DEPARTMENT OF ECONOMICS

Working Paper

International Environmental Agreements
-The Role of Foresight

Effrosyni Diamantoudi and Eftichios S. Sartzetakis

Working Paper No. 2002-10

ISSN 1396-2426

UNIVERSITY OF AARHUS • DENMARK

INSTITUT FOR ØKONOMI

AFDELING FOR NATIONALØKONOMI - AARHUS UNIVERSITET - BYGNING 350
8000 AARHUS C - ☎ 89 42 11 33 - TELEFAX 86 13 63 34

WORKING PAPER

International Environmental Agreements -The Role of Foresight

Effrosyni Diamantoudi and Eftichios S. Sartzetakis

Working Paper No. 2002-10

DEPARTMENT OF ECONOMICS

SCHOOL OF ECONOMICS AND MANAGEMENT - UNIVERSITY OF AARHUS - BUILDING 350
8000 AARHUS C - DENMARK ☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

International Environmental Agreements
–The Role of Foresight *

Effrosyni Diamantoudi †
Department of Economics, University of Aarhus

Eftichios S. Sartzetakis
Department of Accounting and Finance, University of Macedonia &
Department of Economics, University College of Cariboo

May 2001
(This version June 2002)

Abstract

We examine the formation of International Environmental Agreements (IEAs). We extend the existing literature by endogenizing the reaction of the IEA's members to a deviation by a member or a group of members. We assume that when a country contemplates exiting or joining an agreement, it takes into account the reactions of other countries ignited by its own actions. We identify conditions under which there always exists a unique set of farsighted stable IEAs. The new farsighted IEAs can be much larger than those some of the previous models supported but are not always Pareto efficient. We extend the analysis to allow for coordinated action, that is, groups of countries jointly exiting or entering the agreement and fully characterize the coalitionally farsighted stable IEAs.

*We would like to thank conference participants at CREE 2001 and CTN 2002, and seminar participants at the University of Rochester and CORE for helpful suggestions. We would also like to thank Parkash Chander, Henry Tulkens and Claude D'Aspremont for insightful comments. Parts of this project were completed while Effrosyni Diamantoudi was visiting the Department of Economics at the University of Rochester and CORE. Financial support is acknowledged from the Danish Social Science Research Council.

†Corresponding author: Effrosyni Diamantoudi: *faye@econ.au.dk* Department of Economics, University of Aarhus, Building 322 DK-8000 Aarhus C., Denmark.

1 Introduction

International environmental agreements (IEAs) aim at controlling global pollution (climate change, greenhouse gases, etc.), forming thus, a special case of the public good provision problem. Indeed, although the socially optimum outcome requires that every country participates and ratifies the agreement, each country has an incentive to free ride on the cooperating efforts of every other participant. This incentive stems from the fact that the savings generated by not abating outweigh the marginal environmental damage caused, when every other country agrees to control its emissions.

IEAs are a relatively new brunch of the public good provision theory that goes as far back as at least Lindahl (1919). International agreements differ from a typical public good in that there is no supra-national authority that could possibly enforce the socially optimal outcome. Thus, for an IEA to be ratified and implemented it has to be self-enforcing. Due, primarily, to the latter requirement a number of the theoretical assertions as well as the empirical observations go against the Coasian prediction (1960) that efficient outcomes will prevail.

IEAs are generally signed by a significant number of countries but not all the potential signatories. Typical examples are the Vienna Convention signed by 184 countries, the Montreal Protocol signed by 183 countries, the UNFCCC (United Nations Framework Convention on Climate Change) singed by 186 countries, and lastly the Kyoto protocol signed, so far, by 74 countries. It should be noted that United Nations consists of 189 member states. Although it seems tempting at this point to conclude that, indeed, the grand coalition is (almost) formed and efficiency is (almost) attained, it should be noted that the measures over which each agreement was reached varied significantly both in abatement costs and environmental impact. The fact that the Kyoto protocol was drafted before 1998 but 4 years later is only signed by less than half the countries and has not come into force yet illustrates how the number of signatories may drop significantly if the measures to be undertaken are more substantial.

Given the unclear signals drawn from the empirical observations, there is a wedge drawn between the various theoretical studies on IEAs[1]. On the

[1]For an in depth comparison of the approaches see Tulkens (1998). A more general

one hand there is a series of works initiated by Carraro and Siniscalco (1993) and Barrett (1994) that, through a non-cooperative game theoretic approach, supports the pessimistic view that unless alternative incentives (such as for example technology transfers and/or trade sanctions) are put into force IEAs are futile in the sense that they will be either signed by *very few*[2] countries or, if signed by more, it will be over an agreement without any substance[3]. The disparity between this brunch of the literature and the empirical observations can be attributed to a fundamental assumption embedded in it, namely, the fact that each country upon withdrawing from the agreement assumes that the agreement will remain intact, at least in terms of membership status. It is not surprising that such a hypothesis encourages deviations and undermines the viability of an agreement.

On the other hand, there is a series of works initiated by Chander and Tulkens (1992) and (1997) that, adopting a cooperative game-theoretic framework, asserts the formation of the grand coalition and the attainment of efficiency. Although, closer to the aforementioned empirical findings it does not explain the difficulties encountered in the Kyoto protocol. The basic assumption underlying the Chander and Tulkens (1997) model is that when a country deviates it assumes that the agreement collapses and each country fends for itself. Naturally, the pessimistic expectation of a potential perpetrator deters deviations and encourages the sustainment of more cooperative agreements. Their work, however, sheds light to the theory of IEAs from a normative angle: If an agreement is designed in a way that deviators are indirectly yet effectively punished through the collapse of the agreement then cooperation and social optimality may be attainable after all.

A question that is not new to game theoretic analysis is how credible such a threat is, and if an agreement that is designed along these lines is still self-enforcing. Our work addresses precisely this issue and attempts to

literature review is offered by Ioannidis, Papandreou and Sartzetakis (2000).

[2]Diamantoudi and Sartzetakis (2001) show in the model adopted by Barrett (1994) that the equilibrium size of an IEA is either 2, 3 or 4 countries. Similarly, Finus and Rundshagen (2001) show that in the model adopted by Carraro and Siniscalco (1993) when transfers are not allowed the equilibrium size of an IEA is 2 or 3 countries.

[3]As will show in Section 2, in Barrett's (1994) model, even if we abstract from the specific functional forms and IEAs are signed in equilibrium by a significant number of countries, the agreed upon emission levels are identical to those of the laissez-faire state where each country optimizes individually.

bridge the gap between the two approaches. When a country defects from an agreement it makes no exogenously imposed assumptions regarding the behavior of remaining members of the agreement. Instead, it foresees[4] what their reaction will be, and which equilibrium agreement will result from such a deviation. We offer two analyses: one where countries always act alone whether that is to join or withdraw from an agreement and one where any group of countries may choose to coordinate their actions in either joining or withdrawing from an agreement.

Closer in spirit to the works of Ray and Vohra (1997), (1999) and (2001) we define equilibrium IEAs in a consistent manner. That is, each country, or group of countries, upon deviation anticipates the equilibrium, hence credible result, of its actions and compares that outcome to its status quo, instead of the immediate, yet possibly unstable, one. Although Ray and Vohra's (1997) and (1999) works are general and do not concentrate on the public good provision problem, their (2001) does. The present paper, although closer to their (2001) work thematically, it resembles more their (1997) theoretical approach in the sense that it does not model coalition formation explicitly with an extensive form game but focuses on an equilibrium analysis.

Compared to the concept of equilibrium binding agreements by Ray and Vohra (1997) our analysis imposes two restrictions and relaxes one of their assumptions. The first restriction is that agents are symmetric and although it is, indeed, a serious and limiting assumption it is the most common in the literature, offered if not at the outset at least as a hypothesis in the results. Such an assumption allows us to fully characterize our solution set. The second restriction concerns the set of permissible structures. Instead of considering all possible coalition structures and thus allow for multiple IEAs to develop, we constrain our analysis to the case where only one IEA is allowed to form and the only question remaining is the equilibrium size of this agreement. This assumption is adopted in the works of Carraro and Siniscalco (1993) and Chander and Tulkens (1997) among others. Although this assumption seems rather strong it is actually instigated from our empirical observations. Indeed, IEAs are usually unique and fostered by the United Nations. It is an intriguing question whether this is an equilibrium outcome

[4]Such an analysis is discussed and encouraged by Ecchia and Mariotti (1998) and Carraro and Moriconi (1998).

itself, but this is not a question we address in this project. The assumption adopted by Ray and Vohra (1997) that we relax is that agreements can only shrink in size and never grow. We allow, thus, for the possibility of renegotiation among countries, in the sense that although an IEA may collapse countries can always agree anew upon some larger agreement.

Our analysis starts in Section 2 from Barrett's (1994) model that employs D' Aspremont et. al 's (1983) solution concept of coalitional stability. A major difference between Barrett (1994) and Carraro and Siniscalco (1993) is that the former adopts a leadership model a la Stackelberg where the IEA leads and the non-signatories follow whereas the latter adopts a simultaneous a la Cournot model and considers transfers. Although our analysis can easily be applied to the simultaneous model we chose the leadership one primarily because we wanted our results to be directly comparable to those of Barrett (1994) since we do not consider transfers either[5]. In Subsection 3.1 the solution concept is modified to capture consistency and foresight while unilateral actions are maintained. Subsection 3.2 extends further the notion to allow for coordinated action and the results are significantly closer to those of Chander and Tulkens (1997). Section 4 concludes the paper.

