Self-enforcing International Environmental Agreements Revisited*

Santiago J. Rubio† and Alistair Ulph

University of Valencia, Spain

University of Southampton, UK

22 December 2004

*Research of Santiago J. Rubio was supported by the *Ministerio de Ciencia y Tecnología* (Spanish Ministry of Science and Technology) under grant BEC2000-1432 and by *Fundación BBVA*. We also acknowledge the financial support from the *Instituto Valenciano de Investigaciones Económicas.* Finally we want to thank Begoña Casino for her help with the figures and two anonymous referees for their useful comments. Regarding any remaining inadequacies, the usual caveat applies.

†Corresponding author. Department of Economic Analysis, University of Valencia, Edificio Departamental Oriental, Avda. de los Naranjos s/n, 46022 Valencia, Spain. E-mail address: Santiago.Rubio@uv.es

Abstract

In Barrett's (1994) paper on transboundary pollution abatement is shown that if the signatories of an international environmental agreement act in a Stackelberg fashion, then, depending on parameter values, a self-enforcing IEA can have any number of signatories between two and the grand coalition. Barrett obtains this result using *numerical* simulations in a pollution abatement model where he is not constraining emissions to be non-negative. Recent attempts to use analytical approaches and to explicitly recognize the non-negativity constraints have suggested that the number of signatories of a stable IEA may be very small. The way such papers have dealt with non-negativity constraints is to restrict parameter values to ensure interior solutions for emissions. We argue that a more appropriate approach is to use Kuhn-Tucker conditions to derive the equilibrium of the emissions game. When this is done we show, *analytically*, that the key results from Barrett's paper go through. Finally, we explain why his main conclusion is correct although his analysis can implicitly imply negative emissions.

Keywords: international externalities, self-enforcing international environmental agreements, Stackelberg equilibrium, non-negative emissions constraints

JEL classification:C72, D62, F02, Q20

1 Introduction

Over the last two decades, one of the factors driving an increased sense of interdependence between countries is the need to tackle global environmental problems such as climate change, ozone depletion, loss of biological diversity amongst others. Tackling such problems requires some form of agreement between countries, and the Framework Convention on Climate Change, the Montreal Protocol on Substances that Deplete the Ozone Layer, and the Convention on Biodiversity are important examples of such International Environmental Agreements (IEAs). However, the very different experience of these agreements illustrates the crucial importance of understanding how to design agreements which give countries incentives to both join and abide by such agreements. Economists have emphasized two important features: agreements must be profitable, that is there must be potential gains to all signatory countries; more importantly, in the absence of any international authority, agreements must be self-enforcing, i.e. there must be incentives for countries acting in their own self-interest to want to join or stay in an agreement.

One of the earliest definitions of a self-enforcing agreement was the concept of a stable IEA, which means that no individual signatory country has any incentive to leave the IEA, and no non-signatory country has an incentive to join, taking as given the membership decisions of all other countries.[1] Models based on this concept include Carraro and Siniscalco (1991,1993), Hoel (1992), Barrett (1994), Na and Shin (1998) amongst many others. Carraro and Siniscalco (1991) and Hoel (1992) have shown that if signatory countries act in Cournot fashion with respect to non-signatories, then a

[1]There are a number of other concepts of what makes an agreement self-enforcing. Chander and Tulkens (1995, 1997) draw on cooperative game concepts. See Tulkens (1998) for a systematic comparison of these two approaches. Other concepts, such as far-sightedness, have been developed, for instance, by Ecchia and Mariotti (1997,1998), Ray and Vohra (2001) and Diamantoudi and Sartzetakis (2002b); see Finus (2001) and Wagner (2001) for excellent overviews, and, more recently, Barrett (2003a) for a broad exposition of the strategy of environmental treaty-making.

stable IEA consists of 3 countries when marginal environmental damage is constant (i.e., when the countries' best-replay functions are orthogonal), and of 2 countries when marginal damage increases with emissions (i.e., when the best-replay functions have a negative slope), in both cases irrespective of the number of countries affected.[2] In the paper published by Barrett in 1994, that has had a great influence in the subsequent literature on IEA, it is shown that for a model of abatement with quadratic costs and benefits where the countries in the agreement act in a Stackelberg fashion, an IEA may achieve a high degree of cooperation, but only when the gains of cooperation are small. When the gains are large, a self-enforcing IEA cannot support a large number of signatories (see Prop. 1 on page 886).[3] The rationale for the difference in outcomes between Cournot and Stackelberg models is that if one country was to leave the IEA, with Cournot behaviour, the non-signatories expand their emissions and the remaining signatory countries partially accommodate this by reducing their emissions. On the other hand with Stackelberg behaviour, if a signatory was to leave the IEA the remaining signatories would expand their emissions. Thus the incentives to leave an IEA are greater with Cournot behaviour than with Stackelberg.

[2] Carraro and Siniscalco (1991,1993) also show that the number of signatory countries can be increased by means of self-financed transfers. However, expanding coalitions requires some form of commitment. Petrakis and Xepapadeas (1996) extend this result to the case in which the countries are not identical using an emissions game with orthogonal best-replay functions like the one studied by Hoel (1992). Hoel and Schneider (1997) point out that the prospect of receiving a transfer tends to reduce the incentive a country might have to commit itself to cooperation so that if the disincentive is strong, total emissions will be higher with side payments. More recently, Barrett (2001) has shown that with strong asymmetry side payments become the vehicle for increasing participation in a cooperative agreement.

[3] The results obtained so far in the literature as regards participation vary widely. The number of signatories depends on the specification of the net benefit function; for some specifications, it depends on the parameter values; and, for some specifications, it depends on the Cournot/Stackelberg assumption. For a more comprehensive and technical discussion of the importance of specification, see Finus (2001) and Barrett (2003, Table 7.1). In this paper we focus on Barrett's model (quadratic costs and benefits in combination with the Stackelberg assumption).

Since this model has become something of a workhorse tool to study IEAs, it is important that its properties are well understood. In this respect we think that the early analysis developed by Barret (1994) could be improved in two ways. First, it relied on numerical simulations to derive his Prop. 1.[4] Second, it assumed that aggregate emissions are large enough to avoid the maximal abatement constraint binding (see footnote 4 in Barrett's paper). However, it can be shown that this assumption on aggregate emissions does not guarantee non-negative emissions for each individual country.[5] Recent papers have attempted to generalize Barrett's numerical analysis to evaluate the robustness of his results. Finus (2001) presents an analytical generalization of Barrett's results and shows that the higher the level of environmental damages, the greater the size of the stable IEA. However, his proof assumes interior solutions, and it is easy to show that high environmental damages imply that, unconstrained, emissions will become negative. Diamantoudi and Sartzetakis (2002a) and Rubio and Casino (2001) also use analytical approaches, but recognize the need to ensure that emissions are non-negative.[6] They reach even more pessimistic conclusions - that even with Stackelberg behaviour

[4]To be fair to Barrett's (1994) paper it should be pointed out that three additional models with different specifications of the net benefit function are analytically solved in the paper, although these other models have received less attention in the subsequent literature.

[5]This assumption is a potential problem for the quadratic specification studied by Barrett (1994) but it is not necessarily a problem for other specifications as, for instance, the one studied by Carraro and Siniscalco (1991). These authors develop an analytical solution for a symmetric Cournot equilibrium. In their model they assume an environmental damage function that is quadratic with respect to the local emissions but linear with respect to the imported emissions, i.e., environmental damage depends on the product between local emissions and total emissions that affect the country. As a result of this specification the solution of the game always yields interior solutions. In this paper, we focus on a global environmental problem so that we assume that the environmental damage is a quadratic function of the aggregate emissions.

[6]A referee pointed out that sequestration of carbon dioxide could be viewed as a form of negative emissions. However this arises in the context of a stock pollutant, so that in one period it may be possible that reductions in the stock of greenhouse gases through carbon sequestration might exceed any positive flow of new emissions of greenhouse gases.

the number of signatories of a stable IEA will be small - no greater than four.

However the way Diamantoudi and Sartzetakis deal with the non-negativity constraint is to compute an interior solution and then restrict parameter values to ensure that the resulting emissions are always strictly positive, see their Prop. 1. It is not surprising that this restriction on parameters restricts the number of signatories in a stable IEA. Rubio and Casino go further and restrict parameters to ensure that payoffs are non-negative, which is difficult to justify. We argue that neither of these approaches is appropriate. In this paper we also use an analytical approach and deal with the non-negativity constraint by simply imposing it directly on the choice of emissions by both signatory and non-signatory countries and using Kuhn-Tucker conditions to derive the equilibrium of the game. Then for some parameter values, the emission game will result in corner solutions.

Our findings for the Stackelberg equilibrium of the emissions game show that the number of countries in a stable IEA is directly related to the level of marginal environmental damage so that when marginal environmental damage is low enough a stable IEA consists of at most 3 countries while when marginal environmental damage is large enough the unique stable IEA is the grand coalition.[7] The rationale for this kind of relationship is given by the fact that the interdependence among the countries occurs through the environmental damage function. Thus when the marginal environmental damage is relatively high, the countries in the agreement choose emission levels which induce the non-signatories to select low emissions, making exit from the agreement unprofitable. With lower marginal environmental damage cost, these effects are weakened, so that some countries find it profitable to leave the agreement, i.e. the free-riding cannot be avoided by Stackelberg leadership. Finally, we clarify the previous results in the

[7] We have also shown analytically that when the environmental damage is a quadratic function of the aggregate emissions and the non-negativity constraints are taken into account, the previous results obtained by Carraro and Siniscalco (1991) for the Cournot equilibrium are unaffected. See Rubio and Ulph's (2002a) working paper.

literature which have been derived assuming interior solutions. According to our results restricting parameter values to guarantee interior solutions is a *sufficient* condition to get stable IEAs with a small number of signatories but it is not a *necessary* condition. In this paper we show that a stable IEA with a small number of countries can involve a corner solution. In other words, our results establish that what is really a necessary and sufficient condition to get stable IEAs with a small number of signatories is that the marginal environmental damage is low.

