International Environmental Agreements
–The Role of Foresight *

Effrosyni Diamantoudi †
Department of Economics, Concordia University

Eftichios S. Sartzetakis
Department of Economics, University of Macedonia

May 2001
(This version January 2014)

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 group of members. We assume that when a group of countries 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 the solution always exists and fully characterize the coalitionally farsighted stable IEAs. The new farsighted IEAs can be much larger than those some of the previous models supported and are always Pareto efficient.

Keywords: Environmental Agreements, Foresight, Stable Set *JEL*: D6, Q5, C7

*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.

†Corresponding author: Effrosyni Diamantoudi, Department of Economics, Concordia University, 1455 De Maisonneuve Blvd West, Montreal, Quebec, H3G 1M8 Canada. E-mail: *ediamant@alcor.concordia.ca*

1 Introduction

International environmental agreements (IEAs) aim at controlling global environmental problems, forming thus, a special case of the public good provision problem. Although the socially optimum outcome requires that all countries sign the agreement, each country has an incentive to free ride on the cooperating efforts of the rest of the signatories. This incentive stems from the fact that the costs avoided by not abating outweigh the marginal environmental damage caused on the country, when every other country agrees to control its emissions.

IEAs are a relatively new branch 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 go against the Coasian prediction (1960) that efficient outcomes will prevail. These pessimistic predictions are supported by empirical evidence. The continuing stalemate in the negotiations for a post-Kyoto agreement is an illustrative example of the difficulties faced in trying to achieve large, stable coalitions to resolve environmental problems entailing high costs.

The theoretical literature derives conflicting results depending on the deviating agent's expectations over the reaction of the remaining coalition. The non-cooperative approach supports the pessimistic view that IEAs will be signed by *very few* countries or, if signed by more, it will be over an agreement without any substance.[1] This result can be attributed to the assumption that each country upon withdrawing from the agreement assumes that the agreement will remain intact, at least in terms of membership status. The cooperative approach asserts the formation of the grand coalition and the attainment of efficiency, assuming that when a country deviates it expects that the agreement collapses and each country fends for itself.[2] It

[1] As we will show in Section 2, even if IEAs are signed by a significant number of countries in equilibrium, the agreed upon emission levels are identical to those of the laissez-faire state where each country optimizes individually.

[2] For an in depth comparison of the approaches see Tulkens (1998). Ioannidis et al.(2000) and Finus (2003) review the IEAs literature, whileYi (2003) offers a more general

is clear that 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 bridge the gap between the two approaches. When a group of countries defects from an agreement it does not make any assumption regarding the behavior of remaining members of the agreement. Instead, it foresees what their reaction will be, and which equilibrium agreement will result from such a deviation. In the literature on IEAs, farsightedness has been discussed and encouraged by Ecchia and Mariotti (1998), Carraro and Moriconi (1998) and further developed by Eyckmans (2001) and de Zeeuw (2008).

Following a different modeling approach, Chander (2007) also attempts to reconcile the two approaches by formalizing coalition formation as an infinitely repeated game. He shows that the grand coalition is an equilibrium outcome, because each player can credibly commit not to form a coalition unless it includes all players. That is, he is able to achieve the grand coalition, as in Chander and Tulkens (1992) and (1997), without the need of the "all-or-none expectation" employed by the Core concept. However, since the grand coalition is not the only equilibrium, the author uses Schelling's (1960) focal-point argument to select the grand coalition as the only equilibrium outcome.

Closer methodologically to the works of Ray and Vohra (1997), (1999) and (2001) we define equilibrium IEAs in a consistent manner. That is, each 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. Ray and Vohra's (1997) and (1999) works are general and do not concentrate on the public good provision problem, while 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, it focuses instead, on an equilibrium analysis.

Compared to the concept of equilibrium binding agreements by Ray and

survey on games of coalition formation.

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 main reason we are able to do so is because symmetric agents imply symmetric actions in our context. Hence, all signatories will emit the same level and enjoy the same welfare. Same holds for non-signatories. Thus, payoffs need not be indexed to each agent, instead, they can only be indexed to signatories and non-signatories. Moreover, each agent's payoff no longer depends explicitly on every other agent's action but simply on the size of the coalition. Symmetry allows us to study the monotonicity of the signatories' welfare based on the size of the IEA. Without symmetry we loose monotonicity as the constitution (individual identities) of the membership will affect emission levels. Once monotonicity is lost it becomes very difficult to identify stable coalitions. In the presence of heterogeneity the study of IEA size becomes vacuous. It is irrelevant to study how many agents are signing the IEA. The relevant question is *who* is signing it. Yi (2003) also notes that when the symmetry assumption is relaxed, the main difficulty is that "... it is no longer possible to identify a coalition by its size". Furthermore, Chander (2007) discusses the case of asymmetric agents and although the grand coalition is established as one of the equilibrium outcomes, its uniqueness cannot be shown, for the same reason discussed above.