2 The Model

We assume that there exist n identical countries, $N = \{1, ..., n\}$. Production and consumption in each country generate emissions $e_i \geq 0$ of a global pollutant as an output. Thus, the social welfare of country i, w_i, is expressed as the net between the benefits from country i's emission, $B_i(e_i)$, and the damages $D_i(E)$ from the aggregate emissions $E = \sum_{i \in N} e_i$. Since the countries are assumed to be identical we henceforth drop the subscripts from the individual welfare function:

$$w = B(e_i) - D\left(\sum\nolimits_{i \in N} e_i\right).$$

[5] Secondarily, once the IEA is substantially large, as we will show, it seems natural to further assume that the non-signatories will wait to observe the outcome of negotiations and hence behave as followers.

We further assume that the benefit function is strictly concave, that is, $B(0)=0, \quad B' \geq 0$ and $B'' < 0$, and the damage function is strictly convex, that is, $D(0)=0, \quad D' \geq 0$ and $D'' > 0$.

The ratification of the IEA is depicted by the formation of a coalition. In particular, a set of countries $S \subset N$ sign an agreement and $N \backslash S$ do not. Let the size of coalition be denoted by $|S| = s$, total emissions generated by the coalition by E_s while each member of the coalition emits e_s, such that $E_s = se_s$. In a similar manner, each non-signatory emits e_{ns}, giving rise to a total emission level generated by all non-signatories $E_{ns} = (n-s)e_{ns}$. The aggregate emission level is, $E = E_s + E_{ns} = se_s + (n-s)e_{ns}$.

Non-signatories behave non-cooperatively after having observed the choice of signatories. Therefore, their maximization problem gives rise to an indirect welfare function ω_{ns} as follows:

$$\omega_{ns}(e_s, s) = \max_{e_{ns}} \left[B(e_{ns}) - D(se_s + (n-s-1)e_i + e_{ns}) \right].$$

When operating at the optimum, non-signatories' emissions satisfy the condition

$$B'(e^*_{ns}(e_s)) = D'(se_s + (n-s)e^*_{ns}(e_s)),$$

which yields a best response function $e^*_{ns}(e_s, s)$.

Signatories maximize the coalition's welfare, sw_s, taking explicitly into account N/S's behavior. Similarly, the coalition's maximization problem yields an indirect welfare function ω_s as follows:

$$\omega_s(s) = \frac{1}{s} \max_{e_s} \left[sB(e_s) - sD\left[se_s + (n-s)e^*_{ns}(e_s, s) \right] \right].$$

The optimal signatories' emissions $e^*_s(s)$ satisfy the following condition[6]:

$$B'(e^*_s(s)) = D'(se^*_s(s) + (n-s)e^*_{ns}(e^*_s(s), s)) \left[s + (n-s) \frac{\partial e^*_{ns}(e_s, s))}{\partial e_s} |_{e_s = e^*_s(s)} \right].$$

The following proposition establishes that in the presence of cooperation, that is, in the case where $s \geq 1$ it is not always true that those that do not

[6] Note that in the extreme cases where $s = 0$ the model reduces to a Cournot-type competition whereas if $s = n$ all countries cooperate.

cooperate emit more and enjoy higher welfare than those that cooperate. On the contrary, there exists a critical coalition size, below which the signatories emit more and attain higher welfare than the non-signatories and above which the reverse is true. Moreover, this critical coalition size is almost (due to integer adjustments) the worst, in terms of per member welfare level, in comparison with other coalition sizes. Since this critical size can be greater than one, we also establish that the welfare of a signatory, ω_s, is not necessarily increasing as the number of signatories, s, increases. Let e_{nc} and E_{nc} denote the individual and aggregate emissions when there is no agreement and countries behave a la Cournot.

Proposition 1 *Consider the indirect welfare functions of signatory and non-signatory countries, $\omega_s(s)$ and $\omega_{ns}(e_s^*(s), s)$ respectively. Let*

$$z^{\min} = \frac{B''(e_{nc}) - nD''(E_{nc})}{B''(e_{nc}) - D''(E_{nc})}$$

then,

1. $e_s^*(s) \gtreqless e_{ns}^*(s) \Leftrightarrow s \lesseqgtr z^{\min}$,
2. *if $s = z^{\min}$ then $e_s^*(s) = e_{ns}^*(s) = e_{nc}$ (Rutz and Borek (2000))*[7],
3. *$\omega_s(s)$ increases (decreases) in s if $s > z^{\min}$ ($s < z^{\min}$)*,
4. $z^{\min} = \arg\min_{s \in \Re \cap [0,n]} \omega_s(s)$,
5. $\omega_s(s) \gtreqless \omega_{ns}(e_s^*(s), s) \Leftrightarrow s \lesseqgtr z^{\min}$[8]

Proof. Although in our model s is a non-negative integer smaller than n, for the ease of exposition and calculations in the following proof we use z to denote a real number taking values from $[0, n]$. At the end we convert back to integer s.

[7] The fact that the point of intersection between $\omega_s(s)$ and $\omega_{ns}(s)$, when s is a real number, occurs when emission levels are equal to those of the purely non-cooperative case, where there is no leader and firms compete a la Cournot, is due to Rutz and Borek (2000). We are re-stating it as part of proposition 1 in order to present it in terms of our notation and terminology, since Rutz and Borek's model is specified in terms of abatements.

[8] The point of intersection between $\omega_s(s)$ and $\omega_{ns}(s)$ when s is a real number has been identified independently by Rutz and Borek (2000), but the authors did not identify neither the relation between the two functions beyond the point of intersection nor the fact that the point of intersection is the lowest point of $\omega_s(s)$.

1-2. Let $E^*(z)$ denote the aggregate optimal emission levels given a coalition size z and $e^*_{ns}(z) = e^*_{ns}(e^*_s(z), z)$. Since $B'' < 0$ we have $e^*_s(z) \gtreqless e^*_{ns}(z) \Leftrightarrow B'\left(e^*_s(z)\right) \lesseqgtr B'\left(e^*_{ns}(z)\right)$. But in equilibrium we also have

$$\begin{aligned} B'(e^*_s(z)) &\equiv D'(E^*(z))\left[z + (n-z)\frac{\partial e^*_{ns}(e_s)}{\partial e_s}|_{e_s = e^*_s(z)}\right] \\ &\text{and} \\ B'(e^*_{ns}(z)) &\equiv D'(E^*(z)), \end{aligned}$$

thus in equilibrium,

$$\begin{aligned} e^*_s(z) &\gtreqless e^*_{ns}(z) \Leftrightarrow \\ z + (n-z)\frac{\partial e^*_{ns}(e_s)}{\partial e_s}|_{e_s = e^*_s(z)} &\lesseqgtr 1. \end{aligned}$$

From the first order condition of the non-signatories we have that $B'(e^*_{ns}(e_s)) \equiv D'(ze_s + (n-z)e^*_{ns}(e_s))$ which is an identity and differentiating both sides with respect to e_s yields:

$$\begin{aligned} \frac{\partial e^*_{ns}(e_s)}{\partial e_s} &= \frac{zD''(E(e_s))}{B''(e^*_{ns}(e_s)) - (n-z)D''(E(e_s))} \text{ and} \\ \left.\frac{\partial e^*_{ns}(e_s)}{\partial e_s}\right|_{e_s = e^*_s(z)} &= \frac{zD''(E^*(z))}{B''(e^*_{ns}(z)) - (n-z)D''(E^*(z))} \end{aligned}$$

Substituting the latter into the former inequality results in:

$$z + (n-z)\frac{zD''(E^*(z))}{B''(e^*_{ns}(z)) - (n-z)D''(E^*(z))} \lesseqgtr 1.$$

Which reduces to

$$\begin{aligned} e^*_s(z) &\gtreqless e^*_{ns}(z) \Leftrightarrow \\ z &\lesseqgtr \frac{B''(e^*_{ns}(z)) - nD''(E^*(z))}{B''(e^*_{ns}(z)) - D''(E^*(z))}. \end{aligned}$$

Observe that when $e^*_s(z) = e^*_{ns}(z)$ the non-signatories' first order conditions remains satisfied, i.e.,

$$\begin{aligned} B'(e^*_{ns}(z)) &\equiv D'(se^*_s(z) + (n-z)e^*_{ns}(z)) \Leftrightarrow \\ B'(e^*_{ns}(z)) &\equiv D'(ne^*_{ns}(z)) \end{aligned}$$

which is identical to the first order condition of the pure non-cooperative case where countries compete a la Cournot, hence, $e^*_{ns}(z) = e^*_s(z) = e_{nc}$. Note that due to the strict concavity of the benefit function and the strict convexity of the damage function there exists a unique e_{nc} and, thus, a unique $z^{\min} = \frac{B''(e_{nc}) - nD''(E_{nc})}{B''(e_{nc}) - D''(E_{nc})}$. Reverting the coalition size back to integers yields:

$$e^*_s(s) \gtreqless e^*_{ns}(s) \Leftrightarrow s \lesseqgtr z^{\min}.$$

3-4. Since $\omega_s(e^*_s(z)) \equiv B\left(e^*_s(z)\right) - D\left(E^*(z)\right)$ we have

$$\begin{aligned}\frac{d\omega_s(z)}{dz} &= B'(e^*_s(z))\frac{de^*_s(z)}{dz} \\ &\quad -D'(E^*(z))\left[e^*_s(z) - e^*_{ns}(e_s) + \frac{de^*_s(z)}{dz}z + (n-z)\frac{de^*_{ns}(z)}{dz}\right]\end{aligned}$$

which can be rewritten as follows:

$$\begin{aligned}\frac{d\omega_s(z)}{dz} &= \frac{de^*_s(z)}{dz}\left[B'(e^*_s(z)) - zD'(E^*(z))\right] \\ &\quad -D'(E^*(z))[e_s(z) - e_{ns}(e_s)] - D'(E^*(z))(n-z)\frac{de^*_{ns}(z)}{dz}.\end{aligned}$$