Thus we have shown in this paper that the results derived for the model of stable IEAs in paper by Barrett (1994), which used *numerical* calculations on the linear-quadratic version of the model and ignored the issue of non-negative emissions, carry through when derived *analytically* in a model which takes seriously the need to ensure that emissions are non-negative. The reason why taking account of non-negative emissions does not change the main results of the literature is that, as we shall show, the definition of a stable agreement depends on the sign of the difference between payoffs to signatories and non-signatories as the number of signatories varies. Taking account of the need for emissions to be non-negative obviously changes the value of these payoffs, but not the sign of differences in payoffs. Thus this paper not only derives analytically results for the Barrett (1994) model using an appropriate treatment of non-negative emission constraints, but disproves the claim that taking account of such constraints makes a significant difference to known results.

Finally, we would like to clarify three issues. Firstly, we want to recognize explicitly that our results have been obtained assuming that all countries are identical as in Barrett's (1994) paper. This assumption, although restrictive, allows us to get an analytical solution of the game and thus to advance the analysis of the stability of IEAs.[8] Secondly, we want to highlight that it is important, at least for one case, to extend the analysis in order to consider the possibility of zero emissions not only from a mathematical point

[8] Although countries are symmetric, the equilibrium is asymmetric with signatories and non-signatories polluting in different amounts.

of view but also from an economic point of view. We are thinking about the case of *non essential emissions.* In a *partial equilibrium analysis,* like the one we develop in this paper, emissions are non essential when the marginal benefit from emissions is positive but finite for zero emissions. For this type of emissions if the marginal environmental damage is large enough it makes sense to consider the possibility that emissions can be completely eliminated. This means that there exists a substitute so that if the damage is big enough, emissions of a particular pollutant can be zero. This is the case studied in this paper where we assume that the benefit from emissions is given by a quadratic-linear function so that the marginal benefit is positive for zero emissions but finite. Finally, following the approach adopted by Carraro and Siniscalco (1993), Barrett (1994) and Chander and Tulkens (1997) among others we focus our analysis on the case where only *one* IEA is formed and the only question remaining is the size of the self-enforcing agreement. This approach is justified in this paper because our aim is to review Barrett's (1994) numerical analysis in order to evaluate analytically the robustness of his results when the non-negativity constraints are taken into account. Nevertheless, this assumption may be reasonable for global environmental problems, such as the climate change problem, for which IEAs are usually unique and launched by the United Nations. In the last analysis, it could be interpreted as an institutional constraint.

In section 2 we present the basic model of an international emissions game and introduce the definition of a stable international environmental agreement. In section 3 we derive the stability results for Stackelberg behaviour. Section 4 concludes.[9]

[9]The proofs of all Propositions, Corollaries and Lemmas, other than Proposition 3, are provided in the Technical Appendix.

2 An International Emissions Game

2.1 The Basic Model

In this section we present the basic linear-quadratic model of an international emissions game.[10] There are N identical countries, $i = 1, ..., N$. We define $q_i \geq 0$ as the level of emissions generated by country i, $Q_i \equiv \sum_{k \neq i} q_k$ the total emissions generated by all countries other than i, and $Q = \sum_k q_k = Q_i + q_i$ as the total emissions generated by all N countries. Each country derives a gross benefit from its emissions and also suffers environmental damage which depends on the global level of emissions. Then each country has a net benefit(payoff) function:

$$\pi(q_i, Q_i) \equiv \alpha q_i - \frac{\beta}{2} q_i^2 - \frac{\gamma}{2}(q_i + Q_i)^2.$$

We assume that $\alpha > 0, \beta > 0$ and $\gamma > 0$. It should be clear that w.l.o.g. we can normalize one of the parameters and we choose to normalize by setting $\gamma = 1$. To emphasize this normalization we rewrite the net benefit function as:

$$\pi(q_i, Q_i) \equiv a q_i - \frac{b}{2} q_i^2 - \frac{1}{2}(q_i + Q_i)^2. \tag{1}$$

We shall think of b as β/γ - the ratio of the (absolute) slope of the marginal benefit curve and the slope of the marginal damage cost curve, so a low value of b is to be interpreted as a (relatively) high marginal damage cost.[11]

[10] In this model, as in the early papers discussed in the introduction, we deal with a flow pollutant. For analysis using a stock pollutant see Rubio and Ulph (2002b, 2003).

[11] Notice that while we work with an emission game, it can be shown that this is equivalent to the model of an abatement game developed by Barrett. The only thing that is necessary to go from the emission game to the abatement game or vice versa is to properly define the level of unabated emissions and to propose a variable change. In our linear-quadratic emissions game this level is given by the emissions that maximize the gross benefit, a/b, then the abatement is defined as the difference between this level of emissions and the current emissions $x_i = a/b - q_i$, so that the net benefits can be written in terms of the abatement just by substitution of $q_i = a/b - x_i$ in (1), where x_i stands for the abatement.

2.2 Stable International Environmental Agreements

We model an International Environmental Agreement as a **three**-stage model, in which in the first stage (the *Membership Game*) each country decides whether or not to join an IEA, in the second stage the signatories determine their emissions, and in the third stage the non-signatories determine their emissions. Thus, the countries first play a simultaneous game, the Membership Game, and given the outcome of this game they play a sequential game that we call the Emissions Game. We describe each game briefly, in reverse order.

<u>The Emissions Game</u>

Suppose that, as the outcome of the first-stage game, there are n signatory countries (a typical signatory being denoted by s) and $N-n$ non-signatory countries (a typical non-signatory being denoted by f, for fringe or free-rider). Each non-signatory country takes as given the emissions of all other countries and chooses its emissions to maximize its own net benefits. Using symmetry, this will define a *non-signatory reaction function* linking the emissions of a typical non-signatory country to the emissions of a typical signatory country. Signatory countries choose emissions to maximize the aggregate payoff of the n signatories, recognising how this will affect emissions of non-signatories.

Two issues need further clarification. First, there is the question of the timing, or really commitment, of the emissions of signatories relative to non-signatories. Following Barrett (1994) we argue that membership of an IEA acts as a form of commitment device, which we model by thinking of signatories setting their emissions before non-signatories. In that case signatories can calculate what emissions non-signatories will choose (in terms of the non-signatory reaction function), and choose their own emissions to maximize aggregate net benefit. This is the *Stackelberg* model.[12]

The resulting abatement game is equivalent to the emission game and yields the same results. See Diamantoudi and Sartzetakis (2002a) for an application of this procedure.

[12]We would like to explicitly acknowledge that this notion of equilibrium for the emissions game is the Stackelberg version of the "Partial Agreement Nash Equilibrium (PANE) with respect to a coalition"

The second issue is that emissions by signatories and non-signatories have to be non-negative and we just impose these as constraints on the decision problems of signatories and non-signatories. So for some parameter values the emissions of either a signatory or non-signatory could be zero (a corner solution).

The outcome of the emission game then, is that, for any number of signatories n we can define the equilibrium payoffs to signatory and non-signatory countries: $\pi^s(n), \pi^f(n)$.

The Membership Game

We follow Hoel (1992), Carraro and Siniscalco (1993), Barrett (1994) and others in saying that an IEA is self-enforcing if it is stable, where the concept of stability is borrowed from the literature on cartel stability (d'Aspremont *et al* (1983)). For $2 \leq n \leq N$ we define $\Delta(n) = \pi^s(n) - \pi^f(n-1)$; then:

Definition 1 *An IEA with n signatories is* ***stable*** *if it satisfies the conditions:* ***Internal Stability****:* $\Delta(n) \geq 0$, *i.e.* $\pi^s(n) \geq \pi^f(n-1)$; ***External Stability****:* $\Delta(n+1) \leq 0$, *i.e.*$\pi^f(n) \geq \pi^s(n+1)$.

Internal stability simply means that any signatory country is at least as well off staying in the IEA as quitting, assuming that all other countries do not change their membership decisions. External stability similarly requires that any non-signatory is at least as well off remaining a non-signatory than joining the IEA, again assuming that all other countries do not change their membership decisions.

We can also think of a stable IEA as a Nash equilibrium of a *simultaneous open membership game* where the strategies for each country are to sign or not sign.[13] A country takes as given the membership decisions of all other countries. Suppose these have resulted in a membership of m, $0 \leq m \leq N-1$. Then the payoffs to a country are $\pi^s(m+1)$ if it signs and $\pi^f(m)$ if it does not. So it will join if $\pi^s(m+1) \geq \pi^f(m)$ and not join otherwise. For an IEA with n^* members to constitute a Nash equilibrium of the

defined by Chander and Tulkens (1997)

[13] In open membership games, any player is free to join or leave a coalition. In our case each country chooses one of the two possible strategies and the agreement is formed by all players who choose to sign.

Membership Game, it must have paid each signatory to sign, so $\pi^s(n^*) \geq \pi^f(n^*-1)$. Similarly it must have paid each non-signatory not to join, so $\pi^f(n^*) \geq \pi^s(n^*+1)$. These are just the conditions for Internal and External Stability.

In this next section we analyze stable IEAs for the Stackelberg model with non-negative emissions.

3 Stable Stackelberg IEAs with Non-Negative Emissions

In this section we analyze stable IEAs when signatory countries act collectively as a Stackelberg leader and emissions are restricted to be non-negative. We begin with the emissions game.

3.1 Stackelberg Emissions Game with Non-Negative Emissions

Suppose there are n signatories and $N-n$ non-signatories. A non-signatory country k takes as given Q_{fk} and chooses q_{fk} to solve:

$$\max_{q_{fk}\geq 0} \ \pi_k^f = aq_{fk} - \frac{b}{2}q_{fk}^2 - \frac{1}{2}(q_{fk}+Q_{fk})^2,$$

where q_{fk} stands for the level of emissions generated by non-signatory k, $k=1,...,N-n$, and Q_{fk} for the total emissions generated by all countries other than non-signatory k.