The impact of symmetry (or its lack) is well demonstrated in Diamantoudi and Xue (2007). The authors built a solution concept akin to that of Ray and Vohra (1997) by relaxing the restrictions on permissible deviations. In doing so, Diamantoudi and Xue (2007) are able to establish positive results in terms of existence as well as Pareto efficiency of the outcomes when symmetry is assumed. In its absence, however, neither existence can be establish nor Pareto efficiency. The latter is shown to fail in a counter example.

Lastly, once a coalition consist of heterogenous agents with varying payoffs, the issue of transfer payments arises and can be vital to the stability of the coalition. Consider a simple example with two heterogenous countries. If the two countries were to engage in non-cooperative action, a certain level of global emissions would result and country A would be better off than country

B. If however, the two countries were to cooperate, by construction aggregate welfare would increase. It may very well be the case, however, that country A benefits while country B looses from cooperation. The latter would hamper all self-enforcing cooperative efforts unless country A were to make a transfer to country B, which is possible since total welfare increases. An extension of the analysis we offer in the present paper incorporating heterogeneity would have to address payoff distributions as well as membership within a coalition. The solution concept would have to identify a pair comprising a set of agents and a payoff vector for the set rather than simply the cardinality of the set, which is the approach we take in this work.

The literature on heterogeneity within the context of IEAs is very modest. Most of the papers, such as Botteon and Carraro (2001) and McGinty (2006), use simulations with small number of agents and the only analytical attempts use either a completely linear model (Kolstad (2010)) or a semi-linear one (Fuentes-Albero and Rubio (2010)). The results of the afore mentioned literature are mixed. Most of the papers find that introducing heterogeneous agents may increase the size of IEAs, however, this does not always lead to increases in the aggregate welfare. Unfortunately, heterogeneity renders generalized models untractable and no significant conclusions and results can be drawn.

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. Although this assumption seems rather strong it is very common in the IEA literature and 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 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.[3]

[3]Chander (2007) also allows renegotiation among agents.

Along the same vein, but within the context of cartel formation Thoron (1998) employs Coalition Proof Nash Equilibrium (CPNE) to characterize stable cartels. The author's goal is to examine the impact of coalitional (or group) action on stability. Moreover, the use of CPNE imposes a sense of credibility on potential deviations by examining a deviation's immunity to further deviations by any subset of the original (deviating) group. The recursiveness embedded in CPNE is very similar to that embedded in Ray and Vohra's (1997) Equilibrium Binding Agreements and it is a restriction we relax in our approach.

Our analysis starts in Section 2 with the generalization of the results in Diamantoudi and Sartzetakis (2006) that employs D' Aspremont et al.'s (1983) solution concept of coalitional stability. We adopt a leadership model a la Stackelberg where the IEA leads and the non-signatories follow. Our analysis can easily be applied to the simultaneous a la Cournot model.[4] In Section 3 the new solution concept is introduced to capture consistency and foresight while coordinated actions are allowed. 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:

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

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

[4] 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.

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 $E_{ns} = (n-s)e_{ns}$. It is easy to verify that the best response functions of the non-signatories have no slope outside the interval $[-1, 0]$. Thus, according to Amir (1996) our model admits a unique symmetric equilibrium in emission levels at the non-signatory stage. 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 $N \backslash S$'s behavior. Similarly, the coalition's maximization problem yields an indirect welfare function ω_s as follows:

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

The optimal signatories' emissions $e^*_s(s)$ satisfy the following condition:[5]

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

The following proposition generalizes the result of Diamantoudi and Sartzetakis (2006), establishing 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 cooperate

[5]Note that in the extreme cases where $s = 0$ the model reduces to a Cournot-type competition whereas if $s = n$ all countries 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. This critical size is determined by adjusting, to the lower integer, the value of $z^{\min} = \frac{B''(e_{nc}) - nD''(E_{nc})}{B''(e_{nc}) - D''(E_{nc})}$, where $z^{\min}$ denotes the intersection of ω_s with ω_{ns} that lies right at the minimum value of ω_s. Therefore, this critical coalition size is almost (due to integer adjustments) the worst, in terms of per member welfare level, when compared to other coalition sizes. Since this critical size is greater than one,[6] 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.[7]

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))*[8],
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. $\omega_s(s) \gtreqless \omega_{ns}(e_s^*(s), s) \Leftrightarrow s \lesseqgtr z^{\min}$.[9]

[6] Since $B''(e_{nc}) < 0$ and $D''(E_{nc}) > 0$ we get that $B''(e_{nc}) - D''(E_{nc}) < 0$. Having this in mind and that $n > 1$, the definition of $z^{\min}$ assures that and its value always exceeds unity.

