Climate coalitions cannot avert the crisis only by moving first in choosing emission and abatement targets

Effrosyni Diamantoudi, Concordia University
Eftichios S Sartzetakis, University of Macedonia*
Stefania Strantza, Thompson Rivers University

Abstract

The paper provides a fresh look at the literature on the formation of international environmental agreements by introducing into the classic model emissions and abatement as countries' separate choice variables. The model's structure is kept unchanged, assuming a two stage game in which the internal and external stability conditions define coalition's stability. We illustrate the way in which each of the three components of countries' welfare, benefits from own emissions, damages from aggregate emissions and own abatement costs, interact in determining nonsignatories' equilibrium choices, which in turn, determine the stable coalition's size. We show that, ceteris paribus, as abatement becomes cheaper, nonsignatories become more responsive to signatories' choices, strengthening the signatories' leadership position, allowing thus, largest stable coalitions to be formed. However, when abatement costs are low the same choices are individually rational, that is, forming a coalition does not add much over the Nash. Furthermore, large stable coalitions are possible under high abatement costs, only if damages are high relative to benefits, but such coalitions require negative net emissions, which is impossible. Thus, even if the coalition has leadership power in setting abatement and emission targets, significant reductions in free-riding incentives can be achieved only when it is welfare irrelevant, i.e., when the same targets are individually rational. Given that the current situation is characterized by high damages and high abatement costs, further action is required

*Corresponding author: E.S. Sartzetakis, Department of Economics, University of Macedonia, 156 Egnatia Str., 54636. Email: esartz@uom.edu.gr

to increase the size and effectiveness of coalitions so as to avert the crisis.

Keywords: climate coalitions, size and effectiveness of coalitions, choice over emissions and abatement

JEL: D6, Q5, C7

1 Introduction

Climate change evolves rapidly into a crisis while the international community is straggling to agree on the required coordinated action. A significant literature has developed over the last few decades, using game theory as a tool, and attempting to provide answers to the following twofold problem: can large coalitions form to reduce net emissions and do they attain significant welfare improvements by doing so?. The main body of this literature, to which the present paper contributes, has suggested that if a group of countries forms a coalition and enjoys first mover advantage in setting net emission targets, it can decrease free-riding incentives and support a larger coalition size in the first place. It has also been suggested that such coalitions, small or large, achieve very little welfare improvement, a result coined by Barret (1994) as the paradox of cooperation.

The present paper consolidates both of the above suppositions using general functional forms, and establishes that leadership advantage solely over target specification cannot avert climate crisis. Furthermore, the paper provides a unifying approach integrating the different choice-variable alternatives resulting in clear and intuitive explanations over the underlying forces driving the aforementioned results.

We define individual welfare as the difference between benefits from emissions (production), and the sum of individual abatement cost (technology) and damages from global emissions. We introduce emissions and abatement as countries' separate choice variables instead of restricting countries to choose only one or the other, as it is the case in the existing literature. Formalizing the choice of emissions and abatement independently is not only more realistic, but also allows us to disentangle the role of benefits, damages and abatement cost in determining the size of stable coalitions. In doing so, we explain the no longer paradoxical nature of the paradox of cooperation.

We formulate coalition formation as a two stage game following the majority of the literature. In the first stage, each country decides individually whether to stay in the climate coalition (and thus, commit to the maximization of the coalitional welfare) or exit and act selfishly (by maximizing its own welfare). In the second stage, the climate coalition chooses emissions and abatement first, acting as leader, and the outsiders follow with their re-

spective choices. We use general functional forms to illustrate the difference between the two-choice-variable and the single-choice-variable models.

It is well known in the literature that the size of stable coalitions is associated to the size where the coalition members' welfare equates that of the outsiders. Coalition members' as well as outsiders' welfares change as the coalition size increases and they do so non-monotonically. These complex pathways and the relationship between members' and outsiders' welfare are driven, as McGinty (2020) shows, by two effects: (i) the internalization of net emission damages through collective action, and (ii) the first mover advantage. McGinty (2020) along with a great part of the literature, including the recent contributions by Finus et al. (2021a) and (2021b), use abatement as the only choice variable. That is, countries' welfare is defined as the difference between aggregate abatement benefits and private abatement costs. In these models, the benefit function integrates two sources of welfare improvements from increased abatement: first, the direct effect of decreased environmental damages as net emissions decrease, and second, the increased benefits from higher emissions (production), which become possible as an indirect effect of increased abatement. Although these models capture changes in abatement costs, they are limited in attributing changes in benefits coming from decreased environmental damages (derived from decreased net emissions) versus increased production/emissions (derived from increased emissions).[1] On the other hand, models using emissions as the only choice variable, cannot capture changes in abatement cost. Such models explicitly assume that abatement technology is either non existing or prohibitively expensive.[2]

Given that abatement costs and environmental damages are independent from each other, yet essential in determining welfares, analyzing the impact of their changes on welfare separately, is critical. Keeping the benefit function (i.e. preferences) unchanged, we show that (i) when the abatement cost function is increasing at a slower pace large coalition, including the grand, are possible, and (ii) when the environmental damage function is increasing

[1] These models do not introduce constraints on the benefit function, which will otherwise be required to prevent negative production. See for example Barrett (1994) and more recently McGinty (2020).

[2] Assuming interior solutions, Diamantoudi and Sartzetakis (2006), derive a maximum stable coalition size of four, using quadratic functional forms, while larger coalition sizes become stable at the corner as shown by Ulph and Rubio (2006).

at a higher pace, slightly larger coalitions are possible and countries' net emission choices are forced at the corner. Thus, the only viable way in which significantly large coalitions can form is when abatement becomes really inexpensive. In such a case though it is also individually rational to increase abatement and attain the same welfare. Therefore, coalition formation adds very little if any, welfare improvements, which is not very paradoxical, when the choice variables are untwined.

Finally, as a corollary of our analysis, the differences between sequential and simultaneous moves in the second stage are easily explained. In the absence of leadership advantage, the damage internalization effect will increase with the size of the coalition, increasing thus the welfare of both members and outsiders monotonically. Therefore, the only stable coalition will be close to the Nash, and clearly smaller relative to the case in which the leadership effect is in action.[3]

The above results reflect the current wold situation with respect to climate change in that they capture the challenge countries face in forming coalitions. Indeed, (the perception of) environmental damages are continuously increasing while abatement costs remain high. To avoid deepening the environmental crisis, additional actions are required, such as technology transfers, linkages with trade and other agreements, among others, as the literature has already suggested.[4]

The rest of the paper is organized as follows. Section 2 lays out the model. Section 3 presents the results using general functional forms. Section 4 provides a specific example of quadratic benefit, damage and abatement cost functions and derives complete analytical solutions that enhance the explanation of the previous Section's results. It also provides some simulation results to highlight some important comparative statics. The last Section concludes the paper.

[3]The simultaneous moves game has been examined by Carraro and Siniscalco (1993), Finus and Rundshagen (2001) and Rubio and Casino (2001).

[4]It is well known that linking an agreement over a public good to an agreement over a club good, improves cooperation in the former. For the case of technology agreements, see for example, Barret (2006) and Hoel and de Zeeuw (2010). For the case of trade agreements see for example, Eichner and Pethig (2013) and (2015) and Diamantoudi et. al. (2020). Some of these tools have actually been used in successfully implementing the Montreal Protocol

2 The Model

We assume that there exist n symmetric countries, $N = \{1, \ldots, n\}$. Production and consumption activities in each country i, $i \in N$, yield benefits but they also generate emissions, $e_i > 0$, of a global pollutant. We take the production technology as given, i.e., we assume that the relationship between production and emissions remains constant, and, for simplicity, we further assume that it is a one-to-one relationship. Thus, we can express country i's benefits from production and consumption activities as a function of its emissions, $B_i(e_i)$. Each country suffers damages from aggregate emissions of the global pollutant, $E = \sum_{i=1}^{n} e_i$. Responding to the adverse effects of emissions country i could engage in emissions mitigation activities either by reducing production and consumption, or by engaging in end-of-the-pipe abatement $x_i \geqslant 0$.[5] While country i assumes the complete cost of its mitigation effort, the benefits are spread globally. Therefore, countries participate in a decision game with two choice variables: emissions, which create a negative externality, and abatement, which generate a positive externality.

Country i's social welfare, W_i, is defined as the total benefits country i receives from production and consumption activities, expressed in terms of the corresponding emissions, $B_i(e_i)$, minus the environmental damages, $D_i(NE)$, suffered from the aggregate global net emissions, $NE = E - X$, defined as the difference between global emissions E and global abatement $X = \sum_{i=1}^{n} x_i$. To complete the definition of country i's net benefits, we subtract the cost of abatement, $C_i(x_i)$. Country i's welfare function is,

$$W_i = B(e_i) - D(NE) - C(x_i).$$

We assume that $B(e_i)$ is strictly concave, that is, $B(0) = 0$, $B' \geq 0$ and $B'' < 0$. We further assume that environmental damages $D(NE)$ are strictly convex in net emissions, that is, $D(0) = 0$, $D'(NE) \geq 0$ and $D''(NE) > 0$. Finally, we assume that the cost of abatement is strictly convex, that is $C(0) = 0$, $C' \geq 0$ and $C'' > 0$.

[5] We model only end-of-the-pipe abatement while we do not consider cleaner production technologies, since the latter involve the introduction of new processes and products that take much longer to implement and are mostly market rather than policy driven (see Frondel et. al., 2007). We assume that at any given point in time negotiating countries take products and processes as given, while they can induce end-of-the-pipe abatement by increasing the stringency of environmental policy.

3 Stable coalition size using general functions

In this Section we derive the range of possible stable coalitions' sizes and we compare it to the corresponding range when countries' only choice variable is emissions, using general functional forms as defined in the previous Section. The literature uses one choice variable, either emissions or abatement. Some authors have suggested, directly or indirectly, that country i's single choice variable should be interpreted as net emissions, ε_i. Such an interpretation of the single choice variable though implies that, since net emissions are defined over both emissions and abatement, the benefit function embeds, in the case that the single choice variable is emissions, abatement costs, and in the case of abatement, environmental damages and benefits from emissions. Thus, the benefit function of the single variable model is not comparable to the benefit function of the multivariable model, albeit the same name, since the latter captures solely benefits from emissions. The results of the current paper can be compared only to those of papers in which countries choose either their emissions or abatement only. In the present paper the comparison is made to a model in which the single variable is emissions, assuming that abatement is not a choice, presumably either because of technological restrictions or of extremely high costs.

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 $|S| = s$, the coalition's aggregate net emissions by $NE_s = E_s - X_s$ while each member's net emissions are $\varepsilon_s = e_s - x_s$, such that $NE_s = s\varepsilon_s$. In a similar manner, each nonsignatory's net emissions are $\varepsilon_{ns} = e_{ns} - x_{ns}$, giving rise to a total emission level generated by all nonsignatories $NE_{ns} = E_{ns} - X_{ns} = (n-s)\varepsilon_{ns}$. The aggregate net emission level is, $NE = NE_s + NE_{ns} = s\varepsilon_s + (n-s)\varepsilon_{ns}$.

As mentioned in the Introduction, coalition formation is depicted as a two stage game: In the first. each country decides whether to join the climate coalition, and in the second, countries that form the climate coalition choose collectively both their emissions and abatement levels acting as a leader over free-riders. Following the literature, we define a coalition to be stable if no member wants to exit (internal stability) and no outsider wants to join (external stability) this coalition. In what follows, we solve the game backwards,

defining first nonsignatories' choices. Given the importance of their choices in determining the outcome of the game, we explain in detail the parameters that drive nonsignatories' choices and. in doing so. we also establish and clarify the distinct contribution of defining the model over two separate choice variables: emissions and abatement. Moving to the first stage, in order to determine the size of stable coalitions we use an indirect approach, first determining a critical size of the coalition and then link it to the stable coalition size.

3.1 Nonsignatories' choices

Nonsignatories behave non-cooperatively after observing the choice of signatories. Therefore, the maximization problem of nonsignatory country j, is,

$$\max_{e_{j,ns},x_{j,ns}} w_{j,ns} = B(e_{j,ns}) - D\left[s\varepsilon_s + (n-s-1)\varepsilon_{ns} + \varepsilon_{j,ns}\right] - C(x_{j,ns}).$$

The first order conditions are, $\frac{\partial B(e_{j,ns})}{\partial e_{j,ns}} = \frac{\partial D(NE)}{\partial NE}\frac{\partial NE}{\partial e_{j,ns}}$ and $\frac{\partial C(x_{j,ns})}{\partial x_{j,ns}} = -\frac{\partial D(NE)}{\partial NE}\frac{\partial NE}{\partial x_{j,ns}}$. From the definition of aggregate net emissions, $NE = s\varepsilon_s + (n-s-1)\varepsilon_{ns} + \varepsilon_{j,ns}$, we know that $\frac{\partial NE}{\partial e_{j,ns}} = 1$ and $\frac{\partial NE}{\partial x_{j,ns}} = -1$, and thus the first order conditions become , $\frac{\partial B(e_{j,ns})}{\partial e_{j,ns}} = \frac{\partial D(NE)}{\partial NE}$ and $\frac{\partial C(x_{j,ns})}{\partial x_{j,ns}} = \frac{\partial D(NE)}{\partial NE}$.