The first order condition is:

$$\frac{\partial \pi_k^f}{\partial q_{fk}} = a - bq_{fk} - (q_{fk}+Q_{fk}) \leq 0, \ q_{fk} \geq 0, \ \ q_{fk}\frac{\partial \pi_k^f}{\partial q_{fk}} = 0. \qquad (2)$$

(2) defines the *non-signatory reaction function* for a country k allowing for the fact that emissions must be non-negative, so part of the reaction function has $q_{fk}=0$. Now

signatories are assumed to coordinate in order to maximize their collective net benefits taking into account the reaction function of the followers:

$$\max_{q_{s1},\ldots,q_{sn},q_{f1},\ldots,q_{f,N-n}\geq 0} \Pi^s = \sum_{i=1}^{n} \pi_i^s = \sum_{i=1}^{n} [aq_{si} - \frac{b}{2}q_{si}^2 - \frac{1}{2}(q_{si} + Q_{si})^2],$$

$$s.t. \quad -a + bq_{fk} + (q_{fk} + Q_{fk}) \geq 0,$$

where q_{si} stands for the level of emissions generated by signatory i, $i = 1, ..., n$, and Q_{si} for the total emissions generated by all countries other than signatory i.

Under the assumption of symmetry we have that $q_{f1} = ... = q_{f(N-n)} = q_f$, $q_{s1} = ... = q_{sn} = q_s$ and $Q = nq_s + (N-n)q_f$, so that the previous optimization problem reduces to:

$$\max_{q_s \geq 0,\ q_f \geq 0} n\pi_s = n[aq_s - \frac{b}{2}q_s^2 - \frac{1}{2}(nq_s + (N-n)q_f)^2]$$

$$s.t. \qquad -a + bq_f + nq_s + (N-n)q_f \geq 0, \tag{3}$$

The Lagrange function for the problem is

$$\begin{aligned} L &= n[aq_s - \frac{b}{2}q_s^2 - \frac{1}{2}(nq_s + (N-n)q_f)^2] \\ &\quad +\lambda\left(-a + bq_f + nq_s + (N-n)q_f\right), \end{aligned}$$

and the KTCs are

$$\begin{aligned} \frac{\partial L}{\partial q_s} &= n[a - bq_s - n(nq_s + (N-n)q_f) + \lambda] \leq 0, && (4) \\ q_s &\geq 0, \quad q_s\frac{\partial L}{\partial q_s} = 0, \\ \frac{\partial L}{\partial q_f} &= -n(N-n)(nq_s + (N-n)q_f) + \lambda(b + N - n) \leq 0, && (5) \\ q_f &\geq 0, \quad q_f\frac{\partial L}{\partial q_f} = 0, \\ \frac{\partial L}{\partial \lambda} &= -a + bq_f + nq_s + (N-n)q_f \geq 0, && (6) \\ \lambda &\geq 0, \quad \lambda\frac{\partial L}{\partial \lambda} = 0. \end{aligned}$$

Equilibrium in the emissions game involves solving (4)-(6) simultaneously, taking account of non-negativity constraints. In principle there are three possibilities:

(i) **Interior Solution** ($q_s > 0, q_f > 0$). From (4)-(6) we obtain:

$$q_s = \frac{a\left[b^2 - (N-n)(n-2)b + (N-n)^2\right]}{b[(b+N-n)^2 + bn^2]}, \tag{7}$$
$$q_f = \frac{a\left[b^2 + (N+n^2-2n)b - (N-n)n\right]}{b[(b+N-n)^2 + bn^2]}, \tag{8}$$

so that $q_s > 0, q_f > 0$ iff

$$g(b,n) = b^2 - (N-n)(n-2)b + (N-n)^2 > 0,$$
$$h(b,n) = b^2 + (N+n^2-2n)b - (N-n)n > 0.$$

(ii) **Signatory Corner Solution** ($q_s = 0, q_f > 0$). From (4)-(6) this requires:

$$q_f = \frac{a}{b+N-n}, \quad g(b,n) \leq 0. \tag{9}$$

(iii) **Non-Signatory Corner Solution** ($q_s > 0, q_f = 0$). From (4)-(6) this requires:

$$q_s = \frac{a}{n}, \quad h(b,n) \leq 0. \tag{10}$$

It is easy to show that for the Stackelberg equilibrium a solution $q_s = q_f = 0$ does not satisfy the KTCs. Given the emissions for each kind of solution the net benefits can be obtained by substitution.

(i) **Interior Solution**:

$$\pi^s(n) = \frac{a^2}{2b}\left\{1 - \frac{N^2 b}{(b+N-n)^2 + bn^2}\right\}, \tag{11}$$
$$\pi^f(n) = \frac{a^2}{2b}\left\{1 - \frac{(b+1)N^2(b+N-n)^2}{[(b+N-n)^2 + bn^2]^2}\right\}. \tag{12}$$

(ii) **Signatory Corner Solution**:

$$\pi^s(n) = -\frac{a^2(N-n)^2}{2(b+N-n)^2}, \quad \pi^f(n) = \frac{a^2[b-(N-n)(N-n-2)]}{2(b+N-n)^2}. \tag{13}$$

(iii) **Non-Signatory Corner Solution**:

$$\pi^s(n) = -\frac{a^2(b+n(n-2))}{2n^2}, \quad \pi^f(n) = -\frac{a^2}{2}. \tag{14}$$

Now note that for the interior solution the full-cooperative level of emissions is given by Eq. (7) for $n = N$ and that the full-noncooperative Cournot level of emissions is given by Eq. (8) for $n = 0$. In these two cases we have an interior solution, for this reason we focus in the rest of this Section on $n = 1, 2, ..., N-1$.

We now want to determine more precisely for which parameter values the three different solutions occur. This analysis is developed in the following pages and the final results are presented in Fig. 3. The results clearly depend on the signs of $g(b,n)$ and $h(b,n)$. In that Fig. $b_1(n)$ is the function implicitly defined by $h(b,n) = 0$ and $b_2(n)$ and $b_3(n)$ are the functions implicitly defined by $g(b,n) = 0$ so that these functions allows us to determine for which values (b,n), $g(b,n)$ and $h(b,n)$ yield negative values or zero and then to know when the solution of the problem is a corner solution.[14] The next proposition fixes n and considers for which values of b we get each of the three solutions.

Proposition 1 *For any n, there exists a unique solution to the Stackelberg emissions game with non-negative emissions as follows: (i) for $n \in \{1,2,3\}$, there exists $b_1(n)$ defined above such that for $b \leq b_1(n)$ the equilibrium is the non-signatory corner solution while for $b > b_1(n)$ the equilibrium is the interior solution; (ii) For $n = 4$, we have that $g(N-4, 4) = 0$ and that $b_1(4) < N-4$ so that: (a) for $b \leq b_1(4)$ the equilibrium is the non-signatory corner solution, (b) for $b_1(4) < b < N-4$ the solution is the interior solution, (c) for $b = N-4$ the equilibrium is the signatory corner solution, (d) for $b > N-4$ the equilibrium is again the interior solution; (iii) For $\mathbf{n} \in (\mathbf{4}, \mathbf{N})$, we have that $b_1(n) < b_2(n) < b_3(n)$ so that : (a) for $b \leq b_1(n)$ the equilibrium is the non-signatories corner solution, (b) for $\mathbf{b} \in (\mathbf{b_1(n)}, \mathbf{b_2(n)})$ the equilibrium is the interior solution, (c) for $\mathbf{b} \in [\mathbf{b_2(n)}, \mathbf{b_3(n)}]$ the equilibrium is the signatories corner solution, and (d) for $b > b_3(n)$ the solution is the interior solution.*

This Proposition determines, for a given value of n, the critical values of b that yield

[14]Notice that $g(b,n) = 0$ defines two functions because for a given n the equation has two positive real solutions: $b_2(n) < b_3(n)$. However, $h(b,n) = 0$ has only a positive real solution that defines $b_1(n)$.

one type of solution or another. For instance, we have shown that if b is big enough the Stackelberg equilibrium is an interior solution for any value of n. Notice that the three kind of solutions that the problem can present can be explained by resort to the marginal conditions defined by the f.o.cs. of the problem. Next, we explain the logic behind the non-signatories corner solutions since as will be seen in Subsection 3.2 this kind of solution supports high levels of cooperation. To do that first we rewrite the optimization problem of signatories using the reaction function of non-signatories as an equality

$$\max_{q_s \geq 0} n\pi_s = n\left[aq_s - \frac{b}{2}q_s^2 - \frac{1}{2}\left(\frac{(N-n)a + nbq_s}{b+N-n}\right)^2\right].$$

For this problem the f.o.c. can be written as

$$a - bq_s = \frac{nb(N-n)a}{(b+N-n)^2} + \frac{n^2b^2}{(b+N-n)^2}q_s,$$

where the l.h.s. stands for the marginal benefit of emissions and the r.h.s for the marginal environmental damage. Now we can use this condition to explain why when b is enough low the solution is that non-signatories cut their emissions completely. Looking first at the limit case of $b = 0$ we see that the problem becomes linear and the solution would be $q_s = +\infty$ and $q_f = -\infty$ as (7) and (8) establish. Thus, when b is enough low the optimization problem is "quasi-linear" and both the marginal benefit curve and the marginal environmental damage curve are very "flat". Under these conditions the marginal benefit of emissions decreases slowly and the marginal environmental damage also increases slowly resulting in large emissions for the signatories. Then as for $b = 0$ the emissions tend to infinite there exists a critical value for b given by $b_1(n)$ such that if $b \leq b_1(n)$ the signatories' emissions are large enough to induce the non-signatories to choose zero emissions. At this point the reader should take into account that the signatories enjoy a strategic advantage and that emissions are strategic substitutes which explains why zero emissions can be the optimal policy for the non-signatories if signatories' emissions are large enough.