[7] The proof of Proposition 1 is relegated to an Appendix.

[8] 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.

[9] 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 $\omega_s(s)$.

A direct implication of the above result is that $\omega_s(s) \gtreqless \omega_{ns}(0)$, $\forall s \in \{1, ..., n\}$. In other words, the pure non-cooperative outcome (a la Cournot) yields the lowest welfare to countries. Also, directly from the maximization problem stems that the grand coalition yields the highest payoff per signatory than any other size: $\omega_s(s) < \omega_s(n)$ for all $s = 1, ..., n-1$.

The next natural step is the determination of the size of the stable coalitions. In the IEAs literature 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 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 in terms of coalition participation once a country exits (or enters) the coalition, however, they do adjust their emissions as a response to the new size. In the case of quadratic benefit and damage functions, as shown by Diamantoudi and Sartzetakis (2006), the stable coalition is of size 2, 3 or 4.

3 Coordinated Action under Foresight

A natural extension of the analysis of IEAs thus far is to allow countries to coordinate their actions by permitting coalitional deviations in both directions. In other words, countries can coordinate their actions and exit jointly from an agreement, as well as 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 31st of May 2002. It is, indeed, often the case that countries join an environmental agreement in groups through international conventions and meetings. Allowing countries to coordinate their moves enhances their bargaining power in the sense that a group of countries can cause more damage by exiting an agreement than a single country can and hence it can pose a bigger threat.

Moreover, we endow countries with foresight. That is, we build a solution concept in the spirit of von Neumann and Morgenstern (vN-M) (1944) abstract stable set while amending the dominance relation to incorporate forward lookingness, which gives us a set of stable coalitions that would survive *credible* deviations.[10] 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.

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.[11] 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.[12] Although the notion of the stable set is very appealing exactly due to the afore mentioned properties that attribute consistency[13] it has been criticized on two grounds. Firstly, it does

[10]Diamantoudi and Xue (2007) have applied previously vN-M's stable sets as an equilibrium concept.

[11]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.

[12]This feature of the stable set is known as *External Stability*. The same problem with terminology arises here as well.

[13]Its appeal is captured and improved upon by Greenberg (1990). 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

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.

What we aim to accomplish in this section is to provide a notion that allows countries to move "coalitionally." Countries are endowed with foresight in that they anticipate the ultimate outcomes of their actions, and they do so collectively. Employing thus, von Neumann & Morgenstern's (1944) stable set and Harsanyi's (1974) indirect dominance we consider a "solution" set σ which is defined to be the collection of all *coalitionally farsighted stable* coalitions. Let C_s denote any coalition of size s, then, if $\sigma^* = \{C_s, C_t..., C_f\}$ is a coalitionally farsighted stable set it implies that each of its elements is coalitionally farsighted stable given σ^*.

The first step in building our (farsighted) solution concept is to define the direct dominance relation between any two IEAs. Consider two agreements C_q and C_s, where $q < s$. We argue that a larger agreement, C_s, directly dominates a smaller one, C_q, if a group of non-signatories would benefit by joining C_q and thus, inducing C_s. Similarly, we argue that a smaller agreement, C_q, directly dominates a larger agreement, C_s, if a group of signatories would benefit by withdrawing from the agreement and thus, inducing the smaller agreement C_q. Direct dominance is formalized in the definition below.

Definition 1 *Consider two IEAs C_q and C_s, $s \neq q$. C_q (directly) dominates C_s, i.e., $C_q \succcurlyeq C_s$ if either*

1. *$q > s$ and $\omega_s(q) > \omega_{ns}(s)$ or*

2. *$q < s$ and $\omega_{ns}(q) > \omega_s(s)$.*

Now suppose that two IEAs do not (directly) dominate one another even though there exists a group of countries that prefers one IEA to another.

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.

The reason that direct dominance fails is that the benefiting group cannot directly induce the preferred agreement. It may, however, be possible that this group can *indirectly* induce its preferred agreement by acting farsightedly. For instance, consider an agreement that is slightly smaller than the grand coalition and its signatories would benefit from any further enlargement of their IEA. It is obvious that the signatories of the present agreement cannot induce a larger one. Such an inducement would require that the non-signatories act. To make matters more difficult assume further that the non-signatories of the IEA under consideration do not prefer to be members of any larger agreement. Direct dominance fails in such a situation. It may, nonetheless, be possible that (some of) the signatories can act (by exiting the present agreement) and thus coerce the non-signatories in joining. In such an event we say that the grand coalition indirectly dominates the smaller agreement. Indirect dominance is formalized in the definition below.