But we know that in equilibrium

$$B'(e^*_s(z)) \equiv D'(E^*(z))\left[z + (n-z)\left.\frac{\partial e^*_{ns}(e_s)}{\partial e_s}\right|_{e_s = e^*_s(z)}\right]$$

hence

$$\begin{aligned}\frac{d\omega_s(z)}{dz} &= (n-z)D'(E^*(z))\left[\frac{de^*_s(z)}{dz}\left.\frac{\partial e^*_{ns}(e_s)}{\partial e_s}\right|_{e_s = e^*_s(z)} - \frac{de^*_{ns}(z)}{dz}\right] \\ &\quad -D'(E^*(z))[e^*_s(z) - e^*_{ns}(z)].\end{aligned}$$

We also know from (1-2) above that

$$\left.\frac{\partial e^*_{ns}(e_s)}{\partial e_s}\right|_{e_s = e^*_s(z)} = \frac{zD''(E^*(z))}{B''(e^*_{ns}(z)) - (n-z)D''(E^*(z))}.$$

Moreover, the same first order condition that yields $\frac{\partial e^*_{ns}(e_s)}{\partial e_s}$, in equilibrium becomes

$$B'(e^*_{ns}(z)) \equiv D'(ze^*_s(z) + (n-z)e^*_{ns}(z)).$$

Differentiating both sides with respect to z yields

$$B''(e^*_{ns}(z))\frac{de^*_{ns}(z)}{dz} = D''(E^*(z))\left[e^*_s(z) + z\frac{de^*_s(z)}{dz} - e^*_{ns}(z) + (n-z)\frac{de^*_{ns}(z)}{dz}\right]$$

hence

$$\frac{de^*_{ns}(z)}{dz} = \frac{D''(E^*(z))\left[e^*_s(z) - e^*_{ns}(z)\right] + D''(E^*(z))z\frac{de^*_s(z)}{dz}}{B''(e^*_{ns}(z)) - (n-z)D''(E^*(z))}.$$

Next, we replace $\frac{de^*_{ns}(z)}{dz}$ and $\frac{\partial e^*_{ns}(e_s)}{\partial e_s}$ in $\frac{d\omega_s(z)}{dz}$ which yields

$$\frac{d\omega_s(z)}{dz} = -D'(E^*(z))[e^*_s(z) - e^*_{ns}(z)]\left[\frac{B''(e^*_{ns}(z))}{B''(e^*_{ns}(z)) - (n-z)D''(E^*(z))}\right].$$

Now observe that since $\frac{B''(e^*_{ns}(z))}{B''(e^*_{ns}(z))-(n-z)D''(E^*(z))} > 0$ and $-D'(E^*(z)) < 0$ for all z, the sign of $\frac{d\omega_s(z)}{dz}$ depends solely on $[e^*_s(z) - e^*_{ns}(z)]$. Therefore, given the uniqueness of $z^{\min}$, we can conclude that $\omega_s(s)$ is U-shaped and hence $\left.\frac{d\omega_s(z)}{dz}\right|_{z \lessgtr z^{\min}} \lessgtr 0$. The conversion to integer values of coalition size is trivial.

5. Recall that $\omega_s(s) = B(e^*_s(s)) - D(E^*(s))$ and $\omega_{ns}(s) = B(e^*_{ns}(s)) - D(E^*(s))$. Thus, $\omega_s(s) \lessgtr \omega_{ns}(s) \Leftrightarrow B(e^*_s(s)) \lessgtr B(e^*_{ns}(s))$ and since $B' > 0$ we have $B(e^*_s(s)) \lessgtr B(e^*_{ns}(s)) \Leftrightarrow e^*_s(s) \lessgtr e^*_{ns}(s) \Leftrightarrow s \lessgtr z^{\min}$.

■

The next natural step is the determination of the size of the stable coalitions. In the works of Barrett (1994) and Carraro and Siniscalco (1993) the solution concept used is Nash equilibrium, which examines whether a coalition is immune to unilateral deviations. The aspect of the stability notion that examines the incentives of the existing members of the coalition is internal stability, while the aspect that examines the incentives of the non-signatories is external stability. This notion of stability was first developed by D 'Aspremont et al. (1983) within the context of cartel stability in a price leadership model and was adopted later on by Carraro and Siniscalco (1993) and Barrett (1994) to analyze IEAs. Formally a coalition of size s^* is,

$$\begin{aligned} \text{internally stable if } \omega_s(s^*) &\geq \omega_{ns}(s^*-1) \\ \text{and externally stable if } \omega_{ns}(s^*) &\geq \omega_s(s^*+1). \end{aligned}$$

Note that due to the finite number of countries, once a country exits or enters the coalition, the size of the coalition changes resulting in adjusted emission levels and hence different per member welfare level. Recall that the emissions (and by extension the welfare) of signatories depends on the size of the coalition. Thus, according to the original definition of Nash equilibrium all other agents do not change their behavior once a country exits (or enters) the coalition, however, they do adjust their emissions as a response to the new size. Note that in this general form we do not have an analytical solution. However, in the case of quadratic benefit and damage functions as shown by Diamantoudi and Sartzetakis (2001) in a companion paper the stable coalition is of size 2, 3 or 4.

3 Foresight

3.1 Unilateral Action

Although, it is well established in the literature that the formation of the grand coalition is Pareto efficient (with respect to the agents involved) and that each agent has an incentive to unilaterally free ride on others, the major contribution of D' Aspremont et al. (1983) who first introduced this form of coalitional stability was the observation that if the agents are finite, once a member of the coalition deviates and withdraws from the agreement, the remaining coalition will adjust its behavior. If a country withdraws from the agreement its action will be noticed by the coalition and explicitly taken into consideration. Precisely this consideration leads, in our setting, to the adjustment of their emissions in a manner that maximizes the new aggregate welfare.

Once such an adjustment is captured by the model, the result is that it is not always beneficial for a country to withdraw from the agreement. The increase in its welfare the potential deviant may enjoy by increasing its emission level may be offset by the increase in the coalition's emissions as a result of its adjustment.

In the D' Aspremont et al. (1983) model and its variants within the environmental framework it is assumed that, once a country withdraws, the

coalition will adjust its emissions. In fact, the coalition's adjustment and the deviating agent's ability to foresee this adjustment is the very merit of the model. It is only natural thus, that we allow the deviating country to *fully* foresee what is going to happen after it initiates the disbursement of the coalition.

In this work we endow countries with foresight. In particular we built a solution concept in the spirit of von Neumann and Morgenstern (vN-M) stable set while amending the dominance relation to incorporate forward lookingness, which gives us a set of stable coalitions that would survive *credible* deviations. The issue of credibility and foresight has risen on several occasions in economic models and more fundamentally in solution concepts within an economic or game theoretic context.

In the D' Aspremont et. al (1983) solution concept when an agent contemplates exiting a coalition C_s it compares $\omega_s(s)$, that is, the welfare it enjoys while a member of the coalition, with the welfare it will enjoy once it exits and joins the non-signatories of $N \backslash C_{s-1}$, that is, $\omega_{ns}(s-1)$. This comparison examines the internal stability condition as defined earlier. The agent implicitly assumes that once it deviates, no one else will want to deviate and therefore it will indeed enjoy welfare $\omega_{ns}(s-1)$ with certainty. But this is not always the case, in fact, it is possible that another country may wish to exit coalition C_{s-1}, by now, and join the non-signatories of $N \backslash C_{s-2}$, and so on. Thus, the country should compare its status quo $\omega_s(s)$ with the *final* outcome that will result once it initiates a sequence of events by exiting C_s. This *final* outcome can be characterized as such only if no more countries wish to exit and no more countries wish to join, if, in other words, it is stable itself. Put differently, we can determine whether a coalition is stable or not, only if we know what every other coalition is. Such a circular approach is adopted by the classical notion of the abstract stable set.

The (abstract) stable set originally defined by von Neumann and Morgenstern (1944) is a solution concept that captures consistency. The stable set approach, instead of characterizing each outcome independently, characterizes a solution set, that is, a collection of outcomes that are stable, while those excluded from the solution set are unstable. Moreover, no inner contradictions are allowed, that is, any outcome in the stable set cannot dominate

another outcome also in the stable set.[9] Similarly, every outcome excluded from the stable set is accounted for in a consistent manner by being dominated by some outcome in the stable set.[10] Although the notion of the stable set is very appealing exactly due to the afore mentioned properties that attribute consistency[11] it has been criticized on two grounds. Firstly, it does not always exist, and secondly, it suffers from myopia as well, as depicted by Harsanyi (1974) who suggested that the simple (one step) dominance relation be replaced by indirect dominance that allows agents to consider many steps ahead. His criticism inspired a series of works in abstract environments by Chwe (1994), Mariotti (1997) and Xue (1998) among others.