In order to complete the analysis of the parameter space we now characterize functions $b_1(n)$, $b_2(n)$ and $b_3(n)$ which allow us to determine the shape of the contours $g(b,n) = h(b,n) = 0$ of the functions $g(b,n)$ and $h(b,n)$. For $b_1(n)$ we have that $b_1(0) = b_1(N) = 0$ and it is easy to show that the function presents a unique extreme in the interior of the interval $(1, N-1)$ which is a maximum.[15] Then there exists, at least, one integer, $\hat{n}$, in that interval that maximizes $b_1(n)$ so that $b_1(\hat{n})$ is the maximum value of the function given by an integer in domain $\{1, 2, ..., N-1\}$.[16] Given this behaviour of function $b_1(n)$ we are able to establish for a given b what is the solution to the emissions game for non-signatories for different values of n. For $b \leq b_1(\hat{n})$, define $\mathbf{n}_1 \leq \mathbf{n}_2$ as positive real roots of $b = b_1(n)$ then we find that:

Lemma 1 *(i) If $b > b_1(\hat{n})$ we have that for all n the equilibrium will be an interior solution; (ii) If $b \leq \hat{b}_1(\hat{n})$, there exist n_1, n_2 defined above such that the equilibrium will be an interior solution for n such that $\mathbf{n} \notin [\mathbf{n}_1, \mathbf{n}_2]$ and a corner solution for n when $\mathbf{n} \in [\mathbf{n}_1, \mathbf{n}_2]$.*

Moreover, as $b_1(N-1)$ is the minimum value of $b_1(n)$ for $n = 1, 2, ..., N-1$ we can also establish that:

Corollary 1 *If $\mathbf{b} \in (\mathbf{0}, \mathbf{b}_1(\mathbf{N}-\mathbf{1})]$, $\mathbf{n}_1 < \mathbf{1}$ and $\mathbf{n}_2 \geq \mathbf{N}-\mathbf{1}$ so the only interior solutions are $n = 0$ and $n = N$. In other words, the equilibrium is a corner solution for non-signatories for all n.*

[15] In order to study the behaviour of $b_1(n)$ in that interval we assume that n is a real number and once we know the properties of $b_1(n)$ then we are able to characterize the values of $b_1(n)$ with respect to n but now with n restricted to be an integer number. The same approach is followed to study $b_2(n)$ and $b_3(n)$.

[16] Function $b_1(n)$ has a unique maximum, $\tilde{n}$,in the real interval $(1, N-1)$. If the maximum is an integer the function has the same maximum in the domain $\{1, 2, ..., N-1\}$ that we call $\hat{n}$. If $\tilde{n}$ is not an integer and the function is not symmetric then $\hat{n}$ is the closest integer to $\tilde{n}$ that yields the maximum value for $b_1(n)$. If the function is symmetric we could have two maximums $\hat{n}_1$ and $\hat{n}_2$ such that $\hat{n}_2 = \hat{n}_1 + 1$ provided that $\tilde{n}$ is in the center of interval $(\hat{n}_1, \hat{n}_2)$.

We illustrate these results in Figure 1.[17]

⇒ FIGURE 1 ⇐

The kind of solution for *signatories* emissions depend on whether b belongs to the interval $[b_2(n), b_3(n)]$. It can be shown that $b_2(n)$ is a strictly convex, decreasing function defined in the interval $[4, N]$ with $b_2(4) = N-4$ and $b_2(N) = 0$. Then for $n = 4, 5, ..., N-1$ the maximum value of the function is $N - 4$ and the minimum value is $b_2(N-1)$.

On the other hand, $b_3(n)$ is a strictly concave function with a maximum value equal to $N(N-4)/4$ in interval $(4, N-1)$ with $b_3(4) = N-4$ and $b_3(N) = 0$. Then there exists, at least, one integer, n^*, that maximizes $b_3(n)$ so that $b_1(n^*)$ is the maximum value of the function given by an integer in domain $\{4, 5, ..., N-1\}$ and therefore the function increases for $4 \leq n < n^*$ and decreases afterwards.[18] See Fig. 2. For this behaviour of functions $b_2(n)$ and $b_3(n)$ we are able of establishing for a given b what is the solution to emissions game for signatories for different values of n. For $\mathbf{b} \in [\mathbf{b_2(N-1), N-4}]$, define $\mathbf{n}_3$ as the unique, positive real root of $b = b_2(n)$ and $\mathbf{n}_4 > \mathbf{n}_3$ as the unique positive real root of $b = b_3(n)$. For $\mathbf{b} \in (\mathbf{N-4, b_3(n^*)}]$, define $\mathbf{n}_3 \leq \mathbf{n}_4$ as the positive real roots of $b = b_3(n)$. Then we have:

Lemma 2 *(i) If $b > b_3(n^*)$ we have that for all n the equilibrium will be an interior solution;(ii) If $\mathbf{b} \in [\mathbf{b_2(N-1), b_3(n^*)}]$, there exist $\mathbf{n}_3, \mathbf{n}_4$ defined above depending on whether b is greater or less than $N-4$ such that the equilibrium will be an interior solution for n when $n \notin [\mathbf{n}_3, \mathbf{n}_4]$ and a signatory corner solution for n when $n \in [\mathbf{n}_3, \mathbf{n}_4]$;(iii) If $\mathbf{b} \in (\mathbf{0, b_2(N-1)})$, we have that for all n the equilibrium will be an interior solution.*

[17] In order to simplify the graphical representation we assume that the maximum value of real function $b_1(n)$ is the integer number $\hat{n}$.

[18] Function $b_3(n)$ has a unique maximum, $\tilde{n}$, in the real interval $(4, N-1)$ for $N > 5$. If the maximum is an integer the function has the same maximum in the domain $\{4, ..., N-1\}$ that we call n^*. If $\tilde{n}$ is not an integer and the function is not symmetric then n^* is the closest integer to $\tilde{n}$ that yields the maximum value for $b_3(n)$. If the function is symmetric we could have two maximums n_1^* and n_2^* such that $n_2^* = n_1^* + 1$ provided that $\tilde{n}$ is in the center of interval (n_1^*, n_2^*).

We illustrate this Lemma in Figure 2.[19]

$\Rightarrow$ FIGURE 2 $\Leftarrow$

Now we can completely characterize the parameter space. The next Proposition summarizes Lemmas 1 and 2.

Proposition 2 *(i) If* $b > b_3(n^*)$ *we have that for all* n, *the equilibrium will be an interior solution for signatories and non-signatories; (ii) If* $\mathbf{b} \in (\mathbf{b}_1(\hat{\mathbf{n}}), \mathbf{b}_3(\mathbf{n}^*)]$, *the equilibrium will be an interior solution for non-signatories for all* n, *however for signatories there exist* $\mathbf{n}_3, \mathbf{n}_4$ *defined above such that the equilibrium will be an interior solution for* n *when* $\mathbf{n} \notin [\mathbf{n}_3, \mathbf{n}_4]$ *and a corner solution when* $\mathbf{n} \in [\mathbf{n}_3, \mathbf{n}_4]$;*(iii) If* $\mathbf{b} \in [\mathbf{b}_2(\mathbf{N}-\mathbf{1}), \mathbf{b}_1(\hat{\mathbf{n}})]$, *there exist* $\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3$ *and* $\mathbf{n}_4$ *defined above such that the equilibrium will be an interior solution for non-signatories for* n *when* $n \notin [n_1, n_2]$ *and a non-signatory corner solution for* n *when* $\mathbf{n} \in [\mathbf{n}_1, \mathbf{n}_2]$, *moreover the equilibrium will be an interior solution for signatories for* n *when* $\mathbf{n} \notin [\mathbf{n}_3, \mathbf{n}_4]$ *and a signatory corner solution for* n *when* $\mathbf{n} \in [\mathbf{n}_3, \mathbf{n}_4]$;*(iv) If* $\mathbf{b} \in (\mathbf{b}_1(\mathbf{N}-\mathbf{1}), \mathbf{b}_2(\mathbf{N}-\mathbf{1}))$, *the equilibrium will be an interior solution for signatories for all* n *and a corner solution for non-signatories also for all* n *except for* $n = N-1$;*(v) Finally, if* $\mathbf{b} \in (\mathbf{0}, \mathbf{b}_1(\mathbf{N}-\mathbf{1})]$, *the equilibrium will be an interior solution for signatories for all* n *and a corner solution for non-signatories also for all* n.

We illustrate this Proposition in Figure 3.[20]

$\Rightarrow$ FIGURE 3 $\Leftarrow$

[19] In order to simplify the graphical representation we assume that the maximum value of real function $b_3(n)$ is given by the integer n^*.

[20] In order to simplify the graphical representation we assume that the maximum values of real functions $b_1(n)$ and $b_3(n)$ are given by the integers $\hat{n}$ and n^*. Notice that as $b_2(n)$ is a decreasing, strictly convex function and $b_2(N) = b_1(N) = 0$, $b_1(n)$ must be also a strictly convex function for big enough values of n. The previous results apply for $N > 5$ although with minimal changes they are also valid for $N = 5$.

As can be seen the parameter space is divided into five regions which are defined by four critical values of b. The first one is the maximum value of $b_3(n)$. Above this value the equilibrium will be an interior solution for signatories and non-signatories. Under this value the corner solutions appear. The second critical value is the maximum value of $b_1(n)$. Between these two critical values the equilibrium will be a corner solution for signatories for some values of n. The third critical value is the minimum value of $b_2(n)$, i.e. $b_2(N-1)$. In this third region the type of equilibrium depends on n. For low values of n we will find a corner solution for non-signatories. However for high values of n the signatories will choose zero emissions. Finally, the last critical value is $b_1(N-1)$ which defines a transition region almost identical to the bottom region for which the equilibrium will be an interior solution for signatories and a corner solution for non-signatories.

Thus for any parameters N and b, Proposition 2 indicates for, any number of signatories n, what type of solution there is to the Stackelberg emissions game, with corner solutions to take account of non-negative emissions constraints. Next using these results and the corresponding outputs and equilibrium payoff functions for signatories and non-signatories we conduct the stability analysis.

3.2 Membership Game for Stackelberg Model with Non-Negative Emissions

In this Section we show that the scope of the international cooperation for controlling an environmental problem depends critically on the level of the marginal environmental damage. We begin analyzing the stability for (relatively) high marginal damages, i.e., for low values of b.