Definition 2 *Consider two IEAs C_q and C_s. C_q indirectly dominates C_s, i.e., $C_q \gg C_s$ if $\exists C_{s^1}, ..., C_{s^m}$ such that*

1. *$s^1 = s$ and $s^m = q$,*
2. *$C_{s^m} \succcurlyeq C_{s^{m-1}}$ and*
3. *for any s^j, s^{j+1} where $j = 1, ..., m-2$ we have*
   - *(a) if $s^{j+1} > s^j$ then $\omega_{ns}(s^j) < \min\{\omega_s(s^m), \omega_{ns}(s^m)\}$ and*
   - *(b) if $s^{j+1} < s^j$ then $\omega_s(s^j) < \min\{\omega_s(s^m), \omega_{ns}(s^m)\}$.*

The intuition of the above definition is as follows: C_q indirectly dominates C_s if a sequence of IEAs leading from C_s to C_q can be constructed such that, at every step of the sequence, a group of countries can be identified that wishes to induce the subsequent IEA from the preceding one. The acting group of countries wishes to induce one agreement from another, not by myopically comparing its payoffs under the two scenarios, but, by foreseeing the ultimate outcome of its inducement, namely C_q.

To further explain our definition (especially part 3) it is important to recall that C_q is any coalition of size q. Therefore, when we argue that C_q directly or indirectly dominates C_s we basically say that for every *specific*

coalition of size s one can identify another specific coalition of size q that dominates it. We do *not* require that every coalition of size q dominates (directly or indirectly) every coalition of size s.

In the case of direct dominance, matters are simple and our binary relation is irreflexive in that it can never be the case that $C_s \succcurlyeq C_s$. In the case of indirect dominance, however, we must block reflexivity. To induce an agreement of size s from another agreement of the same size would require at least two steps. To see this point consider a very simple scenario with 6 countries, $\{a, b, c, d, e, f\}$, and concentrate on a specific agreement of size 4, $\{a, b, c, d\}$. Further assume that those outside the agreement under consideration are better off than those inside, i.e., $\omega_{ns}(4) > \omega_s(4)$. Although a and b -any two of a, b, c and d would do the job- would like to induce agreement $\{c, d, e, f\}$, it is impossible to do so in one step. Countries a and b have to exit first, induce agreement $\{c, d\}$ where all 6 countries are worse off and then *hope* that e and f will take the initiative to join in and induce $\{c, d, e, f\}$.[14] Put differently, since the inducement from C_s to C_q occurs in one step, either by a group of countries entering or by a group of countries exiting, there is no uncertainty, as far as the acting group is concerned, over which *specific* C_q will arise. The acting group need not *hope* that others along the way will have to cooperate.

In the case of indirect dominance, since more than one steps may take place, the ultimate IEA, although certain in terms of size, is uncertain in terms of composition. Let us return to the simple example constructed earlier, and observe that there are 14 agreements of size 4 besides the starting agreement $\{a, b, c, d\}$. Now suppose that a and b exit and induce $\{c, d\}$. The number of agreements of size 4 that can immediately follow is reduced to 6, some of them include a and b, some include one of the two and some none of the two. Once $\{c, d\}$ is the status quo any two of the four non-signatories, a, b, e and f may proceed to join in as all four of them prefer to be signatories of 4 than non-signatories of 2. i.e., $\omega_s(4) > \omega_{ns}(2)$. It is, however, too optimistic on behalf of a and b to hope that their most preferred outcome

[14] Recall that we allow countries to coordinate their action in one direction *or* another. We do not consider the situation where some countries will be exiting and some will be entering the agreement simultaneously. But even if we were to allow such an event the example would still be valid as countries e and f would not *switch* positions willingly with a and b, since such an action would make them worse off.

will prevail. In other words, it is a distinct possibility that $\{a, b, c, d\}$ may arise again, in which case a and b are as well off as they were in the very beginning. Our indirect dominance relation requires that if any group of agents acts along the sequence, it does so by anticipating with certainty a strictly positive increase in its payoffs in the end. Thus, if a group of countries acts at some stage of the sequence it compares its status quo to the worse possible outcome that may prevail at the end of the sequence, namely $\min\{\omega_s(s^m), \omega_{ns}(s^m)\}$.