In the spirit of von Neumann & Morgenstern's (1944) stable set, Harsanyi's (1974) indirect dominance and within the context of coalition formation, we consider a set σ which is the collection of all farsighted stable coalitions, i.e., $\sigma = \{C_s, C_t..., C_f\}$. A coalition C_s is farsighted stable given σ and thus, $C_s \in \sigma$ if it is both internally and externally farsighted stable given σ.

A coalition is considered to be internally stable if no country wishes to exit from it. So far, a country compared its current welfare $\omega_s(s)$ with the welfare of the set of non-signatories it would join, $\omega_{ns}(s-1)$. We claim that such a comparison is justified only if C_{s-1} is a stable coalition itself, i.e., $C_{s-1} \in \sigma$ as well, which would imply that if C_{s-1} becomes the status quo it would remain so. Otherwise, if $C_{s-1} \notin \sigma$ once at C_{s-1} some other country may

[9] This feature of the stable set is known as *Internal Stability*, yet we will avoid the terminology since it coincides with the one attributed to coalitions which is entirely different. Characterizing a coalition as internally stable implies that no member wishes to exit the coalition. This latter meaning of internal stability is the one we will maintain throughout the paper as formalized in Definition 1 that follows.

[10] This feature of the stable set is known as *External Stability*. The same problem with terminology arises here as well. We will maintain the meaning of external stability as formalized in Definition 1.

[11] Its appeal is captured and improved upon by Greenberg (1994). In the *Theory of Social Situations (TOSS)*, a unifying approach towards cooperative and non-cooperative game theory, where any behavioral and institutional assumptions are explicitly defined, an equivalence is shown between the von Neumann & Morgenstern (vN-M) stable set and the *Optimistic Stable Standard of Behavior (OSSB)*, a solution concept built in the spirit of vN-M stability, yet with the precise assumption of optimistic behavior explicitly formalized. TOSS amplified the pertinence of stability by recasting the dominance relation into a broader concept beyond the boundaries of a binary relation. In doing so, behavioral assumptions can be imposed on the agents, and more complex institutional settings can be analyzed.

wish to either join or exit. Thus, the very first country when contemplating whether to exit from C_s or not it should compare its welfare while a member of C_s to the final stable outcome that will arise. A parallel process describes external stability. A coalition C_s is externally stable if no country wishes to join in. Again the country makes such a decision by comparing its welfare under the status quo $\omega_{ns}(s)$ with the welfare it will enjoy once it joins the coalition, namely $\omega_s(s+1)$. Such a comparison is justified only if $C_{s+1} \in \sigma$ is a stable coalition itself and thus, no more countries wish to enter or exit. The country should compare its status quo with the final outcome that will arise. Formally,

Definition 1 *A set of coalitions, σ, is a **farsighted stable set** if*

1. *σ is free of inner contradictions:*

   (a) *Every coalition $C_s \in \sigma$ is **internally farsighted stable**, i.e., there does not exist a finite sequence of coalitions $C_{s-1}, ..., C_{s-m}$, where $m \in \{1, ..., s\}$ such that $C_{s-m} \in \sigma$ and $\omega_s(s-j) < \omega_{ns}(s-m)$ for every $j = 0, 1, ..., m-1$.*

   (b) *Every coalition $C_s \in \sigma$ is **externally farsighted stable**, i.e., there does not exist a finite sequence of coalitions $C_{s+1}, ..., C_{s+m}$, where $m \in \{1, ..., n-s\}$, such that $C_{s+m} \in \sigma$ and $\omega_s(s+m) > \omega_{ns}(s+j)$ for every $j = 0, 1, ..., m-1$.*

2. *σ accounts for every coalition it excludes:*

   That is, for every coalition $C_s \notin \sigma$ either

   (a) *C_s is **internally farsighted unstable**, i.e., there exists a finite sequence of coalitions $C_{s-1}, ..., C_{s-m}$, where $m \in \{1, ..., s\}$ such that $C_{s-m} \in \sigma$ and $\omega_s(s-j) < \omega_{ns}(s-m)$ for every $j = 0, 1, ..., m-1$, or*

   (b) *C_s is **externally farsighted unstable**, i.e., there exist a finite sequence of coalitions $C_{s+1}, ..., C_{s+m}$, where $m \in \{1, ..., n-s\}$, such that $C_{s+m} \in \sigma$ and $\omega_s(s+m) > \omega_{ns}(s+j)$ for every $j = 0, 1, ..., m-1$.*

Note that the null coalition with $s = 0$ that contains no members is trivially internally farsighted stable since there does not exist a country to exit, and that a perfectly collusive situation where the coalition C_n contains all the countries is trivially externally farsighted stable, since there do not exist any more countries to join.

We mentioned earlier that one of the major drawbacks of the stable set is its difficulty to exist. However, the following theorem establishes that there exists a stable set σ as long as there exist a unique myopic stable coalition of size s^*[12].

Another problem associated with the stable set is its multiplicity. More precisely the existence of more than one farsighted stable sets σ, suggesting different collections of farsighted stable outcomes. Notice the difference between uniqueness of a farsighted stable coalition and uniqueness of a set of farsighted stable coalitions. The former, is obviously not true as we will illustrate in the following result were we claim that a sequence of coalitions are farsighted stable, whereas the latter is asserted in the following Theorem when s^* is unique.

Theorem 2 *If there exists a unique myopic stable coalition of size s^* there exists a unique farsighted stable set of farsighted stable coalitions, σ.*

Proof. The proof of Theorem 2 consists of three parts. In the first part we construct set σ with the use of an algorithm that identifies the coalitions. Note that although σ is constructed in terms of coalition sizes for simplicity, it in fact contains all the permutations of each size. They obviously do not contradict each other since when one agreement is induced from an other it has to grow or shrink monotonically. In the second part we show that σ is farsighted stable and in the third part we show that σ is the unique farsighted stable set.

[12] Diamantoudi (2000) within the context of cartel stability in a price leadership model establishes a general result that guarantees the existence of σ when the coalition members' payoff is increasing in the size of the coalition, that is, in our context, when $\omega_s(s)$ is increasing in s.

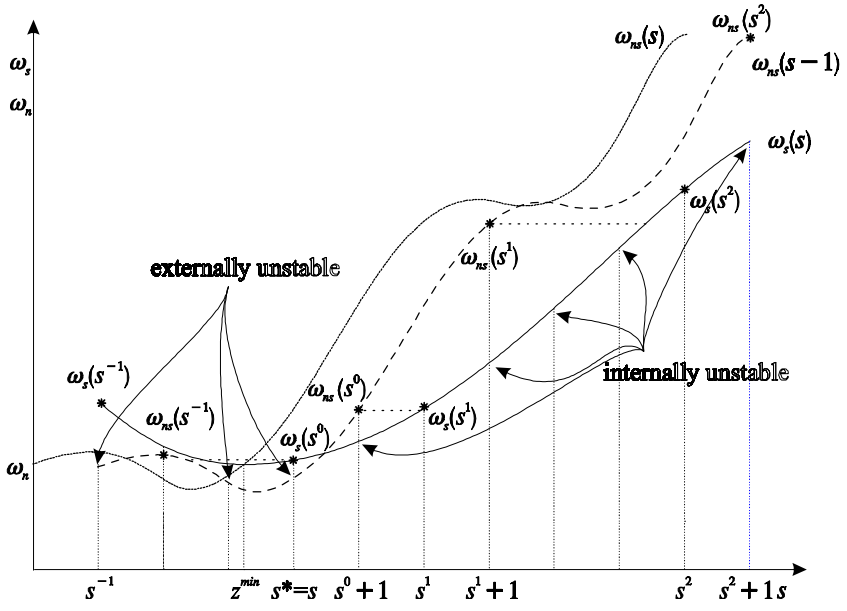

Figure 1

CONSTRUCTION

Step 1: We include C_{s^*} in σ. Let $s^* = s^0$. If $s^0 = n$, then $\sigma = \{N\}$ and we proceed to step 2. Otherwise, we start increasing the coalition size by one until we reach s^1 such that $s^1 > s^0$ and $\omega_s(s^1) \geq \omega_{ns}(s^0)$. In fact, if such a coalition C_{s^1} exists, it does not immediately succeed C_{s^0}, that is $s^1 > s^0 + 1$, otherwise $\omega_s(s^1) = \omega_{ns}(s^0) \Leftrightarrow \omega_s(s^0 + 1) = \omega_{ns}(s^0)$ which would contradict the uniqueness of C_{s^0}. If such a coalition of size s^1 exists, we include it in σ and continue increasing the size of the coalition by one at a time until we reach C_{s^2} where $s^2 > s^1$ and $\omega_s(s^2) \geq \omega_{ns}(s^1)$. If such a coalition of size s^2 exists, we include it in σ. We continue in this manner until we reach N. So far, we have a set $\sigma = \{C_{s^0}, C_{s^1}, ..., C_{s^j}, ..., C_{s^k}\}$ for $j = 0, 1, 2, ..., k$ where $k \leq n-1$, and $s^k \leq n$ such that $\omega_s(s^j + 1) < \omega_{ns}(s^j)$ from the uniqueness of s^* and that $\omega_s(s^{j+1}) \geq \omega_{ns}(s^j)$ for $j \geq 0$ from the construction of σ.