Proposition 3 *If $b \leq b_1(N-1)$ the unique stable IEA of the Stackelberg model with non-negative emissions is the grand coalition.*

Proof: To conclude that the grand coalition is self-enforcing it is only necessary to check if the internal stability condition is satisfied, i.e. if $\triangle(N) = \pi^s(N) - \pi^f(N -$

1) ≥ 0 according to Def. 1. From (11) we know that the net benefits of the full-cooperative equilibrium of the emission game are $\pi^s(N) = a^2/2(b+N^2)$. Moreover, Prop. 2 establishes that for this range of parameter values the equilibrium of the emissions game is the non-signatories corner solution for $n < N$ but for this kind of solution the net benefits of the non-signatories are negative since their emissions are zero, see (14). The result is that the internal stability condition is satisfied: $\triangle(N) > 0$.■

When b is low enough as to induce the non-signatories to cut completely their emissions, the grand-coalition, i.e. the full-cooperative equilibrium, is self-enforcing since no country has incentives to exit from the agreement.

Next we show that the grand coalition cannot be a stable agreement for lower values of damage costs.

Proposition 4 *If $b \in [b_2(N-2), b_2(4) = N-4]$, there exists an upper bound given by the smallest integer no less than n_3 for the number of countries that belong to a self-enforcing IEA. This upper bound decreases when b increases.*

This result establishes that the scope of cooperation is very sensitive to changes in the level of marginal environmental damage. So that we have to expect that a reduction in the marginal damage leads to a reduction in the level of cooperation reached by a self-enforcing IEA. The explanation for this kind of relationship is given by the fact that the interdependence among the countries occurs through the damage function. Thus, when the marginal environmental damage is relatively high (a low b), the leadership of the countries in the agreement is strong and the signatories choose emission levels which induce non-signatories to select low values of emission, making exit from the agreement unprofitable. These effects are weakened as environmental damage costs get smaller.

Finally, we focus on the scope of cooperation when $b > N - 4$.

Proposition 5 *If $b > N - 4$ and $N > 5$ the maximum level of cooperation that can be achieved by a self-enforcing IEA is* 3.

Then taking into account that $N(N-4)/4 > N-4$ when $N > 5$ and that $b > N(N-4)/4$ is a sufficient condition to have an interior solution for signatories and non-signatories we can conclude that:

Corollary 2 *If $b > N(N-4)/4$ then the Stackelberg equilibrium is an interior solution for signatories and non-signatories and the maximum level of cooperation that can be achieved by a self-enforcing IEA is* 3.

It is also easy to show that for a large enough value of b the stable level of cooperation is 2. These conclusions clarify the previous results in the literature which have been derived assuming that there are interior solutions. According to our results restricting parameter values to guarantee interior solutions is a sufficient condition to get stable IEAs with a small number of signatories but it is not a necessary condition. We have obtained that it is enough with $b > N-4$ to have a maximum of three countries in an IEA. But between $N-4$ and $N(N-4)/4$ the Stackelberg equilibrium is a corner solution for different values of n depending on the value of b. This means that what is necessary and sufficient to get a small degree of cooperation is a high value of b and not interior solutions for signatories and non-signatories.

Thus we have shown that even if we take seriously non-negative emission constraints, the Stackelberg model can have stable IEAs as large as the grand coalition and as low as a bilateral agreement depending on the value of the marginal environmental damage.

Finally, although we have shown that, allowing for non-negative emission constraints, it is still possible to get the grand coalition as a stable IEA, it could still be the case that imposing non-negative emissions has a significant effect on the size of a stable IEA in the sense that for any particular set of parameter values the size of IEA is significantly smaller than would be calculated if one simply ignored the constraints. To test this we have taken values of $a = 1000$, values of $N = 10, 20, , 150$, and 1500 values of b. For each set of parameter values we calculated the size of the stable IEA imposing non-negative emission constraints and without imposing such constraints. Three points emerged: (i) first, we confirmed, that, for all N, by varying b the maximum size of stable IEA obtained was the grand coalition, whether or not the non-negative emission constraints were imposed; (ii) for any set of parameter values, the size of the stable IEA with the non-negative emissions constraints imposed was never greater than the size of stable IEA when the constraints were ignored; (iii), but crucially, the maximum difference in the size of stable IEA between restricted and unrestricted emissions is 1. We illustrate this in Table 1 by showing for a range of values of N the average size of stable IEA (averaged over different values of b) with and without the constraints. As can be seen the difference in average size by imposing the constraints is tiny as one expected given that the maximum difference is 1. This means that for a lot of cases there is no difference in the level of cooperation with and without constraints One has to go to the third decimal place to detect a difference in average size.

Table 1

Average Size of Stable IEA

With and Without Constraints

N	With Constraints	Without Constraints
10	6.268	6.272
30	16.677	16.680
50	27.148	27.151
70	37.629	37.631
90	48.110	48.111
110	58.589	58.590
130	69.068	69.069
150	79.545	79.546

4 Conclusions

In this paper we have provided analytical proofs of the main results of the linear-quadratic version of the widely used model of stable IEAs introduced by Carraro and Siniscalco (1991) and Barrett (1994). Moreover, we have shown, analytically, that these results are robust to the introduction of constraints that emissions must be non-negative. While such constraints significantly complicate the analysis, they leave the main findings of the original literature almost completely unaffected. Since it is clearly right that such non-negative constraints should be taken into account, it is important to know the original results are robust and this is one of the contributions of our paper. Another contribution is that our results qualifies the claim by Diamantoudi and Sartzetakis (2002a) that imposing non-negative emissions constraints significantly reduces the size of the stable IEAs. This paper argues that this claim is right only when emissions are restricted to be strictly positive, however if countries can completely cut their emissions the size of the stable IEAs depends directly on the marginal environmental damage so that a high degree of cooperation can be reached if the marginal environmental damage is big enough. Thus, this paper gives an analytical support to the Barrett's conclusion on the scope of cooperation that can be achieved by a self-enforcing IEA.

Of course there are many other respects in which the original models of stable IEAs need to be extended - richer concepts of self-enforcing agreements, asymmetric countries, concepts of fairness, dealing with stock pollutants, allowing for uncertainty and learning - and the authors of the original papers and many others have made important contributions to addressing these extensions (see again Finus (2001) for an excellent survey). We too have addressed some of these extensions (Rubio and Ulph (2002b, 2003), Ulph (2002 a, b)). However, since the basic model continues to attract interest, it is important to make sure that its properties rest on thorough analysis, and this paper contributes to that purpose.

References

[1] Barrett, S. (1994), "Self-enforcing international environmental agreements", *Oxford Economic Papers* **46**: 878-94.

[2] Barrett, S. (2001), "International cooperation for sale", *European Economic Review* **45**:1835-50.

[3] Barrett, S. (2003), *Environmental & Statecraft*, Oxford:O.U.P.

[4] Carraro, C. and Siniscalco, D. (1991), "Strategies for the international protection of the environment", CEPR Discussion Paper 568.

[5] Carraro, C. and Siniscalco, D. (1993), "Strategies for the international protection of the environment", *Journal of Public Economics* **52**: 309-28.

[6] Chander, P. and Tulkens, H. (1995), "A core-theoretic solution for the design of cooperative agreements on transfrontier pollution", *International Tax and Public Finance* **2**:279-93.

[7] Chander, P. and Tulkens, H. (1997), "The core of an economy with multilateral environmental externalities", *International Journal of Game Theory* **26**: 379-401.

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

[9] Diamantoudi, E. and Sartzetakis, E. S. (2002a), "Stable international environmental agreements: An analytical approach", Nota Di Lavoro 7.2002, FEEM, Venice.

[10] Diamantoudi, E. and Sartzetakis, E.S. (2002b), "International environmental agreements: The role of foresight", mimeo, Department of Economics, University of Aarhus.

[11] Ecchia, G. and Mariotti, M. (1997), "The stability of international environmental coalitions with farsighted countries: Some theoretical observations" in C. Carraro (ed.), *International environmental negotiations*, Cheltenham, UK: Edward Elgar.

[12] Ecchia, G. and Mariotti, M. (1998), "Coalition formation in international environmental agreements and the role of institutions", *European Economic Review* **42**:573-82.

[13] Finus, M. (2001), *Game theory and international environmental cooperation*, Cheltenham, UK: Edward Elgar.

[14] Hoel, M. (1992), "International environmental conventions: the case of uniform reductions of emissions", *Environmental and Resource Economics* **2**: 141-59.

[15] Hoel, M. and Schneider, K. (1997), "Incentives to participate in an international environmental agreement", *Environmental and Resource Economics* **9**:153-70.

[16] Na, S.L. and Shin, H.S. (1998), "International environmental agreements under uncertainty", *Oxford Economic Papers* **50**: 173-85.

[17] Petrakis, E. and Xepapadeas, A. (1996), "Environmental consciousness and moral hazard in international agreements to protect the environment", *Journal of Public Economics* 60:95-110.

[18] Ray, D. and Vohra, R. (2001), "Coalition power and public goods", *Journal of Political Economy* **109**:1355-84.

[19] Rubio, S. and Casino, B. (2001), "International cooperation in pollution control", IVIE Working Paper WP-AD 2001-21.

[20] Rubio, S. and Ulph, A. (2002a), "Leadership and self-enforcing environmental agreements with non-negative emissions", Discussion Paper in Economics and Econometrics 0211, Department of Economics, University of Southampton.

[21] Rubio, S. and Ulph, A. (2002b), "A simple dynamic model of international environmental agreements with a stock pollutant", Discussion Paper in Economics and Econometrics 0209, Department of Economics, University of Southampton.

[22] Rubio, S. and Ulph, A. (2003), "An infinite-horizon model of dynamic membership of international environmental agreements", FEEM Working Paper No. 57.2003

[23] Tulkens, H. (1998), "Cooperation versus free-riding in international environmental affairs: two approaches" in N. Hanley and H. Folmer (eds.), *Game theory and the environment*, Cheltenham:Edward Elgar, Chap. 2.

[24] Ulph, A. (2002a), "International environmental agreements with a stock pollutant, uncertainty and learning", Discussion Paper in Economics and Econometrics 0207, Department of Economics, University of Southampton.

[25] Ulph, A. (2002b), "Environmental agreements with uncertainty and learning - the case of stock-dependent unit costs",Discussion Paper in Economics and Econometrics 0208, Department of Economics, University of Southampton.

[26] Wagner, U.J. (2001), "The design of stable international environmental agreements: Economic theory and political economy", *Journal of Economic Surveys* 15:377-411.