A natural worry at this point would be that the introduction of conservative behavior in indirect dominance has gone a bit too far. What if the sequence of coalitions is monotonic in size (increasing or decreasing) and the countries that act at any point of the sequence are certain over their payoff at the end of the sequence? Suppose, for instance, that the sequence is increasing in size and countries only enter while nobody exits. In such an event the payoff of every country that enters along the sequence will end up being $\omega_s(s^m)$, and there is no uncertainty over it. Does our notion force these agents to make unduly conservative choices and perhaps forego a deviation fearing an outcome that can never happen? The answer is no. Consider a situation where a monotonic sequence of coalitions leads from C_{s^0} to C_{s^m} where the latter is larger.[15] All the acting agents are originally non-signatories and become signatories by joining in. Assume that their preferences are such that $\omega_{ns}(s^m) < \omega_{ns}(s^j) < \omega_s(s^m)$ for all j, hence such a sequence would *never* support $C_{s^m} \gg C_{s^0}$. Notice, however, that such a monotonic sequence can always be compacted into one step where the acting set of agents is the union of all the coalitions acting along the sequence. Then, indirect dominance reduces to direct dominance and it only suffices that $\omega_{ns}(s^0) < \omega_s(s^m)$ which holds and we can argue that $C_{s^m} \succcurlyeq C_{s^0}$.

Lastly, note that in Definition 1 we imposed irreflexivity by requiring $q \neq s$ whereas in Definition 2 we introduced conservatism that results in an irreflexive binary relation. In Definition 1 irreflexivity is a result of feasibility constraints: countries cannot directly induce one IEA from another while both have the same size. Although we could have imposed irreflexivity in Definition 3 in the same manner ($q \neq s$) we chose to make the formalization

[15] A similar argument can be developed if C_{s^m} is smaller.

more transparent (albeit more complicated) since irreflexivity is now a result of preferences and not mere feasibility.

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 another, while along the path of these coordinated moves every acting coalition prefers the final agreement to its status quo. Furthermore, an unstable IEA is accounted for in a consistent manner if 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.[16] Consistency is achieved because the "dominating" outcome -the very reason for deviating- is stable (credible) itself.

Definition 3 *A set of IEAs, σ^*, is a* ***coalitionally farsighted stable set*** *if*

1. *σ^* is free of inner contradictions: there do not exist $C_q, C_s \in \sigma^*$ such that $C_q \gg C_s$.*

2. *σ^* accounts for every IEA it excludes: for every $C_t \notin \sigma^*$ there exists $C_q \in \sigma^*$ such that $C_q \gg C_t$.*

Before we proceed with the characterization of our solution concept it is important to mention that there are other approaches to coalitional foresight in the literature such as Suzuki and Muto (2005), Nakanishi (2009) and Kamijo and Nakanishi (2007) among others. Depending on the context (underlying theoretical model) within which the formalizations occur, different dominance relations are adopted offering various advantages (and disadvantages), fitting to the economic environment in question.

[16] We would like to point out that our concept exhibits a degree of optimism in the sense that it suffices that there exists one sequence that leads to a better (stable) outcome. The existence of other sequences that lead to worse (stable) outcomes plays no inhibitive role. Chwe (1994) and Xue (1998) built solution concepts with a conservative aspect in that every sequence that leads to a (stable) outcome must be improving for the deviation to occur. Such notions are normally very large and have little predictive power.

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 Pareto efficient. The set of Pareto efficient agreement sizes is identified in the following remark. A thorough exposition and analysis of Pareto efficiency within the Prisoner's dilemma game can be found in Suzuki and Muto (2005).

Remark 1 *An agreement C_t is Pareto efficient if and only if $\omega_{ns}(t) > \omega_s(n)$. All other agreements C_s are Pareto dominated by the grand coalition since $\omega_{ns}(s) \leq \omega_s(n)$ and $\omega_s(s) < \omega_s(n)$. Note that any two agreements, C_t and $C_{t'}$ that are not Pareto dominated by the grand coalition do not Pareto dominate each other. Assume without loss of generality that $t > t'$. Then $\omega_s(t) > \omega_s(t')$ by Proposition 1*[17]*. So, does C_t Pareto dominate $C_{t'}$? If $\omega_{ns}(t) < \omega_{ns}(t')$ then the entire set of non-signatories in $C_{t'}$ is worse off in C_t. If, however, $\omega_{ns}(t) \geq \omega_{ns}(t')$ the comparison is more subtle: since $t > t'$, there are non-signatories under $C_{t'}$ that switch role and become signatories. This group is worse off under C_t since $\omega_{ns}(t') > \omega_s(n) > \omega_s(t')$.*

Let s^* denote the smallest myopically stable coalition that is larger that $z^{\min}$, as defined in Section 2. 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, $\hat{s} = \arg\max_{s \in \{0,\ldots,s^*\}} \omega_{ns}(s)$. The following result is expressed in term of agreement sizes. Thus, although σ^* may contain only one "size", it supports all agreements (permutations) of that size.

Theorem 2 *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 $\bar{s} < r < z^{\min}$.

[17]Note that all C_t such that $\omega_{ns}(t) > \omega_s(n)$ are located on the increasing part of $\omega_s(s)$ or, put differently, $t > z^{\min}$. If not, then $\omega_s(t) > \omega_{ns}(t) > \omega_s(n)$ and that would contradict the fact that C_n maximizes total welfare.