Since we assume symmetric countries, we can solve for the equilibrium values of any of the individual nonsignatory country's emissions and abatement, $e^*_{ns}(NE_s, s, x_{ns})$ and $x^*_{ns}(NE_s, s, e_{ns})$. Substituting back into the first order conditions at the equilibrium, nonsignatories' emission and abatement satisfy the conditions,

$$B'(e^*_{ns}(NE_s, s, x_{ns})) = \left.\frac{\partial D(NE)}{\partial NE}\right|_{e_{ns}=e^*_{ns}(NE_s,s,x_{ns})}, \tag{1}$$

$$C'(x^*_{ns}(NE_s, s, e_{ns})) = \left.\frac{\partial D(NE)}{\partial NE}\right|_{x_{ns}=x^*_{ns}(NE_s,s,e_{ns})}. \tag{2}$$

Denoting $D'(NE) = \frac{\partial D(NE)}{\partial NE}$, the first order conditions can be written as $B'(e^*_{ns}(NE_s, s, x_{ns})) = D'(NE)|_{e_{ns}=e^*_{ns}(NE_s,s,x_{ns})}$ and $C'(x^*_{ns}(NE_s, s, e_{ns})) = D'(NE)|_{x_{ns}=x^*_{ns}(NE_s,s,e_{ns})}$. Solving the system of (1) and (2) yields the best response functions $Re = e^*_{ns}(NE_s, s)$ and $Rx = x^*_{ns}(NE_s, s)$.

3.1.1 The contribution of introducing two separate choice variables

In order to clarify the contribution of defining countries' emissions and abatement as separate choice variables, we examine nonsignatories' first order conditions. This will allow us also to establish and clarify the difference between this paper's modelling approach and single choice variable specifications. We choose as vehicle of achieving both goals the diagrammatic illustration of nonsignatories' first order conditions evaluated at the equilibrium values, because of the simple, clear and intuitive messages they convey. The left part of Figure 1, illustrates nonsignatory's first order condition with respect to emissions, while the right part the first order condition with respect to abatement. The light blue curve in the left panel, illustrates the individual nonsignatory's marginal benefits $B'(e_{ns})$, assumed positive and decreasing in e_{ns}, i.e., $B' > 0$ and $B'' < 0$. Marginal abatement costs, $C'(x_{ns})$, depicted by the, relatively steeper, light green curve in the right panel, which starts at zero abatement, are assumed positive and increasing in x_{ns}, i.e. $C(0) = 0$, $C' > 0$ and $C'' > 0$.

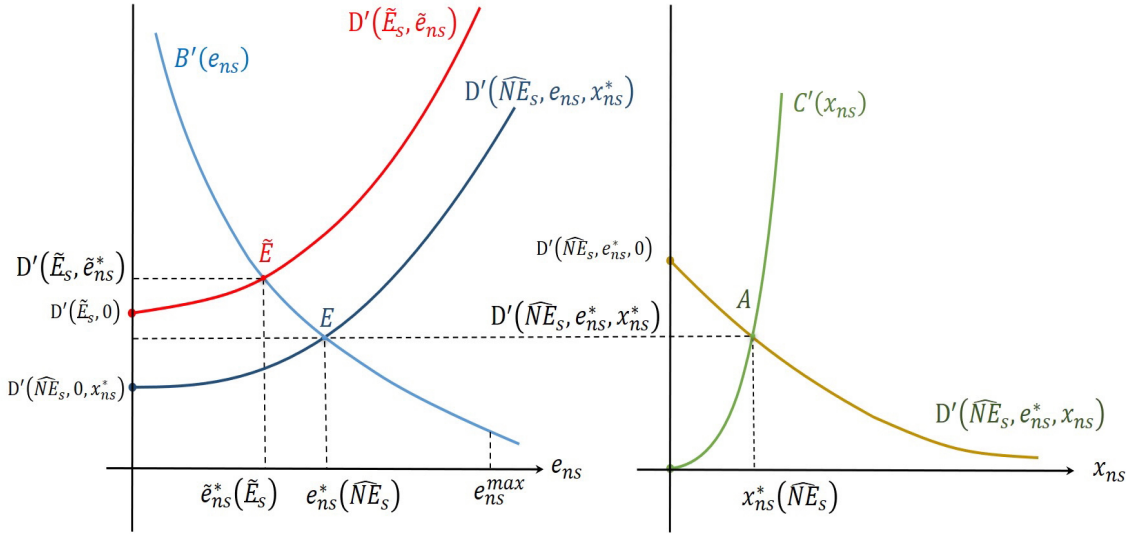


Figure 1: Nonsignatories first order conditions, in the single and two choice variables mode

The particular marginal damage curves illustrated in the two panels, correspond to an arbitrarily given level of signatories' net emissions, $\widehat{NE}_s$, which,

by the solution of signatories' problem, corresponds to a coalition size, $\widehat{s}$. The solid dark blue line in the left panel, depicts the individual nonsignatory's marginal damages with respect to emissions, evaluated at $\widehat{NE}_s$, and also at the equilibrium x^*_{ns}. This curve is increasing in e_{ns}, since we assume $D'' > 0$ and e_{ns} affects NE positively. Its vertical intercept, $D'(\widehat{NE}_s, 0, x^*_{ns})$, is positive, since both $\widehat{NE}_s$ and x^*_{ns} are positive. The brown solid curve in the right panel depicts the individual nonsignatory's marginal damages with respect to abatement, evaluated at $\widehat{NE}_s$ and also at e^*_{ns}. This curve is decreasing in x^*_{ns}, since $D'' > 0$ and x_{ns} affects NE negatively. Its vertical intercept, $D'(\widehat{NE}_s, e^*_{ns}, 0)$, is positive, since both $\widehat{NE}_s$ and e^*_{ns} are positive. Notice that for given $\widehat{NE}_s$, $D'(\widehat{NE}_s, 0, x^*_{ns}) < D'(\widehat{NE}_s, e^*_{ns}, 0)$, since x_{ns} decreases NE, while e_{ns} increases NE. In between these values lies $D'(\widehat{NE}_s, 0, 0)$, the value of marginal damages at zero equilibrium levels of nonsignatories emissions and abatement. Given that, apart from the sign by which e_{ns} and x_{ns} enter NE, which determines whether the slope of the marginal damage curve is positive or negative, the damage function with respect to aggregate net emissions is the same, and thus, the slope of both marginal damage curves is the same.

Given that damage functions are illustrated for the equilibrium values of the other variable,[6] the intersection points, E and A, illustrate the equilibrium values for both choice variables, $e^*_{ns}(\widehat{NE}_s)$ and $x^*_{ns}(\widehat{NE}_s)$, resulting from the solution of the two first order conditions, (1) and (2). At the equilibrium, for the given level of signatories' net emissions $\widehat{NE}_s$, the marginal damage with respect to emissions and with respect to abatement are equal at $D'(\widehat{NE}_s, x^*_{ns}, e^*_{ns})$. It is through marginal damages that nonsignatories' marginal benefit is equated to marginal abatement cost at the equilibrium achieving efficiency in distributing mitigation efforts between emissions reduction and abatement increase. Measuring NE_s on a third axis and evaluating $e^*_{ns}(\cdot)$ and $x^*_{ns}(\cdot)$ for all possible values of NE_s, would allow us to plot nonsignatories' emission and abatement as functions of signatories net emissions, i.e. the nonsignatories' reaction planes.

In the case that emissions is the single choice variable the right part of Figure 1 does not exist. Nonsignatories set marginal benefits, depicted

[6]For example, the marginal benefit curve $D'(\widehat{NE}_s, e_{ns}, x^*_{ns})$, is drawn for values of e_{ns} evaluated at x^*_{ns}.

again by the light blue curve in the left panel, equal to marginal damages $D'(\widetilde{\underline{E}}_s, \widetilde{e}_{ns})$, evaluated at a given value of signatories' emissions, $\widetilde{\underline{E}}_s$, which is depicted by the red curve. Assuming the exact same benefit function as in the two choice variable model, the red curve is an upward shift of the dark blue curve, the two differing only in their vertical intercept. In order to compare the two models, we use the exact same functions for benefits and damages and furthermore, we assume $\widetilde{\underline{E}}_s = \widehat{NE}_s$ in order to be able to illustrate the two models in the same graph. The tilde indicates emission levels derived from the single variable model and the underline in $\widetilde{\underline{E}}_s$ indicates that it is a specific value of signatories ' total emissions. The vertical intercept, $D'(\widetilde{\underline{E}}_s, 0)$, is positive since $\widetilde{\underline{E}}_s > 0$, and lies in between the vertical intercepts of the marginal damage with respect to emissions and with respect to abatement in the two choice variable model, that is, $D'(\widehat{NE}_s, 0, x^*_{ns}) \leq D'(\widetilde{\underline{E}}_s, 0) < D'(\widehat{NE}_s, e^*_{ns}, 0)$, for $x^*_{ns} \geq 0$ and $e^*_{ns} > 0$. If abatement cost becomes prohibitively expensive, yielding $x^*_{ns} = 0$, then $D'(\widehat{NE}_s, 0, 0) = D'(\widetilde{\underline{E}}_s, 0)$, since we assumed $\widetilde{\underline{E}}_s = \widehat{NE}_s$, and the curves depicting marginal damage with respect to emissions in the two, and the one, choice variable models become exactly the same. Thus, we establish a direct comparison to the model with emissions as the only choice variable, assuming abatement is not a choice.

Such a direct comparison with a model using abatement as the single choice variable is not possible. This is because in such a model, with country i's welfare defined as $W_i = B(X) - C(x_i)$, environmental damages and direct benefits from production are not explicitly modeled. They are both embedded in benefits from aggregate (hence the externality) abatement. Diamantoudi and Sartzetakis (2006) has shown the analogy between the single variable models, using quadratic functional forms and relating emissions to abatement through a maximum level of emissions occurring before abatement becomes and option, i.e., $e_i = e_i^{\max} - x_i$. Direct comparison with the model using abatement as a single choice variable would require some heroic assumptions, and thus, we do not further explore it.

Our two choice variable model allows us to perform comparative statics to illustrate the role of preferences, environmental damages and abatement technology in determining nonsignatories' reaction functions. Because of its importance we choose to illustrate, in Figure 2, the effect of changes in

marginal abatement cost. The only difference relative to Figure 1 is the decreased slope of the marginal abatement cost curve. Following the change in abatement cost, there will be a series of adjustments in the choice of emissions and abatement, resulting in the new equilibrium depicted in Figure 2.

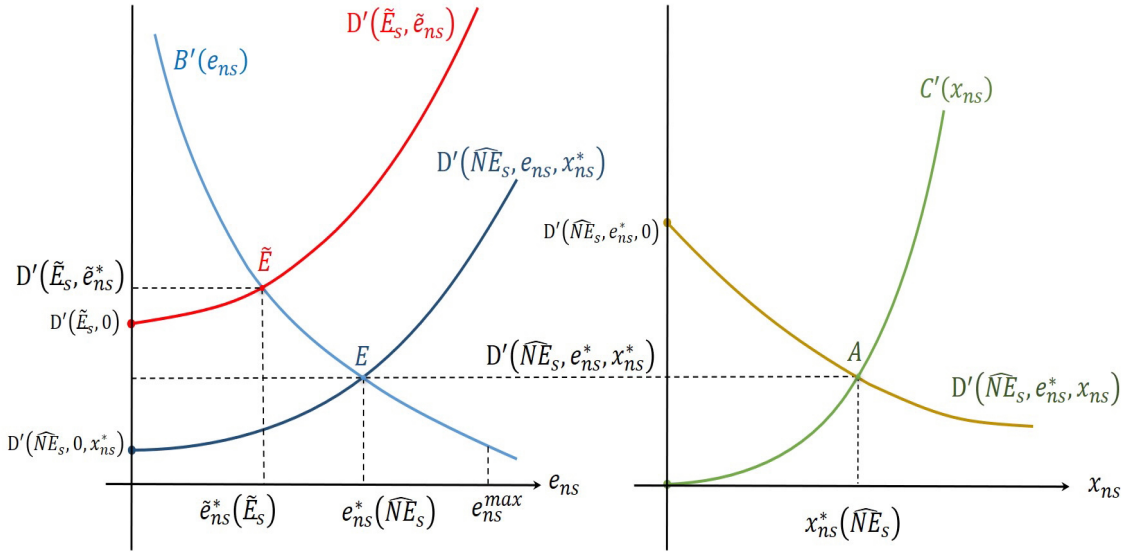


Figure 2: The effect of decreasing abatement costs

At this new equilibrium, nonsignatories choose a higher level of abatement, and thus the intercept of the marginal damage with respect to emissions curve , $D'(\widehat{NE}_s, 0, x^*_{ns})$, in the left panel, moves downwards away from $D'(\widetilde{\underline{E}}_s, 0)$. They also choose a higher level of emissions, which results to a higher vertical intercept for the marginal damage with respect to abatement curve, $D'(\widehat{NE}_s, e^*_{ns}, 0)$, in the right panel, again away from $D'(\widetilde{\underline{E}}_s, 0)$. Both abatement and emissions increase, however, the increase in abatement is higher relative to that in emissions, since we did not change the slope of the marginal benefit curve, leading to a lower level of nonsignatories' net emissions. This is evident by the fact that marginal damages in the new equilibrium are lower relative to the case with steeper marginal abatement cost, depicted in Figure 1. The lower bound of nonsignatories' net emissions is the level of their emissions if abatement is not a choice, that is, $\varepsilon^*_{ns}(\widehat{NE}_s) \leq \widetilde{e}^*_{ns}\left(\widetilde{\underline{E}}_s\right) \leq e^*_{ns}(\widehat{NE}_s)$, with equality when $x^*_{ns} = 0$. The upper

bound of emissions would correspond to the level that maximizes total benefits, accompanied by an equal amount of abatement, which is feasible when the cost of abatement becomes trivial. Therefore, a technological improvement in abatement will allow higher levels of abatement, permitting higher emission level and thus, higher total benefits, without though jeopardizing environmental goals since net emissions decrease, yielding lower marginal damages.