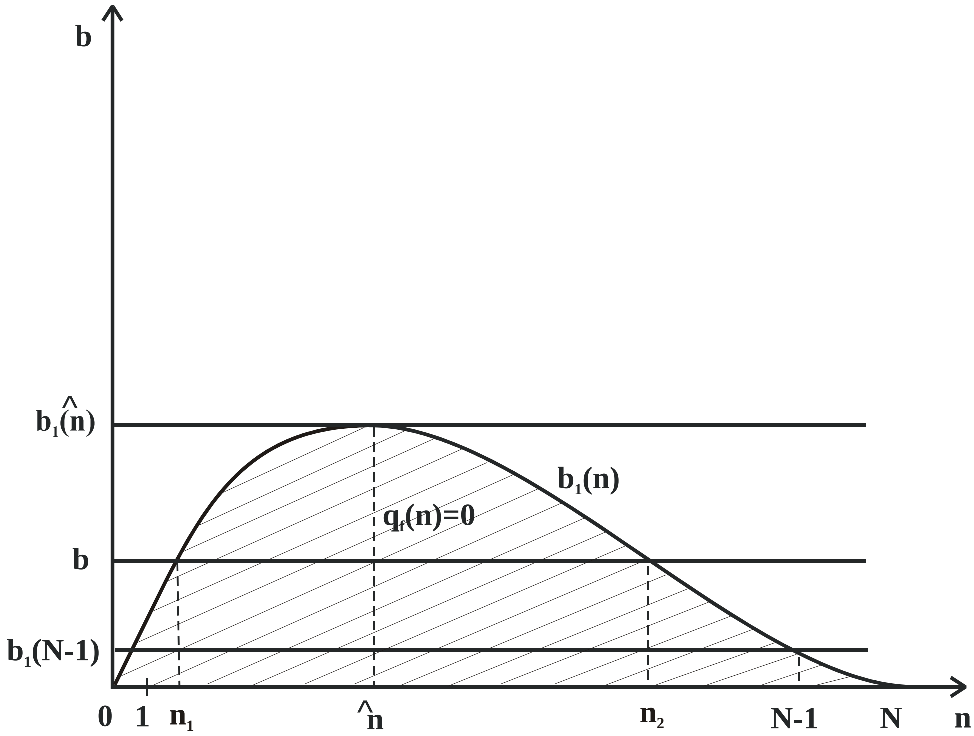

Figure 1. Corner solutions for non-signatories

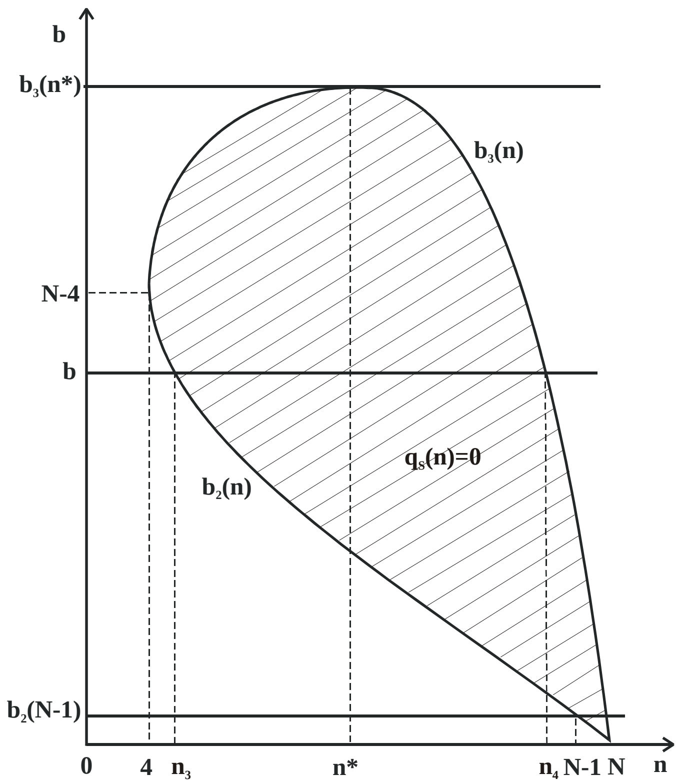

Figure 2. Corner solutions for signatories

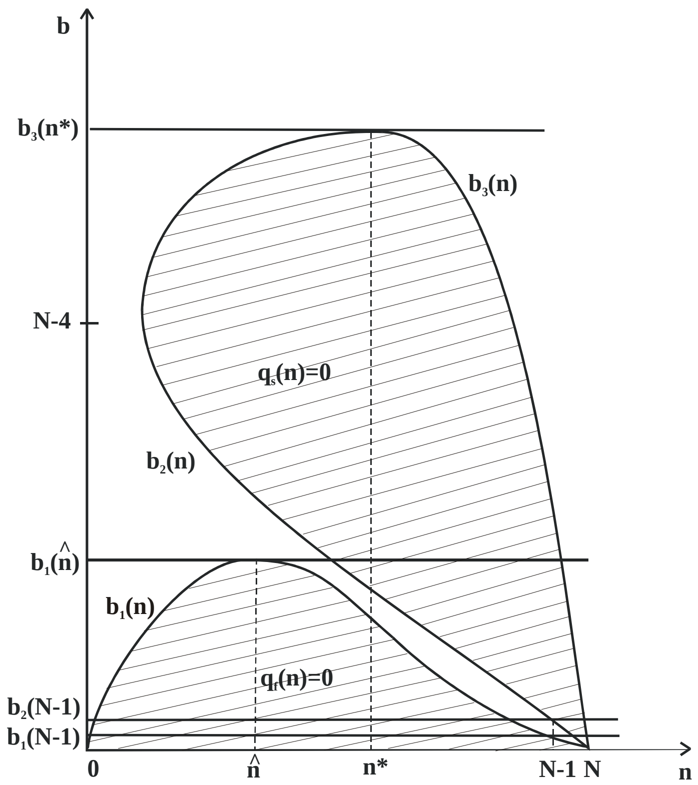

Figure 3. Parameter Space

Technical Appendix

A Proof of Proposition 1

First, we show that there does not exist any value of b such that $g(b,n), h(b,n) \leq 0$ for a given value of n. For given n $h(b,n)$ is strictly convex function with respect to b with a minimum for a negative value of b and an intersection point with the vertical axis also negative. This implies that $h(b,n) = 0$ has only a positive real solution given by

$$b_1(n) = \frac{1}{2}\left\{-(N + n^2 - 2n) + (n^4 - 4n^3 + 2Nn^2 + N^2)^{1/2}\right\}, \tag{1}$$

so that $h(b,n)$ will be negative or zero if $b \in (0, b_1(n)]$ and strictly positive if $b > b_1(n)$.

On the other hand, function $g(b,n)$ is strictly convex with respect to b and presents a minimum for $b = (N-n)(n-2)/2$. For this minimum the value of the function is $n(N-n)^2(4-n)/4$ which implies that $g(b,n) > 0$ for $n = 1, 2, 3$ and $b > 0$. Then we can conclude that there does not exist any value of b such that $g(b,n), h(b,n) \leq 0$ for $n = 1, 2, 3$ since $g(b,n)$ is always positive. For $n = 4$, $g(b,4) = 0$ for $b = N - 4$ and positive for $b \neq N - 4$ but as $b = N - 4$ is bigger than $b_1(4)$ if $N > 4$, $g(b,n), h(b,n)$ cannot be negative or zero at the same time for $n = 4$.[1] For $n > 4$, $g(b,n) = 0$ has two positive real solutions

$$\begin{aligned} b_2(n) &= \frac{1}{2}\left\{(N-n)\left[n - 2 - (n^2 - 4n)^{1/2}\right]\right\}, & (2)\\ b_3(n) &= \frac{1}{2}\left\{(N-n)\left[n - 2 + (n^2 - 4n)^{1/2}\right]\right\} & (3) \end{aligned}$$

such that $b_2(n) < b_3(n)$ and $g(b,n)$ will be negative or zero if $b \in [b_2(n), b_3(n)]$.

Next, we show that $b_1(n)$ is lower than $b_2(n)$. Let's suppose now that $b_1(n) \geq b_2(n)$ which yields

$$\begin{aligned} & -(N + n^2 - 2n) + (n^4 - 4n^3 + 2Nn^2 + N^2)^{1/2} \\ \geq \;& (N-n)(n-2) - (N-n)(n^2 - 4n)^{1/2} > 0, \end{aligned} \tag{4}$$

[1] We assume that N is big enough to satisfy $N > 4$.

simplifying terms we have that

$$(n^4 - 4n^3 + 2Nn^2 + N^2)^{1/2} \geq N(n-1) - (N-n)(n^2-4n)^{1/2} > 0.$$

Then squaring and simplifying terms again we get

$$n(n^2 - (3+N)n + 3N) \geq -(n-1)(N-n)(n^2-4n)^{1/2},$$

where the left-hand side of the inequality is negative for $n \in (4, N)$, then multiplying by -1 we obtain

$$0 < -n(n^2 - (3+N)n + 3N) \leq (n-1)(N-n)(n^2-4n)^{1/2}.$$

Finally, squaring again and simplifying terms we get a contradiction

$$4n(n^2 - 2Nn + N^2) \leq 0,$$

since $n^2 - 2Nn + N^2$ is positive for $n < N$. Consequently, we can establish that $b_1(n) < b_2(n)$ for all $n > 4$ and then $g(b,n)$ and $h(b,n)$ cannot be negative or zero at the same time. This means that for each combination (b,n) only one of the three solution characterized in the paper applies.

