. The deviating country earns a payoff $Q\left(L^{*},\varphi\left(v\right)\right)$. Therefore, in this case, because the countries have the possibility to form a leading coalition they can form the grand coalition, even if the grand coalition is not stable under sharing rule φ: $Q\left(L^{*},\varphi\left(v\right)\right)<P\left(N,\varphi\left(v\right)\right)<Q\left(N\backslash i,\varphi\left(v\right)\right)$.

Second case: $Q\left(L^{*},\varphi\left(v\right)\right)\geq P\left(\varphi\left(N,v\right)\right)$. Consider a coalition $L, L^{*}\subset L$. Note that a L-member's payoff is smaller than the payoff in the grand coalition, since by definition and by assumption:

$$P_{L}\left(M,\phi\left(v,L\right)\right)<P\left(M,\varphi\left(v\right)\right)<P\left(N,\varphi\left(v\right)\right),\forall L\subset N.$$

Therefore, if a member of the leading coalition deviates, choosing $\sigma_i^1 = \varnothing$, in the following sub-game $G_{\varnothing}^{l}$, L^* is formed and the deviator can be independent with a higher payoff $Q\left(L^*, \varphi\left(v\right)\right)$. ■

The last result 3 about the formation of the grand coalition is just a consequence of the sequentiality of the negotiation phase. The stable agreement at the second step of the negotiation phase plays the role of a credible threat during the first step in which the leading coalition is formed.

5 Conclusion

The purpose of this paper has been to analyse the role played by the minimum participation rule which governs the ratification phase of international treaties. Whilst the literature has emphasised the consequences of this rule for the process of ratification, I showed that it must have important consequences for the outcome of the negotiation. Indeed, this paper sustains the argument that negotiators do not anticipate a strategic exploitation of the ratification phase. It is shown that, under a mild assumption which can be understood as a concavity of the benefits drawn from partnership, the negotiating delegates anticipate that the ratification phase will always be completed. However, I showed that the minimum participation rule can be used to modify the result of the negotiation. More precisely, when the rule is defined as a list of countries, these countries can use this rule to signal their leadership commitment, and by so doing, they can induce an expansion of the agreement. The price of this leadership commitment, reflected in the T-value proposed in this paper, is that the leading countries accept a distortion of the sharing rule at their expense, in order to favor the adherence of additional partners.

The rule considered in this paper is a list of countries which must have ratified before the treaty can enter into force. This rule has been used in real international treaties. However other rules may be encountered. The most common rule, which is just a number of countries, clearly cannot introduce any bias into the negotiation. In this case, we may assume indeed, that the role of this rule is to introduce a threshold. The latter will eliminate incentives to avoid ratifying when the benefits of the partnership are below the level at which they become concave. However, recently, more sophisticated rules have appeared,such as, for example, the double rules of the Montreal protocol or the Kyoto protocol. I did not analyse these sophisticated rules in this paper. However, what one may conclude from this paper is that the basic minimum participation rule has been progressively refined precisely because the negotiators use it, not only to facilitate the ratification process but more directly as an instrument during the negotiation to manipulate the outcome.

6 References

Aumann and Dreze, (1974), "Cooperative Games with Coalition Structures", International Journal of Game Theory, Vol 3, Issue 4, 217-237.

Bagnoli M. and B. Lipman (1989), "Provision of Public Goods: Fully Implementing the Core through Private Contributions", Review of Economic Studies, 56, 583-601.

Barrett S., *Environment & Statecraft, The Strategy of Environmental Treaty-making*, Oxford University Press, 2003.

Benedick R. E. (1998), *Ozone Diplomacy: New Directions in Safeguarding the Planet*, Enlarged Edition, Cambridge, MA: Harvard University Press.

Bloch F. (1995), "Endogenous Structures of Association in Oligopolies", Rand Journal of Economics, 26, 537-556.

Bloch F. (1996), "Sequential Formation of Coalitions with Externalities and Fixed Payoff Division", Games and Economic Behavior, 14, 90-123.

Carraro C. and D. Siniscalco (1993), "Strategies for the international protection of the environment", Journal of Public Economics, Vol. 52, 309-328.

Conconi P. and C. Perroni (2003), "Self-enforcing international agreements and domestic policy credibility", CESifo Working Paper No. 988.

Courtois P. and G. Haeringer (2006), "The Making of International Environmental Agreements".

Currarini S. and H. Tulkens (2004), "Stable international agreements on transfrontier pollution with ratification constraints", in C. Carraro and V. Fragnelli (eds.), *Game Practice and the Environment*, Cheltenham, Edward Elgar Publishing, 9-36.

Maskin E. (2003), "Bargaining, coalitions and externalities".

Gupta J. and M. Grubb Editors (2000), *Climate Change and European Leadership, A sustainable role for Europe?* Kluwer Academic Publishers.

Hart and Kurz, (1983), "Endogenous Formation of Coalitions", Econometrica, Vol. 51, No 4, 1047-1064.

Heal G. (1993), "Formation of International Environmental Agreements", in C. Carraro (ed.), *Trade, Innovation, Environment*, Dordrecht: Kluwer.

Ledyard J. (1995), "Public Goods: A survey of Experimental Research", in Kagel, J. and Roth, A. (eds.), The Handbook of Experimental Economics, Princeton University Press.

Littlechild, S. C., and G. F. Thompson, (1977), "Aircraft Landing Fees: A Game Theory Approach", Bell Journal of Economics, Vol. 8, n° 1, 186-204.

Lucas and Maceli, () "Discrete Partition Function Games", in Game Theory and Political Science, 191-213.

Moulin, H. and Yves Sprumont (2005), "Fair Allocation of Production Externalities Recent Results", mimeo.

Palfrey T. and Rosenthal H. (1984), "Participation and the Provision of Discrete Public Goods", Journal of Public Economics, 24, 171-193.

Perez-Castrillo D., D. Wettstein (2001) "Bidding for the Surplus: A Non-cooperative Approach to the Shapley Value", Journal of Economic Theory, 100, 274-294.

Ray D. and R. Vohra (1999), "A Theory of Endogenous Coalition Structure", Games and Economic Behavior, 26, 286-336.

Ray D. and R. Vohra (2001), "Coalitional Power of Public Goods", Journal of Political Economy, Vol 109, n 6.

Roth A. E. (1998), *The Shapley value: Essays in honnor of Lloyd S. Shapley*, Cambridge University Press, Cambridge.

Shapley, (1953), "A Value for n-Person Games", in Contributions to the Theory of Games, Vol. II, ed. By Kuhn and Tucker. Priceton; Princeton University Press, 307-317.

Schröder H. (2001), *Negotiating the Kyoto Protocol, An analysis of negotiation dynamics in international negotiations*, Lit verlag Münster - Hamburg - London.

Thoron, S. (1998), "Formation of a Coalition Proof Stable Cartel", Canadian Journal of Economics, February, Vol. 1.

Thoron, S. (2006), "Threshold-Value: the Price of Leadership Commitment".

Thrall, RM and Lucas, WF, '(1963), "n-Person Games in Partition Function Form", Naval Research Logistics Quarterly 10 , 281-298.

Yi, (1997), "Stable Coalition Structure with externalities", Games and Economic Behavior, 20, 201-237.

Kurz (1988) The Shapley Value, Essays in honor of Lloyd S. Shapley, Cambridge University Press.

Von Neumann and Morgenstern (1944), Theory of Games and Economic Behavior, Princeton, Princeton University Press.

Von Stein, J. (2006), "Institutional Design, Ratification, and Compliance in the International Climate Change Regime".