Step 2: We return to C_{s^0} and start decreasing the size by one until we reach $C_{s^{-1}}$ such that $s^{-1} < s^0$ and $\omega_{ns}(s^{-1}) \geq \omega_s(s^0)$. In a similar manner $C_{s^{-1}}$ cannot immediately precede C_{s^0}, that is $s^{-1} < s^0 - 1$, otherwise the uniqueness of C_{s^0} would be contradicted. If such a coalition $C_{s^{-1}}$ exists we include it in σ and continue decreasing the size of the coalition by one until we reach $C_{s^{-2}}$ where $s^{-2} < s^{-1}$ and $\omega_{ns}(s^{-2}) \geq$

$\omega_s(s^{-1})$. If such a coalition of size s^{-2} exists, we include it in σ. We continue in this manner until we reach the null coalition. We now have a set $\sigma = \{C_{s^{-l}}, ..., C_{s^{-i}}, ..., C_{s^{-1}}, C_{s^0}, C_{s^1}, ..., C_{s^j}, ..., C_{s^k}\}$ for $i = 0, 2, ..., l$ where $l \leq n-1$, and $s^{-l} \geq 0$. Moreover, we know that $\omega_s(s^{-i}) > \omega_{ns}(s^{-i}-1)$ from the uniqueness of s^* and that $\omega_s(s^{-i}) \leq \omega_{ns}(s^{-i-1})$ for $i \geq 0$ from the construction of σ. Note that $\sigma \neq \emptyset$ since, there always exists at least $C_{s^0} \in \sigma$.

STABILITY

Next we argue that σ is farsighted stable. We start by showing that all the elements inside σ are both internally and externally farsighted stable. If $|\sigma| = 1$, that is, it is a singleton then the one coalition in it is trivially both internally and externally farsighted stable since countries have essentially no other alternative, no other *farsighted stable coalition* or non-signatories group, smaller or larger to want to deviate to.

If $|\sigma| > 1$, that is σ contains more than one coalition we start by considering C_{s^0}. It is internally farsighted stable because if its members exit they would *hope* to end up as non-signatories of $C_{s^{-1}}$ which is the next smaller farsighted stable coalition. Indeed $\omega_{ns}(s^{-1}) \geq \omega_s(s^0)$ by construction. However, *at least* at the last step of the sequence no member would wish to exit the coalition any longer since $\omega_{ns}(s^{-1}) < \omega_s(s^{-1}+1)$. C_{s^0} is externally farsighted stable because even though $\omega_s(s^1) \geq \omega_{ns}(s^0)$, by the construction of σ, we also have that $\omega_s(s^1) < \omega_{ns}(s^1-1)$, thus at least at the last step of the sequence that would lead us to C_{s^1} no country wishes to join C_{s^1}.

Next consider some C_{s^j}, where $1 < j < k$. It is internally farsighted stable because $\omega_s(s^j) \geq \omega_{ns}(s^{j-1})$ hence no member wishes to exit and end up at $C_{s^{j-1}}$. Note that if a country exits C_{s^j} it will end up at $C_{s^{j-1}}$ since $\omega_s(s^j-1) < \omega_{ns}(s^{j-1}), ..., \omega_s(s^{j-1}+1) < \omega_{ns}(s^{j-1})$, thus more countries will exit. If a country, member of C_{s^j}, exits its coalition it induces $C_{s^j}-1$, nevertheless, once $C_{s^j}-1$ is induced, another country will join the group of non-signatories of $C_{s^j}-2$, and then another country will join the group of non-signatories of $C_{s^j}-3$, and so on until they reach $C_{s^{j-1}}$, since they all prefer the final outcome $\omega_{ns}(s^{j-1})$ to their current coalition except the very first country that exited C_{s^j}. Note that $C_{s^{j-1}}$ is the terminal outcome since, $C_{s^{j-1}} \in \sigma$, i.e., it is a farsighted stable outcome by construction.

C_{s^j} is externally stable since $\omega_s(s^{j+1}) < \omega_{ns}(s^{j+1}-1)$ even though $\omega_s(s^{j+1}) \geq \omega_{ns}(s^j)$. That is, although members of the group of non-signatories of C_{s^j}

would like to join the coalition $C_{s^{j+1}}$ since they prefer it to their current situation, *at least* at the last step of the sequence that would lead to $C_{s^{j+1}}$ no more members of the non-signatories of $C_{s^{j+1}} - 1$ wish to join the coalition $C_{s^{j+1}}$.

Next consider C_{s^k}, which is externally farsighted stable since there is no other larger farsighted stable coalition for the non-signatories to want to join. C_{s^k} is internally farsighted stable for the same reason C_{s^j} is, that is, because $\omega_s(s^k) \geq \omega_{ns}(s^{k-1})$ and $\omega_s(s^k - 1) < \omega_{ns}(s^{k-1}), ..., \omega_s(s^{k-1} + 1) < \omega_{ns}(s^{k-1})$.

Now consider some $C_{s^{-i}}$. It is internally stable for the same reason C_{s^0} is, that is, if its members exit they would *hope* to end up as non-signatories of $C_{s^{-i-1}}$ which is the next smaller farsighted stable coalition. Indeed $\omega_{ns}(s^{-i-1}) \geq \omega_s(s^{-i})$ by construction. However, at least at the last step of the sequence no member would wish to exit its coalition any longer since $\omega_{ns}(s^{-i-1}) < \omega_s(s^{-i-1} + 1)$. $C_{s^{-i}}$ is externally stable since if a member joins it will end up being a member of $C_{s^{-i+1}}$ which is the next larger farsighted stable coalition and it is not preferred, i.e., $\omega_{ns}(s^{-i}) \geq \omega_s(s^{-i+1})$. Note that the joining member will end up at $C_{s^{-i+1}}$ since $\omega_{ns}(s^{-i} + 1) < \omega_s(s^{-i+1}), ..., \omega_{ns}(s^{-i+1} - 1) < \omega_s(s^{-i+1})$.

Lastly, consider $C_{s^{-l}}$. It is internally stable since there does not exist any other smaller farsighted stable coalition. It is externally stable for the same reason $C_{s^{-i}}$ is.

Next we will argue that every $C_s \notin \sigma$ is accounted for, that is, either its internal or external farsighted stability is violated. Observe that, if $s^{-l} > 0$, all C_s such that $0 \leq s < s^{-l}$ are externally unstable since by construction of σ, $\omega_s(s^{-l}) > \omega_{ns}(s)$ for all $s < s^{-l}$. Similarly, all C_s such that $s^{-i} \leq s < s^{-i+1}$ are externally unstable since by construction of σ we have $\omega_s(s^{-i+1}) > \omega_{ns}(s)$.

Next we consider C_s such that $s^j \leq s < s^{j+1}$. By construction of σ we know that $\omega_s(s^j + 1) < \omega_{ns}(s^j), ..., \omega_s(s^{j+1} - 1) < \omega_{ns}(s^j)$, thus coalitions $C_{s^j} + 1, ..., C_{s^{j+1}} - 1$ are internally farsighted unstable since their members wish to become non-signatories of coalition C_{s^j}.

If $s^k < n$, we consider C_s such that $s^k < s \leq n$. By construction of σ we know that $\omega_s(n) < \omega_{ns}(s^k), ..., \omega_s(s^k + 1) < \omega_{ns}(s^k)$, thus coalitions $C_{s^k+1}, ..., C_n$ are internally farsighted unstable since their members wish to join the group of non-signatories C_{s^k}.

UNIQUENESS

Uniqueness stems from the fact that C_{s^0} is internally and externally farsighted stable regardless of the composition of σ. In particular, since C_{s^0} is unique we know that $\omega_{ns}(s-1) < \omega_s(s)$ for all $s \leq s^0$ and $\omega_{ns}(s) > \omega_s(s+1)$ for all $s \geq s^0$. Therefore, no sequence of coalitions starting from C_{s^0} can lead to a smaller coalition since it will collapse at least at the last step. Similarly, no sequence of coalitions starting from C_{s^0} will lead to a larger coalition since again it will collapse at least at the last step. Once C_{s^0} is included in every σ two groups of coalitions are excluded from every σ, namely those between C_{s^0} and C_{s^1} due to internal instability and those between C_{s^0} and $C_{s^{-1}}$ due to external instability. Continuing in this manner we end up with a unique σ as constructed earlier. ■

It is important to mention that the assumption of uniqueness of C_{s^*} is no too severe. Is is shown in Diamantoudi and Sartzetakis (2001) that for the case where both the benefit and damage functions are quadratic the myopic stable coalition is not only unique but it takes the values of 2, 3 or 4. Moreover, it is known in the literature that in the event of a quadratic benefit and linear damage function the myopic stable coalition is also unique and takes the values of 2 or 3.

The conclusion we can draw from the full characterization of σ is that if some farsighted stable IEA contained in it is proposed it will be adopted. From a normative point of view when a coordinating agency (such as UN for example) puts forth a proposal it can always select the most efficient, hence the largest, among the IEAs included in σ. If, however, some agreement that is not farsighted stable is proposed depending on whether it is internally or externally unstable the final outcome will be a smaller or a larger stable agreement respectively.

3.2 Coordinated Action

A natural extension of our analysis so far is to allow countries to coordinate their actions. In other words to allow for coalitional deviations as opposed to unilateral ones in both directions. That is, countries can coordinate their actions and exit jointly from an agreement, but most importantly countries can coordinate their efforts in joining the agreement. Case in point is the fully coordinated ratification of the Kyoto Protocol by the European Union

member states on the 31 of May 2002. It is, indeed, often the case that countries join an environmental agreement in groups through international conventions and meetings and not one by one.