In a similar way we can perform comparative statics for changes in marginal damages and/or marginal benefits. If new evidence suggest higher damages, leading to a steeper marginal damage curve, ceteris paribus, nonsignatories' choice of emissions will decrease while abatement will increase. Given the increase in abatement, the distance of the marginal damage intercepts between the single and the two variable models, will increase, yielding a much higher difference in nonsignatories' choice of emissions. The single choice model misses the abatement adjustment effect on marginal damages. Similarly, a decrease in the slope of marginal benefits, ceteris paribus, apart from the direct effect of increasing emissions, shown in the left part of the Figures, will also have the indirect effect of increasing abatement, since in the right panel the intercept of marginal damages with respect to abatement will increase. In this case, the increase in emissions will outweigh the increase in abatement, since the slope of the marginal abatement cost curve did not change, leading to higher net emissions and thus, marginal damages. Clearly, the single variable model cannot capture such indirect effects and adjustments in countries' composition of mitigation effort.

3.1.2 Nonsignatories' reaction functions

Return back to the first order conditions (1) and (2) and differentiate both with respect to $NE_s = s\varepsilon_s$, to obtain the slope of nonsignatories' reaction functions around their equilibrium values,

$$Re' = \frac{\partial e^*_{ns}\left(NE_s\right)}{\partial NE_s} = \frac{-1}{\left(n-s\right)\left(1-\psi\right)-\phi}, \tag{3}$$

$$Rx' = \frac{\partial x^*_{ns}\left(NE_s\right)}{\partial NE_s} = \frac{-\psi}{\left(n-s\right)\left(1-\psi\right)-\phi}, \tag{4}$$

where $\phi(s) \equiv \frac{B''(e^*_{ns}(NE_s(s)))}{D''(NE(s))} < 0$ and $\psi(s) \equiv \frac{B''(e^*_{ns}(NE_s(s)))}{C''(x^*_{ns}(NE_s(s)))} < 0$, given that we have assumed, $B'' < 0$, $D'' > 0$, and $C'' > 0$. Since the denominator of both (3) and (4) is positive, the slope of nonsignatories best response functions take values in the following ranges: $\frac{-1}{(n-s)} < Re' \leqslant 0$ and $1 > Rx' \geqslant 0$. Responding to an increase in signatories' net emissions, nonsignatories decrease their emissions and increase their abatement. Note that for $|\psi| > 1$, that is, when $|B''(e^*_{ns}(NE_s))| > |C''(x^*_{ns}(NE_s))|$, the speed of abatement adjustment exceeds that of emissions. Thus, if the slope of marginal benefits exceeds the slope of marginal abatement cost, nonsignatories adjust to a marginal increase in signatories emissions by increasing abatement more than decreasing emissions and vise versa. Combining equations (3) and (4) yields the slope of the best response of nonsignatories' net emissions $R\varepsilon' = \frac{\partial(e^*_{ns}(NE_s) - x^*_{ns}(NE_s))}{\partial NE_s} = \frac{-(1-\psi)}{(n-s)(1-\psi)-\phi} < 0$, given that both ϕ and ψ are negative.

Before turning to signatories' choices, we examine the effect that changes in the slope of marginal damages, $D''(NE)$, and marginal abatement cost, $C''(x^*_{ns}(E_s, X_s))$ have on the slope of nonsignatories' reaction functions, holding preferences constant. If new data and scientific discoveries regarding damages from net emissions indicate a steeper marginal damage curve, i.e., a higher $D''(NE)$, the absolute value of ϕ decreases, yielding a decrease in the denominator of both (3) and (4) and thus, both emissions and abatement reaction functions become more responsive. The overall effect of a change in $D''(NE)$ on the absolute value of $R\varepsilon'$ is positive since $\frac{\partial|R\varepsilon'|}{\partial\phi} = \frac{1-\psi}{[(n-s)(1-\psi)-\phi]^2} > 0$. On the other hand, if investment in abatement reduces the slope of marginal abatement cost, i.e., a lower $C''(x^*_{ns}(E_s, X_s))$, the absolute value of ψ increases and thus, the less responsive the emission reaction function becomes but the more responsive the abatement reaction function becomes. A decrease in $C''(x^*_{ns}(E_s, X_s))$ makes nonsignatories' net emissions reaction function more responsive since $\frac{\partial|R\varepsilon'|}{\partial\psi} = \frac{\phi}{[(n-s)(1-\psi)-\phi]^2} < 0$. Therefore, an increase in the slope of marginal damages or a decrease in the slope of marginal abatement cost make nonsignatories more responsive to changes in signatories choice of net emissions.

3.2 Signatories' choices

Signatories maximize the coalition's welfare, sw_s, taking explicitly into account nonsignatories' behaviour. The coalition's maximization problem yields an indirect welfare function ω_s as follows:

$$w_s(s) = \frac{1}{s} \max_{e_s, x_s} \left[sB\left(e_s\right) - sD\left[s\varepsilon_s + (n-s)\varepsilon^*_{ns}\left(NE_s, s\right)\right] - C(x_s)\right].$$

The signatories' emissions $e^*_s(s)$ and abatement $x^*_s(s)$ at the equilibrium satisfy the following conditions,

$$B'(e^*_s(s, x_s)) = \frac{\partial D(NE)}{\partial NE} \left.\frac{\partial NE}{\partial e_s}\right|_{e_s = e^*_s(s)}, \tag{5}$$

$$C'(x^*_s(s, e_s)) = \frac{\partial D(NE)}{\partial NE} \left.\frac{\partial NE}{\partial x_s}\right|_{x_s = x^*_s(s)}. \tag{6}$$

where, $NE = s\left(e^*_s(s, x_s) - x^*_s(s, x_s)\right) + (n-s)\left(e^*_{ns}(e_s, x_s, s) - x^*_{ns}(e_s, x_s, s)\right)$. The derivatives of aggregate net emissions around the equilibrium values $e^*_s(s)$ and $x^*_s(s)$ are: $\left.\frac{\partial NE}{\partial e_s}\right|_{e_s = e^*_s(s)} = s + (n-s)\frac{\partial(e^*_{ns}(e_s, x_s, s) - x^*_{ns}(e_s, x_s, s))}{\partial e_s}$ and $\left.\frac{\partial NE}{\partial x_s}\right|_{x_s = x^*_s(s)} = -s + (n-s)\frac{\partial(e^*_{ns}(e_s, x_s, s) - x^*_{ns}(e_s, x_s, s))}{\partial x_s}$.

The above two first order conditions can be plotted in a similar way as the nonsignatories' first order conditions in Figures 1 and 2. The corresponding points E and A, will illustrate the equilibrium values of emissions and abatement for a given size $\widehat{s}$ of the coalition, i.e. $e^*_s(\widehat{s})$ and $x^*_s(\widehat{s})$, for which the diagram will be plotted. Plotting s in a third axis, we can illustrate the equilibria points E and A for any value of s, that is, signatories' optimal levels of emission and abatement as functions of the coalition size, $e^*_s(s)$ and $x^*_s(s)$, derived as the solution of the system of equations (5) and (6). Since benefit, damage and abatement cost functions are the same for both signatories and nonsignatories, following the same argument as in the case of nonsignatories, we can establish that $\varepsilon^*_s(\widehat{s}) \leq \widetilde{e}^*_s(\widehat{s}) \leq e^*_s(\widehat{s})$, where the tilde continues to depict values from the single choice variable model. The equality holds for $x^*_s = 0$.

3.3 Size of stable coalitions

Substituting $e^*_s(s)$ and $x^*_s(s)$ back into nonsignatories' reaction functions, we obtain nonsignatories' choices as functions of the coalition's size only,

i.e., $e^*_{ns}(NE^*(s), s)$ and $x^*_{ns}(NE^*(s), s)$. Finally, substituting the equilibrium values of the choice variables into the welfare functions, we obtain signatory and nonsignatory countries' indirect welfare as functions of coalition's size, i.e., $w_s(s)$ and $w_{ns}(s)$.

As to define stable coalitions, the literature uses the notion of internal and external stability. Formally, a coalition of size s^* is,

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

Given that it proves impossible to determine s^* directly from the above conditions, we follow an indirect path. We first determine coalition's critical size $z^{\min}$, at which members and non-members' choices are exactly the same, and then we link this critical size to the size of the stable coalition.

Proposition 1 presents some important properties of countries' choice variables and indirect welfare with respect to the critical value of the coalition's size $z^{\min}$. We establish that there exists a critical coalition size, below which signatories emit more, abate less and attain higher welfare than the nonsignatories and above which the reverse is true. This critical size is determined by adjusting, to the lower integer, the value of $z^{\min}$, where $z^{\min}$ denotes the intersection of e_s with e_{ns} and of x_s with x_{ns} and thus, of w_s with w_{ns} which lies right at the minimum value of w_s. We denote by e_{nc}, x_{nc} and E_{nc} the individual country's emissions and abatement and the aggregate emissions respectively, in the purely non-cooperative case, where there is no leader and firms compete a la Cournot. To simplify the exposition we use the notation ϕ and ψ defined above.

Proposition 1 *Consider the indirect welfare functions of signatory and nonsignatory countries, $w_s(s)$ and $w_{ns}(e^*_s(s), x^*_s(s), s)$ respectively. Let*

$$z^{\min} = \frac{n\left(1-\psi\left(z^{\min}\right)\right) - \phi\left(z^{\min}\right)}{1-\phi\left(z^{\min}\right) - \psi\left(z^{\min}\right)}$$

then,

1. $e^*_s(s) \gtreqless e^*_{ns}(s) \Leftrightarrow s \gtreqless z^{\min}$ *and* $x^*_s(s) \gtreqless x^*_{ns}(s) \Leftrightarrow s \gtreqless z^{\min}$,
2. *if* $s = z^{\min}$ *then* $e^*_s(s) = e^*_{ns}(s) = e_{nc}$ *and* $x^*_s(s) = x^*_{ns}(s) = x_{nc}$,
3. $w_s(s)$ *increases (decreases) in* s *if* $s > z^{\min}$ $(s < z^{\min})$,

4. $z^{\min} = \arg\min_{s \in \Re \cap [0,n]} w_s(s)$,
5. $w_s(s) \gtreqless w_{ns}(e_s^*(s), x_s^*(s), s) \Leftrightarrow s \lesseqgtr z^{\min}$.

Although the discussion of the Proposition's intuition that follows sketches its proof, the Proposition's proper proof is presented in Appendix I. To discuss the intuition of the properties of countries' choice variables presented in Proposition 1, recall that what differentiates signatories' and nonsignatories' behaviour is that each coalition member internalizes the damages its net emissions inflict on all other members, while in the same time they collectively exercise their leadership power over free-riders. These two effects work in opposite directions as the coalition size changes; hence there will be a coalition size at which they offset each other, and thus, the coalition's members' choices become the same as those of free-riders.