Now according to the sign of $h(b,n)$ and $g(b,n)$ it can be established for each combination (b,n) which it is the kind of solution the problem has. For $n = 1, 2, 3$ we have two kinds of solutions. For $b \leq b_1(n)$ we have that $h(b,n) \leq 0$ and the equilibrium yields $q_f = 0$ and $q_s > 0$ whereas for $b > b_1(n)$ $h(b,n) > 0$ and the equilibrium is the interior solution since $g(b,n)$ is also strictly positive. For $n = 4$ we find the three kinds of solutions but the signatory corner solution only for $b = N - 4$ since only for $(b = N-4, n = 4)$ $g(b,n)$ is zero. Finally, as we can order the critical values of $b : b_1(n) < b_2(n) < b_3(n)$, if $b > b_3(n)$ we have that $h(b,n)$ and $g(b,n)$ are strictly positive and the equilibrium yields $q_s > 0, q_f > 0$, if $b \in [b_2(n), b_3(n)]$ $g(b,n)$ is negative or zero and the equilibrium yields $q_s = 0, q_f > 0$, if $b \in (b_1(n), b_2(n))$ $g(b,n)$ and $h(b,n)$ are strictly positive again and then the equilibrium yields $q_s > 0, q_f > 0$, and finally if $b \leq b_1(n)$ $g(b,n)$ is strictly positive and $h(b,n)$ is negative or zero and the equilibrium then yields $q_s > 0, q_f = 0$. The graphical representation of this result appears in Fig. 3

B Proof of Lemma 1

The results in Lemma 1 are shown from the properties of $b_1(n)$. Thus, what we show first is that $b_1(0) = b_1(N) = 0$ and that $b_1(n)$ presents a unique extreme in the interval $(1, N-1)$ which is a maximum.

By substitution it is easy to check that $b_1(0) = b_1(N) = 0$. To show that the unique extreme of $b_1(n)$ is a maximum we use the inverse function of $b_1(n)$. In order to obtain this function we rewrite $h(b,n)$ as

$$h(b,n) = (b+1)n^2 - (2b+N)n + (b+N)b,$$

and then from $h(b,n) = 0$ we get:

$$\mathbf{n}^+(\mathbf{b}) = \frac{2b + N + (N^2 - 4b^3 - 4b^2)^{1/2}}{2(b+1)}, \quad \mathbf{n}^-(\mathbf{b}) = \frac{2b + N - (N^2 - 4b^3 - 4b^2)^{1/2}}{2(b+1)}, \qquad (5)$$

which yield real values for $N^2 - 4b^3 - 4b^2 \geq 0$. Thus these two functions yield real values in the interval $[0, \tilde{b}]$ where $\tilde{b}$ is the unique positive solution of equation $N^2 - 4b^3 - 4b^2 = 0$. Moreover, the sign of their first derivatives is[2]:

$$\begin{aligned} \frac{dn^+}{db} &= -\frac{2b^3 + 6b^2 + 4b + N^2 + (N-2)(N^2 - 4b^3 - 4b^2)^{1/2}}{2(b+1)^2(N^2 - 4b^3 - 4b^2)^{1/2}} < 0, \\ \frac{dn^-}{db} &= \frac{2b^3 + 6b^2 + 4b + N^2 - (N-2)(N^2 - 4b^3 - 4b^2)^{1/2}}{2(b+1)^2(N^2 - 4b^3 - 4b^2)^{1/2}} > 0, \end{aligned}$$

and

$$n^+(\tilde{b}) = n^-(\tilde{b}) = \frac{2\tilde{b} + N}{2(\tilde{b}+1)} = \tilde{n},$$

so that it can be concluded that $n^-(b)$ is the inverse function of $b_1(n)$ in interval $[0, \tilde{n}]$ and that it is increasing in that interval and also that $n^+(b)$ is the inverse function of $b_1(n)$ in interval $[\tilde{n}, N]$ and that it is decreasing in that interval. Then we have that function $b_1(n)$ is increasing in interval $[0, \tilde{n})$ and decreasing in interval $(\tilde{n}, N]$ and therefore it presents a maximum at $n = \tilde{n}$.

[2] It is easy to show that the numerator of dn^-/db is positive.

Notice that

$$\lim_{b \to \tilde{b}^-} \frac{dn^+}{db_1} = -\infty; \quad \lim_{b \to \tilde{b}^-} \frac{dn^-}{db_1} = +\infty,$$

which implies that $db_1/dn = 0$ when $n = \tilde{n}$. Finally, it is immediate to show that $\tilde{n} \in [1, N-1]$. Then given the properties of $b_1(n)$ we can determine for any b the kind of solutions depending on the value of n. Next we explain this.[3]

According to footnote 16 in the main text we define $\hat{n}$ as the integer that maximizes $b_1(n)$ in the domain $\{1, 2, ..., N-1\}$. Then if $b > b_1(\hat{n})$ we have that $b > b_1(n)$ for all n and according to Prop. 1 the equilibrium is the interior solution for all n. If $b < b_1(\hat{n})$ equation $b = b_1(n)$ has two, positive real roots that we call n_1, n_2 such that $n_1 < n_2$. Then as $b_1(n)$ is first increasing and afterwards decreasing we have that $b > b_1(n)$ when $n < n_1$ and $n > n_2$ and accordingly the equilibrium is the interior solution in those two cases. However, when $n \in [n_1, n_2]$ we have that $b \leq b_1(n)$ and Prop. 1 establishes that the equilibrium is the corner solution. Finally, for $b = b_1(\hat{n})$ we find three possibilities. If $\hat{n} = \tilde{n}$ equation $b_1(\hat{n}) = b_1(n)$ yields $n_1 = n_2 = \hat{n}$ and the equilibrium is the interior solution for all n except for $\hat{n}$. If $\hat{n} \neq \tilde{n}$, because $\tilde{n}$ is not an integer and the function is not symmetric, equation $b_1(\hat{n}) = b_1(n)$ yields $n_1 < n_2$ where one of this two values by definition is $\hat{n}$, the closest integer to $\tilde{n}$, see footnote 16. Then given the behaviour of the function around $\tilde{n}$ the distance between n_1 and n_2 must be less than the unity and the equilibrium again is the interior solution for all n except for $\hat{n}$. If the function is symmetric it could present two maximum $\hat{n}_1 = n_1$ and $\hat{n}_2 = n_2$ provided that $\tilde{n}$ is in the center of interval $(\hat{n}_1, \hat{n}_2)$,and the equilibrium would be the interior solution for all n except for $\hat{n}_1$ and $\hat{n}_2$.[4]

[3]Fig. 1 can help to the reader to follow our argument.

[4]Notice that symmetry is a necessary condition to have two maximums but not a sufficient condition.

C Proof of Corollary 1

This result is immediate from Lemma 1 provided that $b_1(N-1)$ is the minimum value of $b_1(n)$ for $n = 1, 2, ..., N-1$ since in this case $n_1 < 1$ and $n_2 \geq N-1$. As $b_1(n)$ is first increasing for $n < \tilde{n}$ and afterwards decreasing, $b_1(N-1)$ is the minimum value if $b_1(N-1)$ is lower than $b_1(1)$.

First we calculate these two values:

$$\begin{aligned} b_1(1) &= \frac{1}{2}\left\{-(N-1) + (N^2+2N-3)^{1/2}\right\}, \\ b_1(N-1) &= \frac{1}{2}\left\{-(N^2-3N+3) + (N^4-6N^3+15N^2-14N+5)^{1/2}\right\}. \end{aligned}$$

Let´s suppose now that $b_1(1) \leq b_1(N-1)$ which yields

$$-(N-1) + (N^2+2N-3)^{1/2} \leq -(N^2-3N+3) + (N^4-6N^3+15N^2-14N+5)^{1/2},$$

simplifying terms we have that

$$0 < N^2 - 4N + 4 + (N^2+2N-3)^{1/2} \leq (N^4-6N^3+15N^2-14N+5)^{1/2}.$$

Then squaring and simplifying terms again we get

$$0 < 2(N^2-4N+4)(N^2+2N-3)^{1/2} \leq 2N^3 - 10N^2 + 16N - 8.$$

Finally, squaring again and simplifying terms we get a contradiction

$$0 \leq -16N^5 + 144N^4 - 512N^3 + 896N^2 - 768N + 256,$$

since the right-hand side of the inequality is negative for $N \geq 3$. Then we can conclude that $b_1(1) > b_1(N-1)$ and from Lemma 1we obtain Corollary 1.

D Proof of Lemma 2

As for Lemma 1, the proof of Lemma 2 derives from the properties of the functions $b_2(n)$ and $b_3(n)$. By substitution in (2) we get that $b_2(4) = N-4$ and $b_2(N) = 0$. Remember that signatories' emissions are always positive for $n \in \{1, 2, 3\}$ and $b > 0$. See Appendix A.

On the other hand, if we take the first derivative of $b_2(n)$ we obtain

$$\frac{db_2}{dn} = \frac{1}{2}\left\{N - 2n + 2 + \frac{2n^2 - (6+N)n + 2N}{(n^2 - 4n)^{1/2}}\right\},$$

that presents the following limits:

$$\begin{aligned}
\lim_{n \to 4} \frac{db^2}{dn} &= -\infty, \\
\lim_{n \to N} \frac{db^2}{dn} &= \frac{1}{2}\left\{-(N-2) + \frac{N(N-4)}{(N^2 - 4N)^{1/2}}\right\} < 0 \text{ for } N > 4.
\end{aligned}$$

Moreover, its second derivative is:

$$\frac{d^2 b}{dn^2} = \frac{1}{2}\left(-2 + \frac{2n^3 - 12n^2 + 12n + 4N}{(n^2 - 4n)^{1/2}}\right).$$

Let ´s suppose that this second derivative is negative or zero. This implies that

$$0 < 2n^3 - 12n^2 + 12n + 4N \leq 2(n^2 - 4n)^{1/2},$$

squaring and reordering terms we obtain the following inequality

$$4n^6 - 48n^5 + 192n^4 + (16N - 288)n^3 - (96N - 140)n^2 + (96N + 16)n + 16N^2 \leq 0.$$

It is easy but tedious to show that the left-hand side of this inequality is positive for $n \geq 4$ yielding a contradiction. So that we can conclude that $d^2 b_2/dn^2 > 0$ which allows us to establish that $b_2(n)$ is a decreasing, strictly convex function in interval $[4, N]$.

Next, we study the properties of function $b_3(n)$. By substitution in (3) we get that $b_3(4) = N - 4$ and $b_3(N) = 0$. Moreover, it is easy to show that $b_3(n)$ is a strictly concave function with a maximum in interval $(4, N-1)$ equal to $N^2/2(N-2)$.[5] So that for $n \in (4, N^2/2(N-2))$ $b_2(n)$ increases and for $n \in (N^2/2(N-2), N)$ decreases. The maximum value for the function is $b_3(N^2/2(N-2)) = N(N-4)/4 > b_2(4) = N - 4$.