Allowing countries to coordinate their moves enhances their bargaining power in the sense that a group of countries can cause more damage than a single country and hence pose a bigger threat. At the same time simultaneous multilateral action may alleviate the disadvantage of the first mover advantage as is the case in the unilateral action framework. Consider, Figure 1 and notice that the farsighted stable IEA of size s^2 is a Pareto improvement on farsighted stable IEA of size s^1, for all agents involved, signatories and non-signatories alike. What stops, therefore, a group of non-signatories to join in? The very fact that they cannot do it together. As argued in the proof of Theorem 2, assume that all countries start entering the agreement sequentially, at least at the last step, the very last country does not have an incentive to enter anymore. The agreement has grown sufficiently and the country is better off remaining a non-signatory and free-riding thus on the efforts of those that moved earlier. This argument is very similar to that developed by Chander (1998).

To formally define our notion we need to introduce some additional notation. In particular, we need to differentiate between the coalition that represents the agreement and the group of countries that coordinate their moves. Let $T \subset N$ denote a deviating group of countries[13] and $|T| = t$ its size.

Next we specify and appropriately denote what a coalition can (directly) induce or bring about. If a group of countries T deviates from within an agreement C_s, i.e., $T \subset C_s$ then all this deviating coalition can induce is a smaller agreement $C_{s-t} = C_s \backslash T$ where $T \subset N \backslash C_{s-t}$. Whereas, if a group of countries T deviates from the set of non-signatories $N \backslash C_s$, i.e., $T \subset N \backslash C_s$ then all this deviating coalition can induce is a larger agreement $C_{s+t} = C_s \cup T$. We will write $C_s \xrightarrow{T} C_q$ to denote that a group of countries T is inducing agreement C_q from agreement C_s. A further specification of whether $T \subset C_s$ or $T \subset N \backslash C_s$ will tell us whether C_q is larger or smaller. Note that *just* for simplicity we do not allow a deviating group of countries to contain

[13] Often to clarify matters further we will precede the notation with either the acronym IEA or the word agreement.

both signatories and non-signatories. Such a restriction has no implications due to the assumption of symmetric countries. Suppose to the contrary that there exists some group of countries T that consists of both types (signatories and non-signatories) and this group prefers some other agreement. Then, by symmetry, all countries prefer this other agreement which can be reached directly via "pure" deviating groups in the following manner: if the preferred agreement is larger then a subset of the non-signatories can join in, if the preferred agreement is smaller then a subset of the signatories can exit.

Lastly, we need to specify and denote coalitional preferences. When a (homogeneous) group of countries T compares its status quo with some agreement it can directly induce, it knows exactly its payoffs. That is, if $T \subset C_s$ and $C_s \xrightarrow{T} C_{s-t}$ then we say that $C_{s-t} \underset{T}{\succcurlyeq} C_s$ if and only if $\omega_{ns}(s-t) \geq \omega_s(s)$. If $T \subset N \backslash C_s$ and $C_s \xrightarrow{T} C_{s+t}$ then we say that $C_{s+t} \underset{T}{\succcurlyeq} C_s$ if and only if $\omega_s(s+t) \geq \omega_{ns}(s)$. If, however, $T \subset C_s$ but T cannot directly induce some agreement C_q from agreement C_s, we will say that $C_q \underset{T}{\succcurlyeq} C_s$ if and only if $\min\{\omega_s(q), \omega_{ns}(q)\} \geq \omega_s(s)$. If $T \subset N \backslash C_s$ and again T cannot induce C_q from S we will say that $C_q \underset{T}{\succcurlyeq} C_s$ if and only if $\min\{\omega_s(q), \omega_{ns}(q)\} \geq \omega_{ns}(s)$. Naturally, all preferences become strict or get reversed when the inequalities become strict or get reversed, respectively. This conservative approach in coalitional preferences is adopted to avoid the (overly optimistic) tendency of changing roles within the same agreement: Consider the case where some IEA, C_s is already assumed stable and C_s is large enough so that it awards its members higher welfare than the purely non-cooperative state a la Cournot, C_{nc}, and that $\omega_{ns}(s) > \omega_s(s)$. Then, members of the agreement will be tempted to break the agreement and induce C_{nc} in the hope that some other group of countries will now induce C_s and the original perpetrators will now become non-signatories of C_s. Under such a scenario σ^* may fail to exist.

In the solution concept that follows we consider a set σ^* to be a collection of coalitionally farsighted stable IEAs if every IEA that belongs to the set is not in conflict with any other IEAs also in the set, moreover, all IEAs excluded from σ^* are accounted for in a consistent manner. In particular, an IEA is not in conflict with an other IEA if there does not exist a sequence of coalitional moves that can eventually induce one agreement from the other, while along the path of these coordinated moves every acting coalition prefers

the final agreement to its status quo. Furthermore, an IEA is accounted for in a consistent manner when there exists a sequence of coalitional moves that can lead to an IEA that is stable itself, i.e., in σ^*, and again every acting coalition prefers the final agreement to its status quo. Consistency is achieved because the "dominating" outcome -the very reason for deviating- is stable (credible) itself.

Definition 2 *A set of IEAs, σ^*, is a **coalitionally farsighted stable set** if*

1. *σ^* is free of inner contradictions:*

   *Every IEA $C_s \in \sigma^*$ is **coalitionally farsighted stable**, i.e., there does not exist a finite sequence of coalitions $T^0, ..., T^m$, and a sequence of IEAs $C_{s^0}, ..., C_{s^m}$ such that $C_{s^0} = C_s$, $C_{s^m} \in \sigma^*$, $C_{s^j} \xrightarrow{T^j} C_{s^{j+1}}$ and $C_{s^j} \underset{T^j}{\prec} C_{s^m}$ for every $j = 0, 1, ..., m-1$.*

2. *σ^* accounts for every IEA it excludes:*

   Every $C_s \notin \sigma^*$ is coalitionally farsighted unstable, i.e., there exist a finite sequence of coalitions $T^0, ..., T^m$, and a sequence of IEAs $C_{s^0}, ..., C_{s^m}$ such that $C_{s^0} = C_s$, $C_{s^m} \in \sigma^*$, $C_{s^j} \xrightarrow{T^j} C_{s^{j+1}}$ and $C_{s^j} \underset{T^j}{\prec} C_{s^m}$ for every $j = 0, 1, ..., m-1$.

In the following result we offer a full characterization of all possible σ^*s. In particular, we establish that the grand coalition can always be supported as a coalitionally farsighted stable set. While all other σ^*, if payoffs are generic, contain exactly one element that is also Pareto efficient. Let $C_{\hat{s}}$ denote the agreement whose *non-signatories* attain the highest payoff among the non-signatories of all agreements where $s \leq s^*$. That is, $\hat{s} = \arg\max_{s \in \{0,...,s^*\}} \omega_{ns}(s)$. The following result, similarly to Theorem 2, is expressed in term of agreement sizes. Thus, although σ^* may contain only one "size", it supports all agreements (permutations) of that size. Note that any two such agreements cannot contradict or dominate each other due to the conservative approach adopted in the coalitional preferences defined earlier.

Theorem 3 *1. $\sigma^* = \{C_n\}$ is a coalitionally farsighted stable set of IEAs.*

2. $\sigma^* = \{C_s\}$ where $C_s \neq C_n$ is a coalitionally farsighted stable set if and only if $\omega_{ns}(s) > \omega_s(n)$ and $\omega_s(s) > \omega_{ns}(\hat{s})$.

3. There exists $\sigma^* = \{C_{\bar{s}}, C_{\tilde{s}}\}$ if and only if $\exists \tilde{s}, \bar{s} \in \{0, \ldots, n\}$ such that $\tilde{s} \neq \bar{s}$, $\omega_s(\tilde{s}) = \omega_{ns}(\bar{s})$, $\omega_{ns}(\tilde{s}) > \omega_s(n)$, $\omega_s(\bar{s}) \geq \omega_{ns}(\hat{s})$ and $\omega_{ns}(\bar{s}) > \omega_{ns}(r)$ for all $\bar{s} < r < z^{\min}$.

4. There does not exist σ^* such that $|\sigma^*| > 2$.

Proof. The figures inserted in each part of the proof are intended to help the reader follow through the various cases. When an agreement C_s is marked with a star it means that is supported by *some* σ^* in one of the three cases. When an agreement C_s is circled, besides the star, then the graph depicts the case where that particular C_s is included in σ^*. Given the specific σ^* the single arrows denote all the unstable IEAs while the double arrows illustrate some of the typical paths that lead from an unstable agreement to a stable one.