Formally, nonsignatories in choosing e_{ns}^* and x_{ns}^* according to (1) and (2), they set $\frac{\partial NE}{\partial e_{ns}} = 1$ and $\frac{\partial NE}{\partial x_{ns}} = -1$, while the signatories, in choosing e_n^* and x_n^* according to (5) and (6), set $\frac{\partial NE}{\partial e_s} = s + (n-s)\frac{\partial (e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s))}{\partial e_s}$ and $\frac{\partial NE}{\partial x_s} = -s + (n-s)\frac{\partial (e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s))}{\partial x_s}$.[7] Given that we assume completely homogeneous countries, signatories and nonsignatories' first order conditions differ only in the response of NE to their choice variables. The slope of the reaction function of all $(n-s)$ nonsignatories' net emissions to changes in one of the signatories' emissions, using (3) and (4) and noting that here we differentiate with respect to e_s and not E_s, is $(n-s)sR\varepsilon' = \frac{-s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$. Therefore, the difference between signatories and nonsignatories optimization conditions, is $(s-1) - \frac{s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$.

A marginal increase (decrease) in one of the signatories' emissions (abatement) increases marginal damages to all other coalition members and thus, the term $(s-1)$ captures the damage internalization effect: The increase in each coalition's member marginal damage coming indirectly from internalizing the marginal damages that all other $(s-1)$ coalition members suffer from a marginal increase (decrease) in that member's emissions (abatement). At the same time, nonsignatories respond to the increase (decrease) in one of the signatories' emissions (abatement) by decreasing their net emissions on

[7] To simplify the exposition, from here on we denote the derivatives at the equilibrium values $\frac{\partial NE}{\partial e_s}\Big|_{e_s = e_s^*(s)}$ and $\frac{\partial NE}{\partial x_s}\Big|_{x_s = x_s^*(s)}$ simply by $\frac{\partial NE}{\partial e_s}$ and $\frac{\partial NE}{\partial x_s}$ and similarly for nonsignatories.

aggregate by $\frac{s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$. This term captures the leadership effect: the decrease in each coalition member's marginal damage coming indirectly through the response of all nonsignatories' net emissions to a marginal increase (decrease) in that member's emissions (abatement). Equating these two effects and solving yields the critical value $z^{\min} = \frac{n(1-\psi)-\phi}{1-\phi-\psi}$. Note that $z^{\min} > 1$[8] and as ϕ decreases an/or ψ increases, that is, as D'' increases and/or C'' decreases relative to B'', the value of $z^{\min}$ increases. The intuition of this result becomes clear if we recall that an increase in D'' and/or a decrease in C'' increases the responsiveness of the non-signatories' net emissions reaction function.

To summarize the above discussion, coalition members, in collectively choosing their net emissions, take into account the damages that each member's net emissions impose on all members, and at the same time, exploit their leadership advantage. The size of these two effects on coalition members' choice of net emissions depends on the size of the coalition. When the coalition has few members, the first effect is smaller since lower aggregate damages are internalized, while the second effect is larger, since the coalition makes gains out of a larger number of free-riders. At the critical size of the coalition, $z^{\min}$, these two effects offset each other. For coalition sizes higher than $z^{\min}$, the reverse is true and thus, the welfare of nonsignatories exceeds that of signatories. For given benefits, the higher the damages and/or the lower the abatement costs, the more responsive nonsignatories' reaction function, yielding a larger $z^{\min}$.

The above results are established in the literature (McGinty, 2020 and Finus et al., 2021) in the case of a single choice variable. Apart from extending the results to general functional forms, providing a formal proof, the main contribution of this paper is that, by allowing countries to choose abatement separately from emissions, we can separate and identify the effect of preferences, environmental damages and abatement technology on countries' equilibrium choices. The difference between the two approaches were illustrated in Figures 1 and 2, considering the first order conditions, and will further extended to the determination of the critical coalition size $z^{\min}$ and the stable coalition size in what follows.

[8]For $z^{\min} > x$, where x is a positive number, $(n-x) > \frac{-(x-1)\phi}{1-\psi}$ which is true for $x = 1$, since $n > 1$, $\phi < 0$ and $\psi < 0$.

3.3.1 Comparison of the critical coalition size to the model with emissions as the only choice variable

Denote by $\widetilde{z}^{\min}$ the critical value of the size of the coalition in the case that emissions is the only choice variable and abatement is not an option, i.e., $x = 0$. Diamantoudi and Sartzetakis (2015) show that in such case the value of the critical coalition size is $\widetilde{z}^{\min} = \frac{n-\widetilde{\phi}}{1-\widetilde{\phi}}$, where here $\widetilde{\phi} \equiv \frac{B''\left(\widetilde{e}^*_{ns}\left(\widetilde{E}_s\right)\right)}{D''\left(\widetilde{E}\right)}$. Similar to the two choice variables case, discussed above, at $s = \widetilde{z}^{\min}$, $\widetilde{e}^*_s(s) = \widetilde{e}^*_{ns}(s) = \widetilde{e}_{nc}$, and thus, $\widetilde{\phi}\left(\widetilde{z}^{\min}\right) = \frac{B''(\widetilde{e}_{nc})}{D''\left(\widetilde{E}_{nc}\right)}$. In the case of the two choice variables model, $\phi\left(z^{\min}\right) = \frac{B''(e_{nc})}{D''(NE_{nc})}$ and $\psi\left(z^{\min}\right) = \frac{B''(e_{nc})}{C''(x_{nc})}$. Following the same reasoning as above, at the full non-cooperation case, for $x_{nc} \geq 0$ we have $e_{nc} \geq \widetilde{e}_{nc}$ and $NE_{nc} \leq \widetilde{E}_{nc}$. Direct comparison of $z^{\min}$ to $\widetilde{z}^{\min}$ yields, $z^{\min} \geq \widetilde{z}^{\min} \Longrightarrow (n-1)\left[\left(\phi - \widetilde{\phi}\right) + \psi\widetilde{\phi}\right] \geq 0$, which, since $n > 1$, is true if $\left(\phi - \widetilde{\phi}\right) + \psi\widetilde{\phi} \geq 0$. Taking into account that ϕ, $\widetilde{\phi}$ and ψ are all negative, rearranging terms yields, $1 - \psi\left(z^{\min}\right) \geq \frac{\left|\phi\left(z^{\min}\right)\right|}{\left|\widetilde{\phi}(\widetilde{z}^{\min})\right|}$. A sufficient condition for this inequality to hold is $\left|\frac{B''(\widetilde{e}_{nc})}{D''\left(\widetilde{E}_{nc}\right)}\right| \geq \left|\frac{B''(e_{nc})}{D''(NE_{nc})}\right|$. Although we know that $e_{nc} \geq \widetilde{e}_{nc}$ and $NE_{nc} \leq \widetilde{E}_{nc}$ comparing the slopes of marginal benefit and marginal damage at different values of emissions would require assumptions on the third order derivatives. For example, it definitely holds if $B''' > 0$ and $D''' < 0$, since then $|B''(\widetilde{e}_{nc})| > |B''(e_{nc})|$ and $D''(\widetilde{e}_{nc}) > D''(e_{nc})$. In the case that $B''' = D''' = 0$, examined latter as an example, the condition again holds holds and thus, $z^{\min} \geq \widetilde{z}^{\min}$. However, if the third order derivatives have the exact different signs, then, $\frac{\left|\phi\left(z^{\min}\right)\right|}{\left|\widetilde{\phi}(\widetilde{z}^{\min})\right|} > 1$, and depending on the value of $\left|\psi\left(z^{\min}\right)\right|$, that is, the value of $C''(x_{nc})$, the condition $1-\psi\left(z^{\min}\right) \geq \frac{\left|\phi\left(z^{\min}\right)\right|}{\left|\widetilde{\phi}(\widetilde{z}^{\min})\right|}$ could, theoretically, be violated.

Return to the sufficient condition and examine the difference between $z^{\min}$ and $\widetilde{z}^{\min}$ at the extreme values of the abatement cost holding benefits and damages unchanged. If abatement becomes prohibitively expensive, so that the marginal abatement cost curve becomes almost vertical, i.e. $C''(x_{nc}) \to \infty$, then $\left|\psi\left(z^{\min}\right)\right| \to 0$. In such case, as explained above, $\widetilde{e}_{nc} = e_{nc}$, which implies $\left|\phi\left(z^{\min}\right)\right| = \left|\widetilde{\phi}\left(\widetilde{z}^{\min}\right)\right|$, since we use the same functions of benefits and damages in the two models, Thus, we have, $\lim_{C''\to\infty} z^{\min} \to \widetilde{z}^{\min}$.

At the other extreme, if abatement becomes very cheap, so that $C''(x_{nc}) \to 0$, then $\left|\psi\left(z^{\min}\right)\right| \to \infty$. In this case, $z^{\min} > \widetilde{z}^{\min}$. Therefore, the critical point is the same for the two models when abatement cost is extremely high and as abatement becomes cheaper, the critical point in the two choice variable model becomes larger relative to the single variable choice model. As abatement cost approaches zero, the critical coalition size $z^{\min}$ approaches n, i.e., $\lim_{C''\to 0} z^{\min} \to n$.[9] The following Proposition summarizes the comparison.[10]

Proposition 2 *The critical size of the coalition $z^{\min}$, at which emissions and abatement choices of signatories and non-signatories are equal, exceeds the corresponding critical size of the coalition, $\widetilde{z}^{\min}$, when emissions is the only choice variable if $B''' > 0$ and $D''' < 0$, or $B''' = D''' = 0$.*
Furthermore, the difference between $z^{\min}$ and $\widetilde{z}^{\min}$ is increasing as the cost of abatement decreases:
(i) As the marginal cost of abatement becomes very steep, that is, $C'' \to \infty$, the critical size of the coalition $z^{\min}$ tends to $z^{\min}_{x=0}$: $\lim_{C''\to\infty} z^{\min} \to \widetilde{z}^{\min}$.
(ii) As the marginal cost of abatement becomes very flat, that is, $C'' \to 0$, the critical size of the coalition $z^{\min}$ tends to n: $\lim_{C''\to 0} z^{\min} \to n$.

To discuss the intuition of the above results, we compare the slopes of nonsignatories' reaction functions in the case that abatement is and in the case it is not an option. Recall that the slope of the best response of nonsignatories' net emissions is $R\varepsilon' = \frac{-(1-\psi)}{(n-s)(1-\psi)-\phi}$. In the case that emissions is the only choice variable, the slope of nonsignatories' best response function is, $R\widetilde{\varepsilon}' = \frac{\partial \widetilde{e}^{*}_{ns}\left(\widetilde{E}_s\right)}{\partial \widetilde{E}_s} = \frac{-1}{(n-s)-\widetilde{\phi}} < 0$.[11] Therefore, as abatement becomes relatively cheap, that is, as C'' decreases, $|R\varepsilon'| > \left|R\widetilde{\varepsilon}'\right|$, which implies that nonsignatories respond to an increase in the coalition's net emissions by decreasing their own net emissions faster than in the absence of the abatement option. As a

[9] It should be noted that, as discussed in detailed in the next Section, $\widetilde{z}^{\min}$ could potentially approach n as well, if $D''\left(\widetilde{E}\right) \to \infty$. However, for given benefit function, as marginal damages become steeper, emissions, the only choice variable, would have to be reduced, reducing also benefits. Eventually, restrictions have to be applied to the slope of marginal damages (relative to the slope of marginal benefits), in order to secure positive emissions.

[10] The proof of Proposition 2 is relegated to Appendix II.

[11] See Diamantoudi and Sartzetakis (2015), p. 543.

result, the coalition's position as a leader is strengthened, which implies that the leadership effect outweighs the damage control effect for larger coalition sizes relative to the case that abatement is not an option. This effect becomes more prominent as the cost of abatement decreases. That is, the smaller is the cost of abatement, the higher is the absolute value of ψ and thus, the higher is the difference $z^{\min} - z^{\min}_{x=0}$.

3.4 Comparison of stable coalition sizes

The critical coalition size $z^{\min}$, determined in Proposition 1, can be used to define the lowest size of a stable coalition. Furthermore, using the comparison between $z^{\min}$ and $\widetilde{z}^{\min}$, established in Proposition 2, we can compare the size of the stable coalition between the case of two choice variables and the case of only one choice variable. As mentioned earlier, in order to define stable coalitions, the literature uses the notion of internal and external stability. Formally, a coalition of size s^* is,

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

As established in Proposition 1, there exists a critical value $z^{\min}$, below which signatories' net emissions and welfare are higher relative to nonsignatories, and above which the reverse is true. Countries' welfare attains its lowest value in the absence of a coalition, the Cournot Nash level w_{nc}. As small size coalitions start forming, their members' net emissions exceed those of nonsignatories, $\varepsilon_s(s) > \varepsilon_{nc}(s) > \varepsilon_{ns}(s)$, and thus, signatories' welfare level exceeds that of nonsignatories, $w_s(s) > w_{ns}(s)$, with $w_{ns}(s)$ dropping below w_{nc}. As the coalition size increases, their members' net emissions decrease relieving the pressure from nonsignatories whose welfare starts increasing while that of signatories decreases. At the critical size $z^{\min}$, net emissions of both signatories and nonsignatories, and thus their welfare level, return to their Cournot Nash levels. For higher coalition sizes nonsignatories' welfare function is always above that of signatories. Although we are able to determine that $w_s(s)$ is monotonically increasing in s after $z^{\min}$, the same is not possible for $w_{ns}(s)$.