[5] In order to have an interval with more than one point we assume that $N > 5$.

Then given the properties of $b_2(n)$ and $b_3(n)$ we can determine for any b the kind of solutions for signatories depending on the value of n. Next we explain this.[6]

According to footnote 18 in the main text we define n^* as the integer that maximizes $b_1(n)$ in the domain $\{4, ..., N-1\}$. Then if $b > b_3(n^*)$ we have that $b > b_3(n)$ for all n and according to Prop. 1 the equilibrium is the interior solution for all n. If $b \in (N-4, b_3(n^*))$ equation $b = b_3(n)$ has two positive real roots that we call n_3, n_4 such that $n_3 < n_4$. Then as $b_3(n)$ is first increasing and afterwards decreasing we have that $b > b_3(n)$ when $n < n_3$ and $n > n_4$ and accordingly the equilibrium is the interior solution in those two cases. However, when $n \in [n_3, n_4]$ we have that $b \leq b_1(n)$ and Prop. 1 establishes that the equilibrium is the corner solution. If $b \in [b_2(N-1), N-4]$, we have that $b = b_3(n)$ has a unique positive solution n_4 that along with n_3 obtained from $b = b_2(n)$ define an interval $[n_3, n_4]$ for which $b \in [b_2(n), b_3(n)]$, then according to Prop. 1 the equilibrium is the corner solution when n belongs to that interval and the interior solution if that is not the case. If $b \in (0, b_2(N-1))$ we have that $b < b_2(n)$ for all n and according to Prop. 1 the equilibrium is the interior solution for all n. Finally, for $b = b_3(n^*)$ we find three possibilities. If $n^* = \tilde{n}$ equation $b_3(n^*) = b_3(n)$ yields $n_3 = n_4 = n^*$ and the equilibrium is the interior solution for all n except for n^*. If $n^* \neq \tilde{n}$, because $\tilde{n}$ is not an integer and the function is not symmetric, equation $b_3(n^*) = b_3(n)$ yields $n_3 < n_4$ where one of this two values by definition is n^*, the closest integer to $\tilde{n}$, see footnote 18. Then given the behaviour of the function around $\tilde{n}$ the distance between n_3 and n_4 must be less than the unity and the equilibrium again is the interior solution for all n except for n^*. If the function is symmetric it could present two maximum $n_1^* = n_1$ and $n_2^* = n_2$ provided that $\tilde{n}$ is in the center of interval (n_1^*, n_2^*),and the equilibrium would be the interior solution for all n except for n_1^* and n_2^*.

[6] Fig. 2 can help to the reader to follow our argument.

E Proof of Proposition 2

Points (i)-(iii) and (v) follow from Lemmas 1 and 2 and Corollary 1 provided $b_1(\hat{n}) < b_3(n^*)$. Point (iv) only occurs if $b_1(1) > b_2(N-1)$ and $b_1(N-2) > b_2(N-1)$ since function $b_1(n)$ is first increasing and afterwards decreasing and $b_1(N-1) < b_2(N-1)$ according to Prop. 1. First we show that $b_1(\hat{n}) < b_3(n^*)$ for $N > 5$. From the proofs of Proposition 1 and Lemma 2 we know that $b_1(4) < b_2(4) = b_3(4) = N - 4 < b_3(n^*)$.

On the other hand, $b_1(n)$ can be increasing or decreasing at $n = 4$ depending on the number of countries, N. Let´s suppose that

$$\frac{db_1(4)}{dn} = -3 + \frac{16 + 4N}{(32N + N^2)^{1/2}} \geq 0.$$

This implies that

$$16 + 4N \geq 3(32N + N^2)^{1/2},$$

that squaring and simplifying yields

$$256 - 160N + 7N^2 \geq 0.$$

It is easy to show that the left-hand side of the inequality is negative for $N \in (5, 21]$. So that we can conclude that $db_1(4)/dn < 0$ which means that $\hat{n} \leq 4$. The previous result also allows us to establish that for $N > 21$, the first derivative of $b_1(n)$ at $n = 4$ is positive which means that $\hat{n} \geq 4$.

Next we suppose that $N \in (5, 21]$. In that case, $\hat{n}$ is an integer in the domain $\{1, 2, 3, 4\}$. If $\hat{n} = 4$, we already know that $b_1(4) < N - 4 < b_3(n^*)$ and it is established that $b_1(\hat{n}) < b_3(n^*)$. If $\hat{n}$ were an integer different from 4, it is easy to show that $b_1(n) < N - 4$ for $n = \{1, 2, 3\}$ so that we can also conclude that $b_1(\hat{n}) < b_3(n^*)$ for this values of n.

Next, we suppose that $N > 21$. In that case, $\hat{n} \geq 4$. For $\hat{n} = 4$ the same argument than the one we have just used is applied. For $\hat{n} > 4$, we know from Proposition 1 that $b_1(\hat{n}) < b_2(\hat{n})$ and from the proof of Lemma 2 that $b_2(\hat{n}) < N - 4 < b_3(n^*)$ so that we find that $b_1(\hat{n}) < b_3(n^*)$ as we wanted to establish.

Finally, we show that $b_1(1) > b_2(N-1)b$ and $b_1(N-2) > b_2(N-1)$. Let´s suppose first that $b_1(1) \leq b_2(N-1)$. This implies that

$$0 < \left((N-1)^2 + 4(N-1)\right)^{1/2} \leq 2N - 4 - \left((N-1)^2 - 4(N-1)\right)^{1/2}.$$

Squaring yields

$$2(2N-4)\left((N-1)^2 - 4(N-1)\right)^{1/2} \leq (2N-4)^2 - 8(N-1),$$

squaring again, simplifying and reordering terms we obtain the following inequality

$$32N^3 - 240N^2 + 448N - 256 \leq 0.$$

This inequality yields a contradiction for $N > 5$ so that we can conclude that $b_2(N-1) < b_1(1)$.

Next, we suppose that $b_1(N-2) \leq b_2(N-1)$. This implies that

$$0 < (N^4 - 10N^3 + 41N^2 - 72N + 48)^{1/2} \leq -4N + N^2 + 5 - (N^2 - 6N + 5)^{1/2}.$$

Squaring yields

$$2(-4N + N^2 + 5)(N^2 - 6N + 5)^{1/2} \leq 2N^3 - 14N^2 + 26N - 18,$$

squaring again, simplifying and reordering terms we obtain the following inequality

$$16N^4 - 144N^3 + 400N^2 - 464N + 176 \leq 0$$

This inequality yields a contradiction fro $N > 5$ so that we can conclude that $b_2(N-1) < b_1(N-2)$.

F Proof of Proposition 4

For $b \in [b_2(N-2), N-4]$, we have that $n_4 \in (N-1, N)$ so that we have a signatory corner solution for $n \in [n_3, n_4]$. This is a consequence of the fact that $b_3(N-1) > b_2(4) = N-4$.[7]

[7]This is very easy to show so that we omit it.

Now given a value of b we select n such that n and $n-1 \in [n_3, n_4]$ and we check if the internal stability condition can be satisfied. In this case we have that both n and $n-1$ are signatory corner solutions so that according to (13) from the main text the internal stability condition is given by

$$\begin{aligned}\triangle(n) &= -\frac{a^2(N-n)^2}{2(b+N-n)^2} - \frac{a^2[b-(N-n+1)(N-n-1)]}{2(b+N-n+1)^2} \\ &= -\frac{a^2[b^3+L(n)b^2+M(n)b+P(n)]}{2(b+N-n)^2(b+N-n+1)^2},\end{aligned}$$

where

$$L(n) = 2(N-n)+1 > 0, \quad M(n) = 3(N-n)^2 + 2(N-n) > 0$$

$$P(n) = 2(N-n)^3 + 2(N-n)^2 > 0$$

So, $\triangle(n)$ is negative, in fact, is negative for all $b > 0$. Consequently if there exists a self-enforcing IEA the number of countries in the agreement must be equal to or less than n_3, in fact, as it is not guarantee that n_3 is an integer, the upper bound for the number of countries in a self-enforcing IEA must be defined as the smallest integer no less than n_3. Finally, from Lemma 2 we know that $b_2(n)$ is a strictly convex, decreasing function defined in interval [4, N], then as n_3 is defined as the unique, positive real root of equation $b = b_2(n)$, we can conclude that n_3 decreases when b increases and that, consequently, the smallest integer no less than n_3 also decreases.

G Proof of Proposition 5

According to Proposition 4 if $b = N-4$ the maximum number of countries in a self-enforcing IEA is four then the question to answer now is whether an agreement of four countries can be self-enforcing for $b > N-4$. For $b > N-4$ and $n = 4$, according to Proposition 1, the equilibrium is the interior solution for signatories and non-signatories. Then using (11) and (12) from the main text we get

$$\triangle(4) = -\frac{a^2N^2k(b)}{2[(b+N-3)^2+9b]^2[(b+N-4)^2+16b]},$$

where

$$
\begin{aligned}
k(b) &= 3b^4 + (6N+48)b^3 + (2N^2 - 40N + 101)b^2 \\
&\quad -(2N^3 - 7N^2 - 10N + 39)b - (N^4 - 14N^3 + 73N^2 - 168N + 144).
\end{aligned}
$$

It can be easily shown that $k''(b)$ is a strictly convex function with a minimum at $b = (8+N)/2$. For this minimum we have that the second derivative is positive for all N. This implies that $k'(b)$ is increasing for all $b > 0$ and consequently is also increasing for $b > N-4$. Then as $k'(N-4) = 252N^3 - 1212N^2 - 3408N + 16716$ is positive for $N > 5$ we can conclude that the first derivative is positive for $b > N-4$, which implies that $k(b)$ is increasing for $b > N-4$. Finally, as $f(N-4) = 8N^4 - 99N^3 + 362N^2 - 207N - 676$ is positive for $N > 5$ we have that $k(b)$ is positive for $b > N-4$ so that $\triangle(4)$ is negative and the internal stability condition is not satisfied. This show that $n = 4$ cannot be the Nash equilibrium of the membership game.