1. Inner contradictions are obviously not present since σ^* contains only one element.

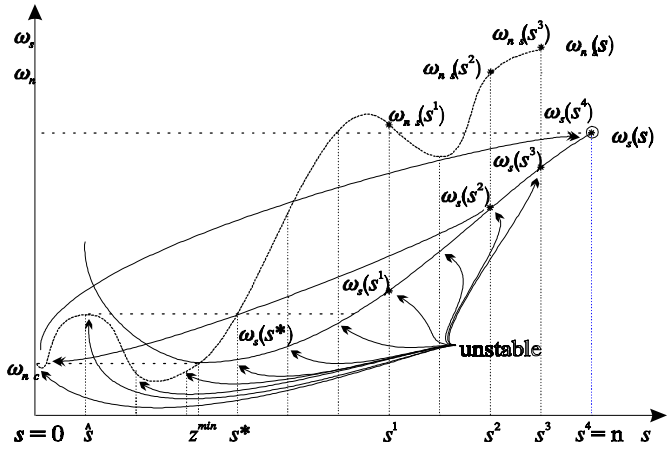

Figure 2

To see how every $C_s \notin \sigma^*$ is accounted for note that $\omega_s(n) > \omega_{ns}(\hat{s})$, otherwise, since $\omega_s(\hat{s}) > \omega_{ns}(\hat{s})$ we would have $\omega_s(n) \leq \omega_{ns}(\hat{s}) < \omega_s(\hat{s})$ and $\omega_s(n)$ would not be maximizing the total welfare. Now consider some IEA C_s such that $\omega_{ns}(s) < \omega_s(n)$. Such an IEA is coalitionally farsighted unstable since there exists coalition $T^0 = N \backslash C_s$ that prefers $C_n \in \sigma^*$ to C_s and can directly induce C_n from C_s by joining C_s. If there exists C_s such that $\omega_{ns}(s) \geq \omega_s(n)$ it must be the case that $\omega_s(s) < \omega_s(n)$. In this case, there exists coalition $T^0 = C_s$ that prefers $C_n \in \sigma^*$ to C_s and can induce $C_0 = C_{nc}$ (the state where $s = 0$ and all countries behave a la Cournot) by exiting the agreement. Then, coalition $T^1 = N$ will induce $C_n \in \sigma^*$ from C_0, since C_n is both inducible from and preferred to C_0 via coalition T^1.

2. When $\sigma^* = \{C_s\}$ obviously there are no inner contradictions.

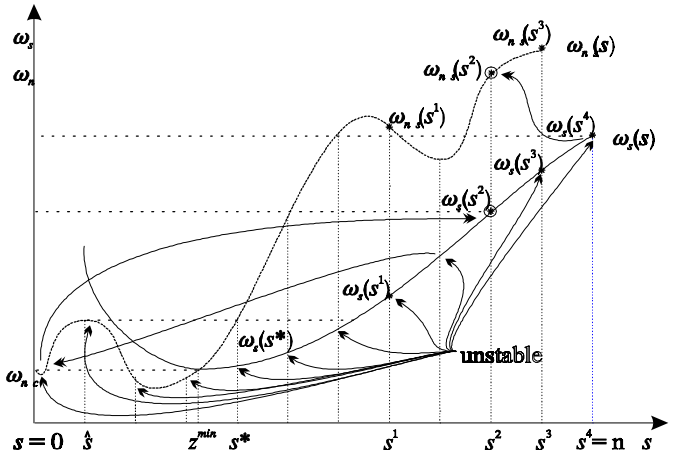

Figure 3

Note that since C_s is such that $\omega_{ns}(s) > \omega_s(n)$ it must be that $s > z^{\min}$. All C_r such that $r > s$ are coalitionally farsighted unstable since $\omega_{ns}(s) > \omega_s(n) > \omega_s(r)$ and members of C_r will exit the agreement and directly induce C_s. All C_r such that $r \leq z^{\min}$ are coalitionally farsighted unstable since $\omega_s(s) > \omega_{ns}(\hat{s}) \geq \omega_{ns}(r)$ and members of $N \backslash C_r$ will join the agreement and directly induce C_s. All C_r such that $z^{\min} < r < s$ are coalitionally farsighted unstable since $\omega_{ns}(s) > \omega_s(s) > \omega_s(r)$ and

coalition $T^0 = C_r$ can induce C_0 from C_r and then coalition $T^1 = C_s$ can and would like to (directly) induce coalition C_s.

Observe that if $\sigma^* = \{C_s\}$ is a coalitionally farsighted stable set and $C_s \neq C_n$ then it must be that $\omega_{ns}(s) > \omega_s(n)$ for the exclusion of C_n to be accounted for. Since $\omega_{ns}(s) > \omega_s(n)$ it is also implied that $C_s \neq C_{\hat{s}}$, hence $\omega_s(s) > \omega_{ns}(\hat{s})$ is necessary for the exclusion $C_{\hat{s}}$, recall that $\omega_s(\hat{s}) > \omega_{ns}(\hat{s})$.

3. Firstly note that if $\omega_{ns}(\tilde{s}) > \omega_s(n)$ then it must be the case that $\tilde{s} \in \{s^*, ..., n\}$. Then, $\omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$ implies that $\bar{s} \in \{0, ..., s^* - 1\}$. Note also that it may be the case that $C_{\bar{s}} = C_{\hat{s}}$.

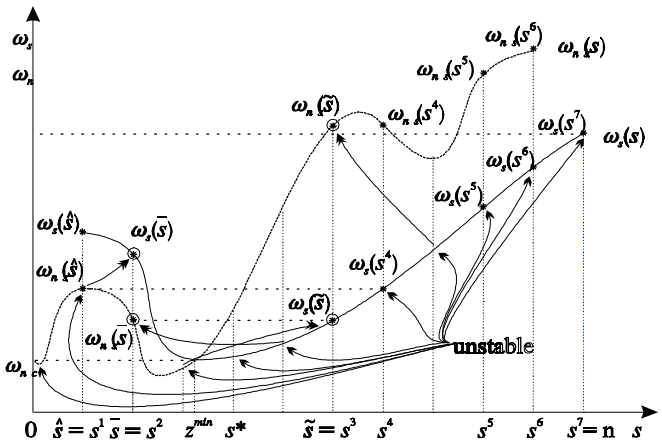

Figure 4

First we show that there are no inner contradictions: No signatories of $C_{\bar{s}}$ can directly induce $C_{\tilde{s}}$ and since $\omega_s(\bar{s}) \geq \omega_{ns}(\bar{s}) = \min\{\omega_s(\tilde{s}), \omega_{ns}(\tilde{s})\}$ no member will initiate an action. No non-signatories of $C_{\bar{s}}$ can directly induce $C_{\tilde{s}}$ either and since $\omega_{ns}(\bar{s}) = \min\{\omega_s(\tilde{s}), \omega_{ns}(\tilde{s})\}$ they will not deviate. Similarly, no members of $C_{\tilde{s}}$ wish to join $N \backslash C_{\bar{s}}$ because $\omega_s(\tilde{s}) = \omega_{ns}(\bar{s})$. As for the non-signatories of $C_{\tilde{s}}$, they will not initiate an action either since they cannot directly induce $C_{\bar{s}}$ and $\omega_{ns}(\tilde{s}) \geq \omega_s(\tilde{s}) = \min\{\omega_s(\bar{s}), \omega_{ns}(\bar{s})\}$.

Next we show that every excluded outcome is accounted for:

All agreements larger than $C_{\tilde{s}}$ are unstable since signatories will exit and directly induce $C_{\tilde{s}}$ since $\omega_{ns}(\tilde{s}) > \omega_s(n) > \omega_s(\tilde{s})$ for all $s \geq \tilde{s}$. Any agreement C_s such that $z^{\min} \leq s < \tilde{s}$ is also unstable: its members wish to break apart and join the non-signatories of $C_{\bar{s}}$ since $\omega_s(s) < \omega_s(\tilde{s}) = \omega_{ns}(\bar{s})$. Any agreement C_s such that $\bar{s} < s < z^{\min}$ is unstable since the non-signatories can directly induce $C_{\tilde{s}}$ since $\omega_{ns}(s) < \omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$. Lastly, any agreement C_s such that $s < \bar{s}$ is unstable since its non-signatories will directly induce $C_{\bar{s}}$ since $\omega_{ns}(s) < \omega_{ns}(\hat{s}) < \omega_s(\bar{s})$.

Lastly we will argue that all the conditions presented in (3) so far are necessary: Observe from the argument presented in (1) that C_n cannot coexist with any other agreement in σ^*. But if $C_n \notin \sigma^*$ it must be that $\exists C_s \in \sigma^*$ such that $\omega_{ns}(s) > \omega_s(n)$ for the exclusion of C_n to be accounted for. Let this IEA be $C_{\tilde{s}}$ and note that the latter inequality implies that $\tilde{s} \in \{s^*, ..., n\}$. Next observe that for $C_{\bar{s}} \in \sigma^*$, $C_{\bar{s}} \neq C_{\tilde{s}}$ it must be the case that $\omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$. Otherwise, if $\omega_{ns}(\bar{s}) > \omega_s(\tilde{s})$ the members of $C_{\tilde{s}}$ would exit and induce $C_{\bar{s}}$, while if $\omega_{ns}(\bar{s}) < \omega_s(\tilde{s})$ then members of $N \backslash C_{\bar{s}}$ would join the IEA and induce $C_{\tilde{s}}$. Note that $\omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$ implies that $\bar{s} \in \{0, ..., z^{\min}\}$. If $\bar{s} = \hat{s}$ then the rest of the conditions are immediately implied by the definition of $C_{\hat{s}}$. If $\bar{s} \neq \hat{s}$ then from the definition of $C_{\hat{s}}$ we have $\omega_{ns}(\hat{s}) > \omega_{ns}(\bar{s})$ and hence $\omega_{ns}(\hat{s}) > \omega_{ns}(\tilde{s})$. Therefore, since $C_{\tilde{s}}$ cannot be directly induced by $C_{\hat{s}}$ and $\omega_s(\hat{s}) > \min\{\omega_s(\tilde{s}), \omega_{ns}(\tilde{s})\}$, $C_{\tilde{s}}$ does not account for $C_{\hat{s}}$'s exclusion. For $C_{\bar{s}}$ to account for the exclusion of $C_{\hat{s}}$, since $\omega_s(\hat{s}) > \omega_{ns}(\hat{s}) > \omega_{ns}(\bar{s})$ it must be the case that $\omega_s(\bar{s}) > \omega_{ns}(\hat{s})$. Lastly, note that for all C_r such that $\bar{s} < r < z^{\min}$ to be accounted for it must be the case that $\omega_{ns}(r) < \omega_{ns}(\bar{s})$. Otherwise, if $\omega_{ns}(r) \geq \omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$, $C_{\tilde{s}}$ cannot account for C_r's exclusion since $\omega_s(r) > \omega_{ns}(r) \geq \omega_s(\tilde{s})$ and $N \backslash C_{\tilde{s}}$ is not directly inducible by either C_r or $N \backslash C_r$. Similarly, $C_{\bar{s}}$ cannot account for the exclusion of C_r either since $\omega_{ns}(\bar{s}) \leq \omega_{ns}(r) < \omega_s(r)$ and $C_{\bar{s}}$ is not directly inducible by either C_r or $N \backslash C_r$.