If one member exits the coalition of size s, the welfare of nonsignatories

becomes $w_{ns}(s-1)$, which is a function with the same properties as $w_{ns}(s)$ shifted by one. That is, the $w_{ns}(s-1)$ function will cut from below the $w_s(s)$ function at a coalition size higher than $z^{\min}$. Adjusting this critical size to the lower integer yields the size of a stable coalition, denoted by s_1^*, since below that $w_s(s_1^*) \geq w_{ns}(s_1^*-1)$, and above that $w_{ns}(s_1^*) \geq w_s(s_1^*+1)$. For general functional forms, it is impossible to either define the value of s_1^* or claim its uniqueness: Depending on the slope of $w_s(s)$ around $z^{\min}$, the size of stable coalition s_1^* could be very close of far away from $z^{\min}$. Furthermore, given that we cannot determine that $w_{ns}(s)$ is monotonically increasing, the $w_{ns}(s-1)$ curve could intersect from below $w_s(s)$ also at higher coalition sizes. This is the reason why the literature uses specific functional forms to determine the size of stable coalitions. The next Section provides an example. However, we can discuss the effect that changes in the benefit, damage and abatement cost have on the smallest size of a stable coalition, if there are more than one, utilizing the results of Proposition 2, and thus, we can define the range of the stable coalitions' size.

It is evident, from the expressions $z^{\min} = \frac{n(1-\psi)-\phi}{1-\psi-\phi}$ and $\widetilde{z}^{\min} = \frac{n-\widetilde{\phi}}{1-\widetilde{\phi}}$, that if abatement is not an option, or an extremely expensive one, driving ψ to zero, the only determinant of $\widetilde{z}^{\min}$ is the ration $\widetilde{\phi}\left(\widetilde{z}^{\min}\right) = \frac{B''(\widetilde{e}_{nc})}{D''\left(\widetilde{E}_{nc}\right)}$. When marginal benefits are decreasing much faster than marginal damages are increasing, i.e. $\left|\widetilde{\phi}\right|$ is very large, $\widetilde{z}^{\min}$ becomes smaller since damages are not considered important relative to benefits and thus, there is no value gained from an agreement. When damages become more prominent, driving $\widetilde{\phi}$ to zero, $\widetilde{z}^{\min}$ tends to n: as the evaluation of damages relative to benefits increases, even the grand coalition becomes an option. However, as damages become very high relative to benefits, large coalitions require that their members reduce their emissions drastically to internalize very large externalities over many countries. This could require drastically reducing production and consumption, since abatement is not a viable option, or even drive them to negative values depending on the functional forms of benefits and damages. For example, in the case of quadratic benefit and damage functions, Diamantoudi and Sartzetakis (2006), using quadratic specifications for benefits and damages, show that any coalition with more than four members would require from their members negative emission levels. Larger coalitions are stable if the positivity constraint is ignored, which is the case with a number

of papers using abatement as the only choice variable.

When abatement becomes a viable option, separate from emissions, the ratio $\psi = \frac{B''(e_{nc})}{C''(NE_{nc})}$ plays an important role in determining the value of $z^{\min}$, in addition to ϕ. As noticed already, when the rate of the increase in abatement costs is much higher than the rate of the increase in the benefits from production and consumption, that is as $\psi \to 0$, then $z^{\min} \to \tilde{z}^{\min}$. When abatement is very expensive we have the problems described above, restricting the size of stable coalitions. However, when the cost of abatement decreases such that ψ becomes much higher than ϕ, then larger stable coalitions are stable. That is, for given benefits, the less costly abatement is relative to damages, the higher is the stable coalition.

Although the low cost abatement option can lead to large stable coalitions, the welfare gains that such coalitions attain over the case of no-cooperation, are very small. This is because $z^{\min}$ yields the lowest possible welfare level for both signatories and nonsignatories and thus, s_1^*, regardless of how much larger than $z^{\min}$ is, yields a welfare level that cannot be significantly larger than w_{nc}, under the assumptions of concave benefits and convex damages and abatement costs.

In the case that the coalition has no leadership power, coalition members of any size choose lower emissions relative to nonsignatories, since they internalize the externalities on all the coalition members and do not have any power over nonsignatories. Therefore, starting from a welfare level w_{nc} for all countries, as a coalition of two countries forms, the welfare of signatories is lower than that of nonsignatories and their difference increases as the size of the coalition increases. Given that in this case both w_s and w_{ns} are monotonically increasing in s, only one stable coalition s^* exists, which lies to the right of $z^{\min} = 1$. In the case of quadratic functions the literature, using a single choice variable, has shown that the maximum size of a stable coalition is two. Since there is no leadership effect, nothing changes when we model abatement separately from emissions. Therefore, large coalitions can be stable only if the coalition acts as a leader and abatement costs are not very high relative to damages from emissions. The main points of the above discussion are summarized in the following Proposition.

Proposition 3 *When emissions and abatement are modeled separately, larger*

coalitions become stable if the coalition can exercise leadership power over free-riders. For given benefits, the lower is the abatement cost relative to damages the larger is the stable coalition, and even the grand coalition could become stable. However, large stable coalitions offer very little improvement to global welfare over the non-cooperative level.

The latter result from the above Proposition has been termed in the literature as the "green paradox" or "paradox of cooperation" (see Barret, 1994 and Finus et al., 2021). However, in light of the above discussion, it is not paradoxical at all. Countries facing large damages from net emissions and having available cheap abatement, engage extensively in abatement, eliminating most, or even all, emissions and enjoy the benefits from production and consumption. However, under such conditions, engaging in high abatement levels is individually rational, that is, a collective agreement attains very little, if any, welfare improvement. When abatement costs are low relative to damages, the difference between the welfare achieved when all countries act as singletons and the welfare achieved by the grand coalition becomes very small. When abatement is very cheap, countries can increase their production and consumption, since they can eliminate all generated emissions costlessly: all countries could join such a coalition that requires no effort and attains the highest possible welfare, which though they can also achieve by acting independently.

4 Specific example: quadratic functions

For an illustrative example of the above analysis, we consider the case of quadratic functions most often used in this literature. That is, we assume quadratic benefits, $B_i(e_i) = b(ae_i - \frac{1}{2}e_i^2)$, with $a > 0$ and $b > 0$ and quadratic damages from aggregate net emissions, $D_i(NE) = \frac{1}{2}c(NE)^2$, with $c > 0$. We further assume that country i's cost of abatement is quadratic, $C_i(x_i) = \frac{1}{2}dx_i^2$, with $d > 0$. Therefore, country i's social welfare is,

$$W_i = b(ae_i - \frac{1}{2}e_i^2) - \frac{1}{2}c(NE)^2 - \frac{1}{2}dx_i^2. \tag{7}$$

In the second stage, each nonsignatory country i chooses e_i and x_i in order to maximize its own welfare W_i, given in (7), taking the other countries'

emission and abatement levels as given. The first order conditions, $\frac{\partial W_i}{\partial e_i} = 0 \implies ba - be_i = cNE$, and $\frac{\partial W_i}{\partial x_i} = 0 \implies cNE = dx_i$, yield country i's emission and abatement as functions of the rest of the countries' net emissions, $NE_{-i} = \sum_{j \neq i}(e_j - x_j)$,

$$Re_i = e_i^*(NE_{-i}) = \frac{(1+\delta)\, a - \gamma\delta NE_{-i}}{1+\delta+\gamma\delta} \ , \tag{8}$$

$$Rx_i = x_i^*(NE_{-i}) = \frac{a + NE_{-i}}{1+\delta+\gamma\delta} \ , \tag{9}$$

respectively, where $\gamma = \frac{c}{b}$ and $\delta = \frac{d}{c}$. The parameter γ is the ratio between the constant slope of marginal damages from net emissions and the constant slope of marginal benefits from emissions and δ is the ratio between the constant slope of marginal abatement cost and the constant slope of marginal damages from net emissions. We use a different notation from the one used in the previous Section, to avoid inappropriately confusing this particular example, in which third order derivatives are assumed zero, to the general case examined before. Notice that in this particular example, $B''(e_i) = -b$, $D''(NE) = c$, and $C'''(x_i) = d$ at any level of emissions and abatement and thus, we can write $\phi = \frac{-1}{\gamma}$ and $\psi = \frac{-1}{\gamma\delta}$. However, ϕ and ψ take different values depending on the functional forms used and, in general, are not independent of emissions and abatement.

Country i's best reaction to an increase in the rest of the countries' net emissions is to decrease its own net emissions both by decreasing its emissions,[12] and by increasing its abatement effort[13]. As we noted in the general case, the speed of emissions reaction exceeds that of abatement if $\gamma\delta > 1 \implies d > b$, that is, when the slope of the marginal cost of abatement exceeds the slope of the marginal benefits from emission. If benefits from emissions are relatively higher, country i adjusts mostly by increasing abatement, while when abatement costs are relatively higher, it adjusts by primarily reducing emission.

Appendix III present countries' equilibrium choices in the two benchmark cases, assuming that in the first case there is either no cooperation or complete cooperation.

[12] The slope of the emission reaction function is $Re_i' = \frac{\partial e_i^*(NE_{-i})}{\partial NE_{-i}} = -\frac{\gamma\delta}{1+\delta+\gamma\delta}$.

[13] The slope of the abatement reaction function is $Rx_i' = \frac{\partial x_i^*(NE_{-i})}{\partial NE_{-i}} = \frac{1}{1+\delta+\gamma\delta}$.

4.1 Coalition formation

We now move to examine the stable size of the coalition. Using the notation established in Section 3, aggregate net emissions are, $NE = E - X = s\,(e_s - x_s) + (n-s)\,(e_{ns} - x_{ns})$. Substituting total net emissions of the rest of the countries, $\sum_{j \neq i}(e_j - x_j) = NE_s + (n-s-1)\,(e_{ns} - x_{ns})$, where $NE_s = s\,(e_s - x_s)$, into the reaction functions (8) and (9), we calculate nonsignatories' reaction functions, $Re = e_{ns}(e_s, x_s, s)$ and $Rx = x_{ns}(e_s, x_s, s)$. For brevity, we only report nonsignatories net emissions' reaction function,[14]

$$R\varepsilon = \varepsilon_{ns}(NE_s) = \frac{\delta a - (1+\gamma\delta)\,NE_s}{\delta + (1+\gamma\delta)\,(n-s)}. \tag{10}$$

It is clear that when all countries act as singletons, $s = 0$, the above collapses to the Nash equilibrium, $R\varepsilon|_{s=0} = R\varepsilon_i$, where $R\varepsilon_i$ is defined by subtracting Rx_i given in (9) from $Re_i(NE_{-i})$ in (8). Furthermore, in the absence of damages, $c = 0$, $R\varepsilon|_{c=0} = a$, while when abatement is costless, $d = 0$, $R\varepsilon|_{d=0} = 0$. Finally, if abatement was not an option, or extremely expensive, that is, $d \to \infty$, then $x_{ns} = x_s \to 0$ and $R\varepsilon|_{d\to\infty} \to \frac{a-\gamma s e_s}{1+\gamma(n-s)}$, which is exactly the same reaction function as the one reported in Diamantoudi and Sartzetakis (2006).

We can compare the slopes of the net emission reaction functions with and without the option of abatement to verify and extend the discussion in Section 3. It is clear that when technological advancements reduce substantially the cost of abatement, nonsignatory countries adjust their net emissions, responding to a change in the coalition's emissions, faster relative to when abatement is not an option. Denote the slope of the reaction function in (10) by $R\varepsilon' = \frac{\partial \varepsilon_{ns}(NE_s)}{\partial NE_s} = -\frac{(1+\gamma\delta)}{\delta+(1+\gamma\delta)(n-s)}$ and the slope in the absence of abatement option by $R\varepsilon'|_{x_s=0} = \frac{\partial e_{ns}(E_s)}{\partial E_s} = -\frac{\gamma}{1+\gamma(n-s)}$.[15] Then, $|R\varepsilon'| - \left|R\varepsilon'|_{x_s=0}\right| = \frac{1}{[1+\gamma(n-s)][\delta+(1+\gamma\delta)(n-s)]} > 0$, which goes to zero as $d \to \infty$ and thus $\delta \to \infty$, and becomes larger the smaller is the slope of the marginal cost of abatement. Thus, nonsignatories adjust faster to the signatories ' choices as the abatement option becomes cheaper, which implies that the leading coalition can exercise more pressure on singletons when abatement

[14] The emission and abatement reaction functions have the same numerators as (8) and (9) respectively and the same denominator as the one in the following expression (10).

[15] See Diamantoudi and Sartzetakis (2006), p. 251.

becomes cheaper.