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. The solid line depicts the per member payoffs of the signatories, $\omega_s(s)$, while the dotted line depicts the per member payoffs of the non-signatories, $\omega_{ns}(s)$. When an agreement C_s is marked with a star it means that it 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 (circled) σ^* the single arrows denote all the unstable IEAs while the double arrows illustrate some of the representative paths that lead from an unstable agreement to a stable one.

1. Consider $\sigma = \{C_n\}$. Inner contradictions are obviously not present since σ^* contains only one element.

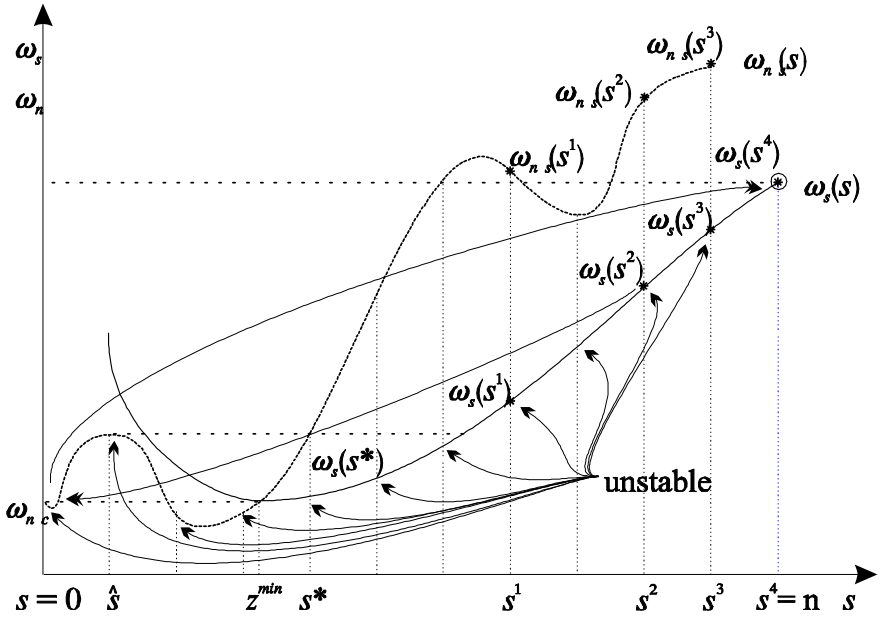

Figure 1

To see how every $C_s \notin \sigma^*$ is accounted for, note that it must be the case 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 any IEA C_s such that $\omega_{ns}(s) < \omega_s(n)$, since $n > s$ we have that $C_n \succ C_s$ (non-signatories join in) and any such IEA is coalitionally farsighted unstable. 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 such a case, $C_n \gg C_s$ via the sequence C_s, C_0, C_n: all the members of C_s exit, inducing the pure non-cooperative state, C_0, and then the grand coalition enters, forming C_n. In Figure 1 it is shown how $C_{s^4} = C_n$ accounts for the exclusion of C_{s^2}: an arrow leads from $\omega_s(s^2)$ to ω_{nc} and another arrow leads from ω_{nc} to $\omega_s(s^4)$.