4. Suppose in negation that there exists σ^* such that $|\sigma^*| > 2$. It is obvious from the arguments presented in (1-2) that such a σ^* will not contain either C_n or C_s if $\omega_s(s) > \omega_{ns}(\hat{s})$. Suppose now that it contains some C_s such that $\omega_s(s) \leq \omega_{ns}(\hat{s})$. Then from the arguments presented in

(3) σ^* cannot contain $C_{\hat{s}}$ or any other $C_{\bar{s}}$ where $\omega_{ns}(\bar{s}) = \omega_s(s)$ either. Finally, in (3) we also argued that C_s cannot coexist with any other $C_{s'}$ if $\omega_{ns}(s') > \omega_s(\tilde{s})$ or $\omega_{ns}(s') < \omega_s(\tilde{s})$.

■

As can be easily seen from the characterization of the two analyses, unilateral and coordinated, the latter supports more socially beneficial IEAs than the former. When actions can be coordinated, countries can use the complete collapse of the agreement as a threat to sustain it. The insights drawn from the results of the coordinated action analysis can have a significant effect from a normative perspective. In particular, the rules for entry into force of the Kyoto Protocol require 55 Parties to the Convention to ratify the Protocol, including enough countries accounting for 55% of a certain groups carbon dioxide emissions in 1990. It remains an interesting puzzle whether had the rules for participation been stronger, we would be observing such a reluctance from a number of countries to ratify the agreement. In other words, has the commitment (through the ratification process) of about half the countries encouraged the other half to free-ride? An answer to this question is far from trivial, as the theoretical restrictions along with several (economic and political) factors that are not even discussed in this work play a crucial role in the negotiation process.

4 Epilogue

The present paper examines the problem of deriving the size of farsighted stable IEAs. We assume that when a country or a group of countries contemplates exiting (or entering) an agreement, it takes into account the actions of other groups of countries ignited by its own. That is, it compares its current welfare, as part of the IEA (or outside the IEA) with its ultimate welfare resulting from its own exit (or entry) and the subsequent reactions by other countries. We provide two solution concepts allowing for unilateral and coordinated actions that capture exactly such a decision-making process and show that the new (coalitional) farsighted stable coalitions are much larger than those the previous models supported. In this manner, we explain better the fact that IEAs are already ratified by a large number of countries while

we provide formal arguments that would encourage an even larger number of countries to join in.

The first and foremost step in extending this work is the study of the heterogeneous country case. Although both our definitions can be trivially extended to accommodate asymmetric decision makers and existence results may be possible to attain, a full characterization of the solution set like the ones offered in this work would be very difficult to obtain.

5 References

1. D' Aspremont, C.A., Jacquemin, J. Gabszeweiz, J., and Weymark, J.A. (1983), "On the stability of collusive price leadership." *Canadian journal of Economics*, **16**, 17-25.

2. Barrett, S. (1994), "Self-enforcing international environmental agreements." *Oxford Economic Papers*, **46**, 878-894.

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

4. Carraro, C. and Moriconi, F. (1998), "International Games on Climate Change Control." FEEM working paper, 56.98.

5. Chander, P. (1999), "International Treaties on Global Pollution: A Dynamic Time-Path Analysis." Raut,-Lakshmi-K., eds. Trade, growth and development: Essays in honor of Professor T. N. Srinivasan. Contributions to Economic Analysis, vol. 242. Amsterdam; New York and Oxford: Elsevier Science, North-Holland.

6. Chander, P. and H. Tulkens (1992), "Theoretical Foundations of Negotiations and Cost-sharing in Transfrontier Pollution Problems." *European Economic Review* **36**, 388-398.

7. Chander, P. and H. Tulkens (1997), "The Core of an Economy with Multilateral Environmental Externalities." *International Journal of Game Theory* **26**, 379-401.

8. Chwe, M. S.-Y. (1994). "Farsighted Coalitional Stability." *Journal of Economic Theory* **63**, 299-325.

9. Coase, R. (1960). "The Problem of Social Cost." *Journal of Law and Economics* **3,** 1-44.

10. Diamantoudi, E. (2000). "Stable Cartels Revisited." University of Aarhus, working paper 2001-09.

11. Diamantoudi, E. and Sartzetakis, E. (2001). "Stable International Environmental Agreements: An Analytical Approach." University of Aarhus, working paper 2001-10 .

12. Ecchia, G. and Mariotti, M. (1998). "Coalition Formation in International Environmental Agreements and the Role of Institutions." *European Economic Review*, **42**, 573-582.

13. Finus, M. and Rundshagen, B. (2001), "Endogenous Coalition Formation in Global Pollution Control." FEEM, working paper, 43.2001.

14. Greenberg, J. (1990). *The Theory of Social Situations: An Alternative Game-Theoretic Approach.* Cambridge University Press.

15. Harsanyi, J. C. (1974). "Interpretation of Stable Sets and a Proposed Alternative Definition." *Management Science* **20**, 1472-1495.

16. Ioannidis, A., Papandreou, A. and Sartzetakis E. (2000), "International Environmental Agreements: A Literature Review." GREEN working paper, Universite Laval 00-08.

17. Lindahl, E. (1919), "Just Taxation- A Positive Solution." reprinted in *Classics in the Theory of Public Finance*, (1967), Eds. R. Musgrave and A. Peacock, Martins Press, New York.

18. Mariotti, M. (1997), "A Model of Agreements in Strategic Form Games." *Journal of Economic Theory*, **74**, p.196-217.

19. Ray, D. and Vohra, R. (1997). "Equilibrium Binding Agreements." *Journal of Economic Theory* **73**, 30-78.

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

21. Ray, D. and Vohra, R. (2001). "Coalitional Power and Public Goods." *Journal of Political Economy* ***109***, 1355-1384.

22. Rutz, S. and Borek, T. (2000). "International Environmental Negotiations: Does Coalition Size Matter?" WIF ETH working paper 00/20.

23. Tulkens, H. (1998), "Cooperation versus Free-Riding in International Environmental Affairs: Two Approaches." In Nick Hanley and Henk Folmer (editors): *Game Theory and the Environment*, E. Elgar, Cheltenham, UK.

24. von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

25. Xue, L. (1998). "Coalitional Stability under Perfect Foresight." *Economic Theory* ***11*** *603-627.*

Working Paper

2001-9: Effrosyni Diamantoudi: Stable Cartels Revisited.

2001-16: Bjarne Brendstrup, Svend Hylleberg, Morten Nielsen, Lars Skipper and Lars Stentoft: Seasonality in Economic Models.

2001-17: Martin Paldam: The Economic Freedom of Asian Tigers - an essay on controversy.

2001-18: Celso Brunetti and Peter Lildholt: Range-based covariance estimation with a view to foreign exchange rates.

2002-1: Peter Jensen, Michael Rosholm and Mette Verner: A Comparison of Different Estimators for Panel Data Sample Selection Models.

2002-2: Torben M. Andersen: International Integration and the Welfare State.

2002-3: Bo Sandemann Rasmussen: Credibility, Cost of Reneging and the Choice of Fixed Exchange Rate Regime.

2002-4: Bo William Hansen and Lars Mayland Nielsen: Can Nominal Wage and Price Rigidities Be Equivalent Propagation Mechanisms? The Case of Open Economies.

2002-5: Anna Christina D'Addio and Michael Rosholm: Left-Censoring in Duration Data: Theory and Applications.

2002-6: Morten Ørregaard Nielsen: Efficient Inference in Multivariate Fractionally Integration Models.

2002-7: Morten Ørregaard Nielsen: Optimal Residual Based Tests for Fractional Cointegration and Exchange Rate Dynamics.

2002-8: Morten Ørregaard Nielsen: Local Whittle Analysis of Stationary Fractional Cointegration.

2002-9: Effrosyni Diamantoudi and Licun Xue: Coalitions, Agreements and Efficiency.

2002-10: Effrosyni Diamantoudi and Eftichios S. Sartzetakis: International Environmental Agreements - The Role of Foresight.