Signatories maximize the coalition's welfare, sW_s, explicitly taking into account nonsignatories' reaction. Thus, aggregate net emissions depend only on signatories' choices, $NE(e_s, x_s) = s\,(e_s - x_s) + (n-s)(e_{ns}\,(e_s, x_s) - x_{ns}\,(e_s, x_s))$. The first-order conditions yield signatories emission and abatement effort levels, $e_s\,(s)$ and $x_s\,(s)$. Substituting $e_s\,(s)$ and $x_s\,(s)$ into the non-signatories' reaction functions, we derive the non-signatories' emission and abatement. As shown in Appendix V, $e_{ns} \underset{>}{\overset{<}{\equiv}} e_s \Leftrightarrow \delta s \underset{>}{\overset{<}{\equiv}} \Omega$, $x_{ns} \underset{<}{\overset{>}{\equiv}} x_s \Leftrightarrow \delta s \underset{>}{\overset{<}{\equiv}} \Omega$ and $\varepsilon_{ns} \underset{>}{\overset{<}{\equiv}} \varepsilon_s \Leftrightarrow \delta s \underset{>}{\overset{<}{\equiv}} \Omega$, where $\Omega = \delta + (n-s)\,(1+\gamma\delta) > 0$. We will explore this condition further when we compare the welfare of signatories to that of non-signatories. At this point, it is worth noting that since for small coalition sizes $\Omega > \delta s$ regardless of the value of the parameters, when coalition size is small non-signatories emit less and abate more than the signatories. For small coalition sizes, the leadership effect dominates the damage reduction effect. Furthermore, Appendix IV states the restrictions for $e_s > 0$ and $e_{ns} > 0$ and presents the indirect welfare function of signatories $w_s\,(s)$ and non-signatories $w_{ns}(s)$.

Utilizing the specification of the indirect welfare functions presented in Appendix IV, Proposition 3 establishes their properties, applying the results in Proposition 1 to the specific example of quadratic functions. The proof of Proposition 3 follows the same steps of the general case applied in the quadratic functions.[16,17]

Proposition 4 *We consider the indirect welfare function of signatories and non-signatories,* (w_s) *and* (w_{ns})*, respectively. If we define* $s^{\min} = \frac{\delta+(1+\delta\gamma)n}{1+\delta+\delta\gamma}$*, then,*

(i) $s^{\min} = \arg\min_{s\in\Re\cap[0,n]} w_s(s)$*, that is,* $s^{\min}$ *is the* s *at which* w_s *is minimized,*

(ii) $w_s(s)$ *increases in* s *if* $s > s^{\min}$ *and it decreases in* s *if* $s < s^{\min}$*,*

(iii) $w_{ns}(s) \lesseqgtr w_s(s)$ *for all* $s \lesseqgtr s^{\min}$*.*

[16]The proof is available to the interested reader upon request. The steps are the same as in Diamantoudi and Sartzetakis (2006) although they use emissions as the only choice variable.

[17]Notice that if we substitute the values $\phi = \frac{-1}{\gamma}$ and $\psi = \frac{-1}{\gamma\delta}$, into the definition of $z^{\min}$ defined in Proposition 1, yields the definition of $s^{\min}$.

The above defined properties of the indirect welfare function imply that the indirect welfare function of nonsignatories cuts the indirect welfare function of signatories from below at its minimum, defined by $s^{\min}$. Note that $s^{\min}$ solves $\Omega = \delta s$, that is, the welfare of the signatories takes its minimum value at the coalition size that equalizes the emission and abatement of signatories and non-signatories.

We know from Section 3 that there are two effects that make coalition members' choices different from those of outsiders: the effect of internalizing damages of all members and the leadership effect which work in opposite directions. At the critical size of coalition $s^{\min}$ these two effects offset each other and thus, signatories and nonsignatories choices are the same as in the absence of a coalition. The damage internalization effect is $(s-1)$ and, in this particular example, the leadership effect is $\frac{-s(n-s)(1+\gamma\delta)}{(n-s)(1+\gamma\delta)+\delta}$. The sum of these two effects is $\frac{\delta s-\Omega}{(n-s)(1+\gamma\delta)+\delta}$, which becomes zero for $\Omega = \delta s$.

In the case that countries choose only their emission level, Diamantoudi and Sartzetakis (2006) find that the critical coalition size at which $w_{ns}(s) = w_s(s)$, is $\widetilde{s}^{\min} = \frac{1+\gamma n}{1+\gamma}$, which is clearly increasing in γ, for $n > 1$. Furthermore, Diamantoudi and Sartzetakis (2006) associate the size of the stable coalition to the integer closer to $\widetilde{s}^{\min}$ and after restricting the admissible values of γ in order to have positive emissions, the size of stable coalitions is limited to $\widetilde{s}^{\min} \in \{2, 3, 4\}$. In the present case, it is clear that when abatement costs are very low, $\lim_{\delta \to 0} z^{\min} \to n$ as expected from Proposition 2. The case that abatement costs are very high, $\delta \to \infty$ corresponds to the case that abatement is not an option and we go back to the results obtained in Diamantoudi and Sartzetakis (2006).

The following Corollary compares the case in which emission is the only choice variable with the case we develop in the present paper.

Corollary 1 *Allowing countries to choose abatement separately from emission level increases the size of the stable coalition for any admissible values of benefits, damages and cost parameters. The coalition size, when abatement is an option, is increasing as the cost of abatement decreases.*

Direct comparison of the two cases reveals that $s^{\min} > \widetilde{s}^{\min}$, for $\delta > 0$ and $n > 1$. Furthermore, $\frac{\partial s^{\min}}{\partial \delta} = \frac{1-n}{(1+\delta+\delta\gamma)^2} < 0$, that is, recalling that $\delta = \frac{d}{c}$, as

either the cost of abatement decreases or environmental damages increase, the higher is the size of stable coalition. In what follows, we illustrate the above results by considering numerical examples.

4.2 Illustration of the results using simulations

To facilitate the comparison to the case that abatement is not a separate choice variable, we choose the same parameter values used in Diamantoudi and Sartzetakis (2006). That is, we assume the following values for the parameters: $n = 10$, $a = 10$, $b = 6$, and $c = 0.39999$, which results in $\gamma = 0.066665$. These values satisfy the restrictions set in *Lemma 1*, since the parameter γ is less than $\gamma < \frac{4}{n(n-4)} < 0.066667$. We also choose a small value for the parameter δ, such that $c > d$, which means that marginal damage increases faster than marginal abatement cost. We choose $d = 0.239994$, so that given $c = 0.39999$, we have $\delta = 0.6$.

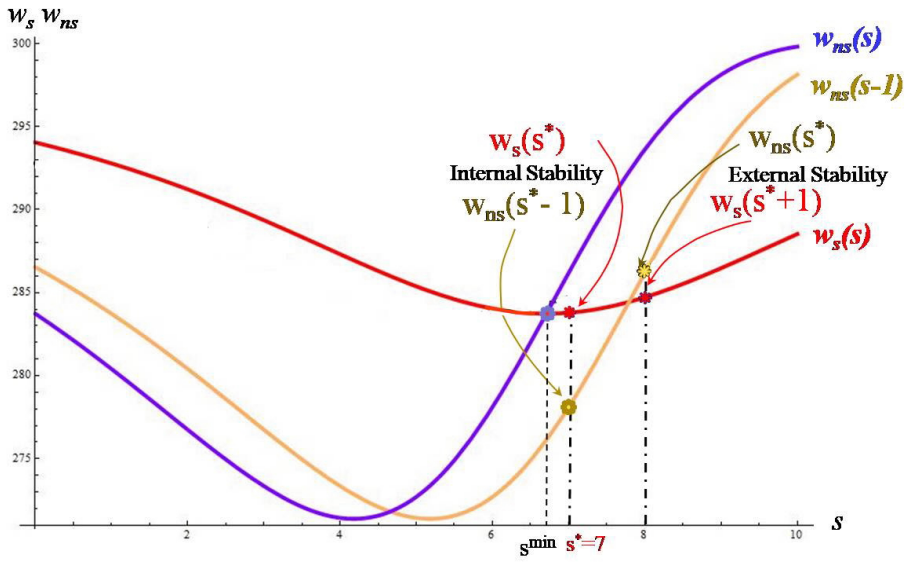
In Figure 3, we plot, using the above defined parameter values, the indirect welfare functions against different coalition sizes s. The red curve depicts $w_s(s)$, the purple curve $w_{ns}(s)$ and the orange curve $w_{ns}(\ s-1)$. Notice that $w_{ns}(s-1)$ is a horizontal shift of $w_{ns}(s)$.

As Figure 3 illustrates, no country wants to exit or join a coalition of seven, that is, a coalition of seven is stable, $s^* = 7$. The internal condition $w_s(s^*) \geq w_{ns}(s^* - 1)$ is satisfied since $w_s(7) > w_{ns}(6)$, as the $w_s(s)$ curve is above the $w_{ns}(s-1)$ curve. Moreover, coalition $s^* = 7$ is externally stable i.e. $w_s(s^*+1) \leq w_{ns}(s^*)$ since at $s = s^*+1 = 8$ the $w_{ns}(s-1)$ curve is above the $w_s(s)$ curve. Therefore, the coalition of size $s^* = 7$ is stable.

Similar to the analysis in Diamantoudi and Sartzetakis (2006) the stable coalition size is the higher integer following the size of the coalition for which the welfare of the signatories is at its minimum, $s^{\min}$ and for which $w_s = w_{ns}$. In our example, $s^{\min} = 6.70732$.

The above specified parameter values satisfy the constraints for $e_s > 0$, $e_{ns} > 0$, $e_s - x_s > 0$ and $e_{ns} - x_{ns} > 0$ and also the internal and external stability conditions. Note that for the same parameter values, the optimal size of the coalition in Diamantoudi and Sartzetakis (2006) is $s^* = 3$. The following Table summarizes the results of the simulations. The first column reports the equilibrium values of emissions, abatement, net emissions and

1



3.jpg

Figure 3: Defining the size of a stable IEA

welfare for signatories, the second column for non-signatories and the last column reports the aggregate values.

	$s^* = 7$	$n - s^*$	$s^* + (n - s^*)$
e	9.622	9.665	96.347
x	9.456	8.376	91.322
$e - x$	0.165	1.289	5.025
w	283.8	286.2	2,845.2

Table 1. Coalition formation with abatement for $\delta = 0.6$

These values confirm that signatories emit less and abate more than non-signatories, i.e., $e_s < e_{ns}$ and $x_s > x_{ns}$. Moreover, the net emissions are significantly smaller for the signatories, $e_s - x_s$, than for the non-signatories, $e_{ns} - x_{ns}$. Total net emissions, $\sum_{i=1}^{n}(e_i - x_i)$, include the activities from both signatories and non-signatories.

Total net emissions are slightly smaller relative to the non-cooperative case, $NE_{nc} = 5.4545$, but much larger than the full cooperation case, $NE_c = 0.5736$. The welfare level each of the seven members of the coalition attains is slightly over the one they attain at the non-cooperative case $w_{nc} = 283.736$ and, thus, the total welfare improvement attained by the coalition is almost negligible and derives from the gains of free-riders.

4.2.1 Comparative statics with respect to abatement cost

In the simulations presented above, we used $\delta = 0.6 < 1$, indicating that the abatement cost parameter d is smaller than the emission damage parameter c. In order to show the effect of changes in the abatement cost that change the ratio δ, we simulate the model for two extreme values of δ, namely, $\delta = 0.0000001$ and $\delta = 1.637$. The first case indicates that abatement cost is negligible relative to environmental damages, while in the second, we assume that abatement costs exceed environmental damages. Table 2 presents the results of the simulations: on the left, the case with significantly high abatement cost is presented, which results in $s^* = 5$ and on the right, the case with almost zero abatement cost is presented, which yields the grand coalition, $s^* = 10$.

	$\delta = 1.637$			$\delta = 1 * 10^{-7}$		
	$s^* = 5$	$n - s^*$	$s^* + (n - s^*)$	$s^* = 10$	$n - s^*$	$s^* + (n - s^*)$
e	9.079	9.192	91.354	10	–	100
x	8.439	7.406	79.229	10	–	100
$e - x$	0.639	1.785	12.124	$1 * 10^{-8}$	–	$1 * 10^{-7}$
w	244.7	250.7	2,476.0	300.0	–	3,000.0

Table 2. Coalition formation with abatement under different values of δ

The results of the simulations presented in Tables 1 and 2 reveal that the value of the parameter δ is crucial in determining the size of the stable IEA. When $\delta = 0.6$, the size of the stable coalition is $s^* = 7$. When δ takes a very low value, i.e., $\delta = 0.0000001$, the number of signatory countries increases, reaching the grand coalition, $s^* = 10$. When abatement costs are negligible, all countries generate the maximum emissions which they completely abate, $\lim_{\delta \to 0} e = \lim_{\delta \to 0} x = a = 10$, to attain the maximum benefits. On the contrary, when δ takes a very high value, i.e. $\delta = 1.637$, the number of signatory countries decreases to $s^* = 5$. Notice, though, that even when the abatement cost parameter takes the highest value allowed by the model's constraints, presented in Lemma 1, the size of the stable coalition is higher than the case in which countries have only one choice variable.