2. When $\sigma^* = \{C_s\}$ where $s \neq n$, obviously there are no inner contradictions.

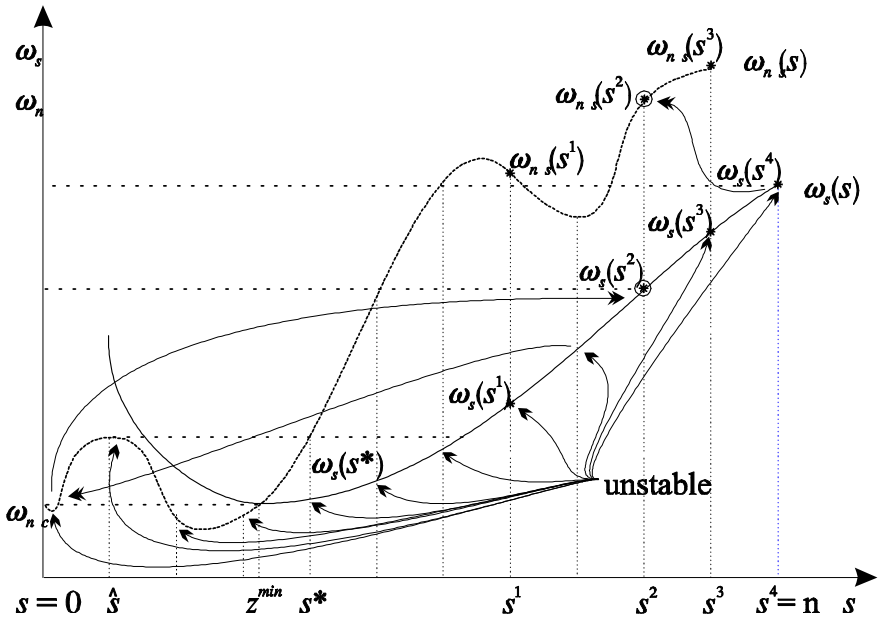

Figure 2

Note that since C_s is such that $\omega_{ns}(s) > \omega_s(n)$ it must be that $s > z^{\min}$. In Figure 2, s is depicted by s^2. 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, i.e., $C_s \succ C_r$. In Figure 2 such an inducement is captured by the movement from $\omega_s(s^4)$ to $\omega_{ns}(s^2)$. 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 non-signatories of C_r will join the agreement and directly induce C_s, again $C_s \succcurlyeq C_r$. In Figure 2 such an inducement is captured by the movement from ω_{nc} to $\omega_{ns}(s^2)$. All C_r such that $z^{\min} < r < s$ are coalitionally farsighted unstable since $\omega_{ns}(s) > \omega_s(s) > \omega_s(r)$ and all the members of C_r will exit the agreement and induce C_0 from C_r, then a coalition of size s will form and induce C_s. That is, $C_s \gg C_r$ via sequence C_r, C_0, 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 $C_{\bar{s}} = C_{\hat{s}}$.

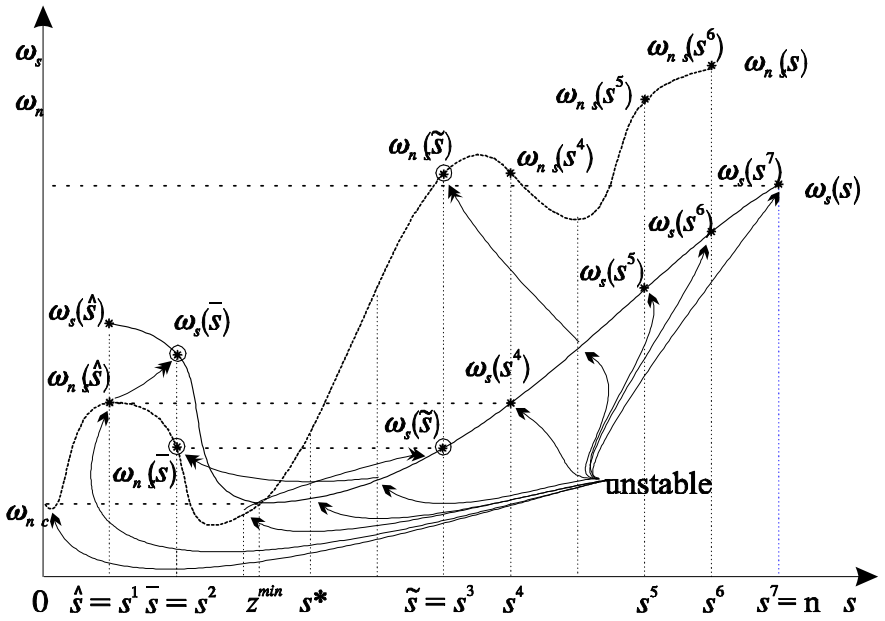

Figure 3

In Figure 3 $\tilde{s}$ is depicted by s^3 while $\bar{s}$ is depicted by s^2. 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 any sequence that would lead to $C_{\tilde{s}}$. No non-signatories of $C_{\bar{s}}$ wish to directly induce $C_{\tilde{s}}$ either since $\omega_{ns}(\bar{s}) = \min\{\omega_s(\tilde{s}), \omega_{ns}(\tilde{s})\}$. Similarly, no members of $C_{\tilde{s}}$ wish to exit $C_{\tilde{s}}$ 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 $s \geq \tilde{s}$, i.e., $C_{\tilde{s}} \succcurlyeq C_s$. 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 $\omega_s(s) < \omega_s(\tilde{s}) = \omega_{ns}(\bar{s})$, i.e., $C_{\bar{s}} \succcurlyeq C_s$. Any agreement C_s such that $\bar{s} < s < z^{\min}$ is unstable since the non-signatories can directly induce $C_{\tilde{s}}$ since $\omega_{ns}(s) < \omega_{ns}(\bar{s}) = \omega_s(\tilde{s})$, i.e., $C_{\tilde{s}} \succcurlyeq C_s$. Lastly, any agreement C_s such that $s < \bar{s}$ is unstable since its non-signatories will directly induce $C_{\bar{s}}$ since $\omega_{ns}(s) < \omega_{ns}(\hat{s}) < \omega_s(\bar{s})$, i.e., $C_{\bar{s}} \succcurlyeq C_s$. (Observe the double arrows in Figure 3).