When δ approaches zero and the grand coalition emerges, countries' net emissions are the lowest possible, achieving the highest welfare. Notice,

though, that as abatement becomes very inexpensive, i.e., as δ approaches zero, we also have that, $\lim_{\delta \to 0} e_{nc} = \lim_{\delta \to 0} x_{nc} \to a = 10$ and thus, $\lim_{\delta \to 0} w_{nc} \to 300$. That is, countries' choices are the same whether or not they cooperate. When the abatement cost becomes negligible, indeed, the grand coalition is stable, but it delivers no additional welfare gains relative to the case that countries decide independently. The following Corollary restates the general result presented in Proposition 3.

Corollary 2 *When the coalition acts as a leader and countries choose emissions and abatement separately, larger coalitions can be stable as abatement costs decrease, including the grand coalition when costs are negligible. However, very small welfare gains over the non-cooperative equilibrium are attained.*

When technology drives abatement cost to zero, it is individually rational for the countries to abate all their emissions, and thus, they emit the highest possible amounts to attain the highest welfare from production and consumption. As the abatement costs increase, free-riding incentives become significant, and the size of stable coalitions decreases but remains larger compared to the case where abatement is not an option. All stable coalition cases, though, yield equilibrium values for signatory and non-signatory countries' welfare, i.e., almost the same as the ones they receive acting non-cooperatively. Modelling abatement separately from emissions allows for higher stable coalitions because free-riding incentives are reduced, which implies that the larger coalitions yield no welfare gains over the non-cooperative case.

5 Conclusions

The present paper offers a fresh look at the literature on the formation of international environmental agreements, by introducing emissions and abatement as countries' separate choice variables. The classic model's structure is kept unchanged, assuming a two stage game in which the internal and external stability conditions define coalition's stability. Given the structure of the game, nonsignatories' decision making is very important in determining the

size of stable coalitions. For this reason we focus on disentangling the effect of changes in benefits, damages and abatement cost on nonsignatories' reaction to signatories' net emission choices. We illustrate diagrammatically how the three components of countries' welfare interact in determining nonsignatories' equilibrium choices and we perform comparative statics. We show, for example, that an improvement in abatement technology induces directly an increase in abatement but also indirectly an increase in emissions, with the former outweighing the later, leading to increased benefits and reduced damages. Such an analysis cannot be performed using a model defined over a single choice variable, be it abatement or emissions. Furthermore, we show that nonsignatories' reaction to changes in the coalition's net emissions becomes more responsive as abatement becomes cheaper, which strengthens the coalition's position as a leader.

We identify two effects that determine signatories' choices: the damage internalization effect and the leadership effect. As the size of the coalition increases, the former increases while the latter decreases. There is a critical coalition size at which the two effects offset each other and thus, signatories and nonsignatories make the exact same choices. We are able to relate this critical coalition size to the stable coalition size.

Connecting in this way nonsignatories' and signatories' decision making to the stable coalition size, reveals the importance of nonsignatories' reaction functions. We show that, ceteris paribus, as abatement becomes cheaper, nonsignatories become more responsive, enhancing the leadership effect, allowing thus, largest stable coalitions to be formed. Because of its importance, we choose to present comparative statics with respect to abatement costs. However, it is clear that static analysis can be performed with respect to the other two components of countries' welfare or combinations of them.

Given that the critical coalition size, at which signatories' and nonsignatories' choices equate, corresponds to the choices at the Nash equilibrium, this coalition yields the Nash welfare for all. Since the size of stable coalition is defined close to the critical size, any stable coalition, regardless of its size, will not offer significant welfare improvement over the Nash. This is because, in this model's set up, the only way in which large coalitions can become stable is by reducing abatement costs relative to benefits. However, in such case the same choices are individually rational, that is, forming a coalition does

not add much welfare over the Nash. Furthermore, large stable coalitions are possible if abatement costs are high, only if damages are high relative to benefits, but such coalitions require negative net emissions.

Finally, it is clear that under this model's specification of collisional stability, if the second stage game is simultaneous, only very small coalitions are stable. Larger coalitions are stable only when the coalition acts as a leader, regardless of the abatement cost or the level of damages.

Although in our model, countries' emissions and abatement choices are assumed to be taken simultaneously, it is easy to show that our results hold if we assume that countries choose either emissions or abatement at a separate stage and then the coalition is formed over net emissions at a later stage, as long as there are no strategic interactions between these two stages. Thus, our model unifies and compares all existing, and even those implied, modeling approaches of coalition formation.

Within the limitations of this model, which are well known and clearly stated in the paper, the modification proposed in this paper of introducing emissions and abatement as separate choice variables could open new directions for future research. The two-choice-variable model can be used to analyze different, and potentially richer, structures of the game, allowing, for example, more stages in the decision making process and different assumptions regarding countries' information. Another potentially important extension could be to allow for heterogeneity in countries' benefits, damages and cost of abatement. The two-choice-variable model allows to examine the effect of heterogeneity in both damages and abatement cost, not possible using a single-variable model.

6 References

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

Barrett, S. (2006) "Climate treaties and breakthrough technologies." *American Economic Review,* **96**, 22–25.

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

Diamantoudi, E. and Sartzetakis, E. (2006). "Stable International En-

vironmental Agreements: An Analytical Approach." *Journal of Public Economic Theory* ***8****, 247–263.*

Diamantoudi, E. and E. Sartzetakis (2018) "International Environmental Agreements-The Role of Foresight." *Environmental & Resource Economics*, **71**(1), 241-257.

Diamantoudi, E., E. Sartzetakis and S. Strantza (2020) "International Environmental Agreements and Trading Blocks: Can Issue Linkage Enhance Cooperation?" *Strategic Behavior and the Environment*, **8**(3), 269-310.

Eichner, T. and Pethig, R. (2013) "Self-enforcing environmental agreements and international trade." *Journal of Public Economics*, **102**, 37–50.

Eichner, T. and Pethig, R. (2015) "Self-enforcing international environmental agreements and trade: taxes versus caps." *Oxford Economic Papers*, **67(4)**, 897–917.

Finus, M. and Rundshagen B. (2001), "Endogenous coalition formation global pollution control." *working paper, FEEM,* Nota di Lavoro 43.2001.

Finus, M., F. Furini and A. V. Rohrer (2021a) "The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership." *Journal of Environmental Economics and Management*, **109**, doi.org/10.1016/j.jeem.2021.102461.

Finus, M., F. Furini and A. V. Rohrer (2021b) "International Environmental Agreements and the Paradox of Cooperation: Revisiting and Generalizing Some Previous Results." Graz Economics Papers 2021-05, University of Graz, Department of Economics.

Frondel, M., Horbach, J. and Rennings, K. (2007) "End-of-pipe or cleaner production? An empirical comparison of environmental innovation decisions across OECD countries." *Bussiness Strategy and the Environment*, **16**, 571-584.

Hoel, M. and De Zeeuw A., (2010). "Can a Focus on Breakthrough Technologies Improve the Performance of International Environmental Agreements?" *Environmental and Resource Economics,* **47**(3), 395-406.

McGinty, M. (2020) "Leadership and Free-Riding: Decomposing and Explaining the Paradox of Cooperation in International Environmental Agreements." *Emvironmental and Resource Economics*, **77**, 449-474.

Rubio, J. S. and Casino, B. (2001), "International Cooperation in Pollution Control" *mimeo.*

Ulph, A & Rubio, S (2006) "Self-enforcing Environmental Agreements Revisited", *Oxford Economic Papers*, **58(2)**, 233-263.

7 Appendices

7.1 Appendix I: Proof of Proposition 1

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.

[**Parts 1-2**] For a coalition size z, we denote the aggregate net emissions level at the equilibrium by $NE^*(z) = s\left(e_s^*(z) - x_s^*(z)\right) + (n-s)\left(e_{ns}^*(z) - x_{ns}^*(z)\right)$, where $e_{ns}^*(z) = e_{ns}^*(e_s^*(z), x_s^*(z), z)$ and $x_{ns}^*(z) = x_{ns}^*(e_s^*(z), x_s^*(z), z)$. Given the assumptions regarding countries' benefit function, $B'' < 0$ we have $e_s^*(z) \gtreqless e_{ns}^*(z) \Leftrightarrow B'\left(e_s^*(z)\right) \lesseqgtr B'\left(e_{ns}^*(z)\right)$. In equilibrium we also have,

$$\begin{aligned} B'(e_s^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} \\ &\text{and} \\ B'(e_{ns}^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial e_{ns}}\right|_{e_{ns}=e_{ns}^*(z)}. \end{aligned}$$

where the derivative of aggregate net emissions around the equilibrium value $e_s^*(s)$ is: $\left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} = z + (n-z)\left.\frac{\partial e_{ns}^*(e_s,x_s))}{\partial e_s}\right|_{e_s=e_s^*(z)} - (n-z)\left.\frac{\partial x_{ns}^*(e_s,x_s))}{\partial e_s}\right|_{e_s=e_s^*(z)}$.

Thus, recalling that $D'(E^*(z)) > 0$ and $\left.\frac{\partial NE^*(z)}{\partial e_{ns}}\right|_{e_{ns}=e_{ns}^*(z)} = 1$, in equilibrium we have,

$$\begin{aligned} e_s^*(z) &\gtreqless e_{ns}^*(z) \Leftrightarrow \\ \left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} &\lesseqgtr 1. \end{aligned} \qquad \text{(A I. 1)}$$

Furthermore, given the assumption regarding countries' abatement cost function, $C'' > 0$ we have that $x_s^*(z) \gtreqless x_{ns}^*(z) \Leftrightarrow C'\left(x_s^*(z)\right) \gtreqless C'\left(x_{ns}^*(z)\right)$. In

equilibrium we also have,

$$C'(x_s^*(z)) \equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)}$$

and

$$C'(e_{ns}^*(z)) \equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial x_{ns}}\right|_{x_{ns}=x_{ns}^*(z)}.$$

where the derivative of aggregate net emissions around the equilibrium value $x_s^*(s)$ is: $\left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)} = -z+(n-z)\left.\frac{\partial e_{ns}^*(e_s,x_s))}{\partial x_s}\right|_{x_s=x_s^*(s)} -(n-z)\left.\frac{\partial x_{ns}^*(e_s,x_s))}{\partial x_s}\right|_{x_s=x_s^*(s)}$. Thus, recalling again that $D'(E^*(z)) > 0$ and $\left.\frac{\partial NE^*(z)}{\partial x_{ns}}\right|_{x_{ns}=x_{ns}^*(z)} = -1$, in equilibrium we have,

$$\begin{aligned} x_s^*(z) &\gtreqless x_{ns}^*(z) \Leftrightarrow \\ \left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)} &\gtreqless -1. \end{aligned} \qquad \text{(A I. 2)}$$

The first order condition of the non-signatories with respect to their emissions, yield the identity $B'(e_{ns}^*(e_s)) \equiv D'[z\,(e_s - x_s) + (n-z)(e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s))]$. Differentiating both sides of this identity with respect to e_s yields,

$$\frac{\partial e_{ns}^*(e_s,x_s)}{\partial e_s} = \frac{zD''(NE(e_s)) - (n-z)D''(NE(e_s))\frac{\partial x_{ns}^*(e_s,x_s)}{\partial e_s}}{B''(e_{ns}^*(e_s)) - (n-z)D''(NE(e_s))}\,. \qquad \text{(A I. 3)}$$

Similarly, the first order condition of nonsignatories with respect to their abatement, yield the identity $C'(x_{ns}^*(e_s)) \equiv D'[z\,(e_s - x_s)+(n-z)(e_{ns}^*(e_s,x_s)- x_{ns}^*(e_s,x_s))]$. Differentiating both sides of this identity with respect to e_s yields,

$$\frac{\partial x_{ns}^*(e_s,x_s)}{\partial e_s} = \frac{zD''(NE(e_s)) + (n-z)D''(NE(e_s))\frac{\partial e_{ns}^*(e_s,x_s)}{\partial e_s}}{C''(x_{ns}^*(e_s)) + (n-z)D''(NE(e_s))}\,. \qquad \text{(A I. 4)}$$

The solution of the system of equations (A I. 3) and (A I. 4) yields,

$$\left.\frac{\partial e_{ns}^*(e_s,x_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z}{(n-z)\,(1-\psi)-\phi}\,, \qquad \text{(A I. 5)}$$

and

$$\left.\frac{\partial x_{ns}^*(e_s,x_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z\psi}{(n-z)\,(1-\psi)-\phi}\,. \qquad \text{(A I. 6)}$$

where $\phi = \frac{B''(e^*_{ns}(E_s,X_s))}{D''(NE)} < 0$ and $\psi = \frac{B''(e^*_{ns}(E_s,X_s))}{C''(x^*_{ns}(E_s,X_s))} < 0$, given $B'' < 0$, $D'' > 0$, and $C'' > 0$. From (A I. 5) and (A I. 6) it is clear that the slope of nonsignatories' emission reaction with respect to a change in signatories' emissions is negative around the equilibrium, $\frac{\partial e^*_{ns}(e_s,x_s)}{\partial e_s} < 0$, while the slope of the abatement reaction function is positive $\frac{\partial x^*_{ns}(e_s,x_s)}{\partial e_s} > 0$.