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 non-signatories of $C_{\bar{s}}$ 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 non-signatories of $C_{\tilde{s}}$ are not directly inducible by either C_r or non-signatories of 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 signatories or non-signatories of 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 $\omega_{ns}(s') < \omega_s(\tilde{s})$.

■

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 that fifty five Parties to the Convention ratify the Protocol and in addition the ratifying countries account for at least 55% of the 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 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 coalitionally farsighted stable IEAs. We assume that when a group of countries contem-

plates 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 a solution concept allowing for coordinated actions that captures exactly such a decision-making process and we 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 our definition 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 one offered in this work would be very difficult to obtain.

5 References

1. Amir, R. (1996). "Cournot Oligopoly and the Theory of Supermodular Games." *Games and Economic Behavior* **15**, 132-148.

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

3. Botteon, M., Carraro, C. (2001). "Environmental coalitions with heterogeneous countries: Burden-sharing and carbon leakage." In: Ulph, A. (editor), *Environmental Policy, International Agreements, and International Trade*, Oxford University Press, Oxford.

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." In Raut,-Lakshmi-K. (editor), *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. (2007). "The gamma-core and Coalition Formation." *International Journal of Game Theory* **35**, 539-556.

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

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

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

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

11. de Zeeuw A. (2008). "Dynamic effects on the stability of international environmental agreements." *Journal of Environmental Economics and Management* **55**, 163-174.

12. Diamantoudi, E. and Xue, L. (2007). "Coalitions, agreements and efficiency." *Journal of Economic Theory* **136**: 105-125.

13. Diamantoudi, E. and Sartzetakis, E. (2006). "Stable International Environmental Agreements: An Analytical Approach." *Journal of Public Economic Theory **8**, 247–263.*

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

15. Eyckmans, J. (2001). "On the Farsighted Stability of the Kyoto Protocol." Working Paper Series, Faculty of Economics and Applied Economic Sciences, University of Leuven, No. 2001-03

16. Finus, M. (2003). "Stability and design of international environmental agreements: the case of transboundary pollution." In H. Folmer and T. Titetenberg (editors), *The International Yearbook of Environmental and Resource Economics 2003-04*, Edward Elgar, Cheltenham, pp 82-158.

17. Fuentes-Albero, C., Rubio, S.J. (2010) "Can international environmental cooperation be bought?" *European Journal of Operational Research* **202**, 255–264.

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

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

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

21. Kamijo, Y. Nakanishi, N. (2007). "Stability of Price Leadership Cartel with Endogenous Pricing." University of Kobe, working paper, 706.

22. Kolstad, C. D. (2010). "Equity, Heterogeneity and International Environmental Agreements." The B.E. Journal of Economic Analysis & Policy, **10** (2), Article 3.

23. 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.

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

25. McGinty, M., (2007). "International environmental agreements among asymmetric nations." *Oxford Economic Papers* **59**, 45–62.

26. Nakanishi, N. (2009). "Noncooperative farsighted stable set in an n-player prisoners' dilemma." *International Journal of Game Theory*, **38**, 249–261.

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

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

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

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

31. Schelling T.C. (1960). *The Strategy of Conflict.* Harvard University Press, Cambridge

32. Suzuki, A. and Muto, S. (2005). "Farsighted Stability in an n-Person Prisoner's Dilemma." *International Journal of Game Theory* **33**, 431-445.

33. Thoron, S. (1998). "Formation of a Coalition-Proof Stable Cartel." *Canadian Journal of Economics* **31**: 63-76.

34. 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.

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

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

37. Yi, S.-S. (2003). "Endogenous Formation of Economic Coalitions: A Survey of the Partition Function Approach." In: Carraro, C. (editor), *Endogenous Formation of Economic Coalitions*, E. Elgar, Cheltenham, UK, ch. 3, pp. 80-127.

6 Appendix 1: Proof of Proposition 1

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.

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

$$\frac{d\omega_s(z)}{dz} = (n-z)D'(E^*(z))\left[\frac{de_s^*(z)}{dz}\left.\frac{\partial e_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} - \frac{de_{ns}^*(z)}{dz}\right] - D'(E^*(z))[e_s^*(z) - e_{ns}^*(z)].$$

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, $\omega_s(s) \gtrless \omega_{ns}(s) \Leftrightarrow B(e_s^*(s)) \gtrless B(e_{ns}^*(s))$ and since $B' > 0$ we have $B(e_s^*(s)) \gtrless B(e_{ns}^*(s)) \Leftrightarrow e_s^*(s) \gtrless e_{ns}^*(s) \Leftrightarrow s \lessgtr z^{\min}$.

■