Substituting (A I. 5) and (A I. 6) into the inequality (A I. 1) yields,

$$z - \frac{z\,(n-z)\,(1-\psi)}{(n-z)\,(1-\psi)-\phi} \lessgtr 1.$$

Which reduces to

$$\begin{aligned} e^*_s(z) &\lessgtr e^*_{ns}(z) \Leftrightarrow \\ z &\lessgtr \frac{n\,(1-\psi)-\phi}{(1-\psi)-\phi}. \end{aligned} \tag{A I. 7}$$

In a similar manner as above, differentiating with respect to x_s the first order conditions of the non-signatories with respect to their emissions and abatement yields,

$$\frac{\partial e^*_{ns}(e_s,x_s)}{\partial x_s} = \frac{-zD''(NE(e_s)) - (n-z)D''(NE(e_s))\frac{\partial x^*_{ns}(e_s,x_s)}{\partial x e_s}}{B''(e^*_{ns}(e_s)) - (n-z)D''(NE(e_s))}\,. \tag{A I. 8}$$

$$\frac{\partial x^*_{ns}(e_s,x_s)}{\partial x_s} = \frac{-zD''(NE(e_s)) + (n-z)D''(NE(e_s))\frac{\partial e^*_{ns}(e_s,x_s)}{\partial x_s}}{C''(x^*_{ns}(e_s)) + (n-z)D''(NE(e_s))}\,. \tag{A I. 9}$$

The solution of the system of equations (A I. 8) and (A I. 9) yields,

$$\left.\frac{\partial e^*_{ns}(e_s,x_s)}{\partial x_s}\right|_{x_s=x^*_s(z)} = \frac{z}{(n-z)\,(1-\psi)-\phi}, \tag{A I. 10}$$

and

$$\left.\frac{\partial x^*_{ns}(e_s,x_s)}{\partial x_s}\right|_{x_s=x^*_s(z)} = \frac{z\psi}{(n-z)\,(1-\psi)-\phi}. \tag{A I. 11}$$

Nonsignatories react to an increase in the signatories' abatement by increasing their emissions, $\frac{\partial e^*_{ns}(e_s,x_s)}{\partial x_s} > 0$, and decreasing their abatement $\frac{\partial x^*_{ns}(e_s,x_s)}{\partial x_s} < 0$.

Substituting (A I. 10) and (A I. 11) into the inequality (A I. 2) yields,

$$-z + \frac{(n-z)\,z\,(1-\psi)}{(n-z)\,(1-\psi)-\phi} \gtrless -1.$$

Which reduces to

$$\begin{aligned} x_s^*(z) &\gtreqless x_{ns}^*(z) \Leftrightarrow \\ z &\gtreqless \frac{n(1-\psi)-\phi}{(1-\psi)-\phi}. \end{aligned} \tag{A I. 12}$$

As expected, working either from the comparison of signatories to nonsignatories emissions $e_s^*(z) \gtreqless e_{ns}^*(z)$, or from the comparison of their abatement $x_s^*(z) \gtreqless x_{ns}^*(z)$, we get the same condition presented in (A I. 7) and (A I. 12). Observe that when $e_s^*(z) = e_{ns}^*(z)$ the non-signatories' first order conditions remains satisfied, i.e.,

$$\begin{aligned} B'(e_{ns}^*(z)) &\equiv D'\left[z\left(e_s - x_s\right) + (n-z)\left(e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s)\right)\right] \Leftrightarrow \\ B'(e_{ns}^*(z)) &\equiv D'(ne_{ns}^*(z)), \\ C'(x_{ns}^*(e_s)) &\equiv D'\left[z\left(e_s - x_s\right) + (n-z)\left(e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s)\right)\right] \Leftrightarrow \\ C'(x_{ns}^*(z)) &\equiv D'(ne_{ns}^*(z)). \end{aligned}$$

These are 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 both the damage and the abatement cost functions, there exists a unique set (e_{nc}, x_{nc}) and, thus, a unique $z^{\min} = \frac{n(1-\psi)-\phi}{(1-\psi)-\phi}$. Reverting the coalition size back to integers yields:

$$\begin{aligned} e_s^*(s) &\lesseqgtr e_{ns}^*(s) \Leftrightarrow s \lesseqgtr z^{\min}, \\ x_s^*(z) &\gtreqless x_{ns}^*(z) \Leftrightarrow s \lesseqgtr z^{\min}. \end{aligned}$$

[Parts 3-4] Since $w_s(e_s^*(z)) \equiv B\left(e_s^*(z)\right) - D\left(NE^*(z)\right) - C\left(x_s^*(z)\right)$ we have

$$\frac{dw_s(z)}{dz} = B'(e_s^*(z))\frac{de_s^*(z)}{dz} - D'(E^*(z))\frac{\partial NE^*(z)}{\partial z} - C'\left(x_s^*(z)\right)\frac{dx_s^*(z)}{dz} \tag{A I. 13}$$

where, $\frac{\partial NE^*(z)}{\partial z} = e_s^*(z) - x_s^*(z) - (e_{ns}^*(e_s) - x_{ns}^*(z)) + \left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right)z + (n-z)\left(\frac{de_{ns}^*(z)}{dz} - \frac{dx_{ns}^*(z)}{dz}\right)$. Denoting by $\Delta NE^* = e_s^*(z) - e_{ns}^*(e_s) - (x_s^*(z) - x_{ns}^*(z))$,

(A I. 13) can be rewritten as follows:

$$\begin{aligned}\frac{dw_s(z)}{dz} &= \frac{de_s^*(z)}{dz}\left[B'(e_s^*(z)) - zD'(NE^*(z))\right] \\ &\quad -\frac{dx_s^*(z)}{dz}\left[C'(e_s^*(z)) + zD'(NE^*(z))\right] - D'(E^*(z))\Delta NE^* \\ &\quad -D'(E^*(z))(n-z)\left[\frac{de_{ns}^*(z)}{dz} - \frac{dx_{ns}^*(z)}{dz}\right]. \qquad \text{(A I. 14)}\end{aligned}$$

From signatories' first order conditions, we know that in equilibrium,

$$\begin{aligned}& B'(e_s^*(z)) - zD'(E^*(z)) \\ \equiv\ & D'(E^*(z))(n-z)\left[\left.\frac{\partial e_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)}\right], \\ & C'(e_s^*(z)) + zD'(NE^*(z)) \\ \equiv\ & D'(E^*(z))(n-z)\left[\left.\frac{\partial e_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)}\right].\end{aligned}$$

Furthermore, (A I. 5) and (A I. 6) yield $\left.\frac{\partial e_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z(1-\psi)}{(n-z)(1-\psi)-\phi}$ while (A I. 8) and (A I. 9) yield $\left.\frac{\partial e_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} = \frac{z(1-\psi)}{(n-z)(1-\psi)-\phi}$. Therefore, the first two terms in (A I. 14) can be written as,

$$\begin{aligned}& \frac{de_s^*(z)}{dz}\left[B'(e_s^*(z)) - zD'(NE^*(z))\right] \\ & -\frac{dx_s^*(z)}{dz}\left[C'(e_s^*(z)) + zD'(NE^*(z))\right] \\ =\ & D'(E^*(z))(n-z)\frac{z(1-\psi)}{(n-z)(1-\psi)-\phi}\left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right) \qquad \text{(A I. 15)}\end{aligned}$$

Differentiating nonsignatories' first order conditions at the equilibrium, $B'(e_{ns}^*(z)) \equiv D'[z(e_s^*(z) - e_s^*(z) + (n-z)e_{ns}^*(z)$ and $C'(x_{ns}^*(z)) \equiv D'(ze_s^*(z) + (n-z)e_{ns}^*(z))$,with respect to z yields,

$$\frac{de^*_{ns}(z)}{dz} = \frac{\Delta NE^* + z\left(\frac{de^*_s(z)}{dz} - \frac{dx^*_s(z)}{dz}\right) - (n-z)\frac{dx^*_{ns}(z)}{dz}}{\phi - (n-z)},$$
$$\frac{dx^*_{ns}(z)}{dz} = \frac{\Delta NE^* + z\left(\frac{de^*_s(z)}{dz} - \frac{dx^*_s(z)}{dz}\right) + (n-z)\frac{de^*_{ns}(z)}{dz}}{\frac{\phi}{\psi} + (n-z)}.$$

From the above, we calculate the difference $\frac{de^*_{ns}(z)}{dz} - \frac{dx^*_{ns}(z)}{dz}$ and thus, the last terms in (A I. 14) can be written as,

$$\begin{aligned} & D'(E^*(z))(n-z)\left[\frac{de^*_{ns}(z)}{dz} - \frac{dx^*_{ns}(z)}{dz}\right] \\ = \; & D'(E^*(z))(n-z)\frac{(\psi-1)\left[\Delta NE^* + z\left(\frac{de^*_s(z)}{dz} - \frac{dx^*_s(z)}{dz}\right)\right]}{(n-z)(1-\psi)-\phi} \end{aligned} \quad \text{(A I. 16)}$$

Substituting (A I. 15) and (A I. 16) into (A I. 14) yields,

$$\frac{dw_s(z)}{dz} = \frac{\phi \Delta NE^*}{(n-z)(1-\psi)-\phi}.$$

We know that $\frac{\phi}{(n-z)(1-\psi)-\phi} < 0$ for all z, and thus, the sign of $\frac{d\omega_s(z)}{dz}$ depends solely on ΔNE^*. From (A I. 7) and (A I. 12) we know that $\Delta NE^* = 0$ at $z^{\min}$. Therefore, given the uniqueness of $z^{\min}$, we can conclude that $\omega_s(s)$ is U-shaped and hence $\left.\frac{d\omega_s(z)}{dz}\right|_{z \lessgtr z^{\min}} \lesseqgtr 0$. The conversion to integer values of coalition size is trivial.

[Part 5] Recall that $w_s(z) = B(e^*_s(z)) - D(NE^*(z)) - C\left(x^*_s(z)\right)$ and $w_{ns}(s) = B(e^*_{ns}(z)) - D(NE^*(z)) - C\left(x^*_{ns}(z)\right)$. Thus, $w_s(z) \gtreqless w_{ns}(z) \Leftrightarrow B(e^*_s(z)) - C\left(x^*_s(z)\right) \gtreqless B(e^*_{ns}(z)) - C\left(x^*_{ns}(z)\right)$ and since $B' > 0$ and $C' > 0$ we have $B(e^*_s(z)) - C\left(x^*_s(z)\right) \gtreqless B(e^*_{ns}(z)) - C\left(x^*_{ns}(z)\right) \Leftrightarrow e^*_s(z) \gtreqless e^*_{ns}(z)$ and $x^*_s(z) \lesseqgtr x^*_{ns}(z) \Leftrightarrow s \lesseqgtr z^{\min}$.

7.2 Appendix II: Proof of Proposition 2

In order to make the analysis is terms of C'' we express $z^{\min}$ in terms of the second derivatives and not the ratios ϕ and ψ, that is,
$z^{\min} = \frac{n[D''(NE_{nc})C''(x_{nc}) - B''(e_{nc})D''(NE_{nc})] - B''(e_{nc})C''(x_{nc})}{D''(NE_{nc})C''(x_{nc}) - B''(e_{nc})D''(NE_{nc}) - B''(e_{nc})C''(x_{nc})}$. The difference $z^{\min} > \widetilde{z}^{\min}$ is decreasing in $C''(x_{nc})$, because $z^{\min}$ is decreasing in $C''(x_{nc})$,

$\frac{\partial z^{\min}}{\partial C''(x_{nc})} = -\frac{(n-1)D''(NE_{nc})[B''(e_{nc})]^2}{[D''(NE_{nc})(C''(x_{nc})-B''(e_{nc}))-B''(e_{nc})C''(x_{nc})]^2} < 0$, since $n > 1$ and $D''(NE_{nc}) > 0$.

(i) From the specification of $z^{\min}$, $\lim_{C''\to 0} z^{\min} \to \frac{-nD''(NE_{nc})B''(e_{nc})}{-D''(NE_{nc})B''(e_{nc})} = n$.

(ii) Using the specification of $z^{\min}$, for moderate values of $B''(e_{nc})$, $\lim_{C''\to\infty}(C''(x_{nc}) - B''(e_{nc})) \to C''(x_{nc})$. Therefore, we have that $\lim_{C''\to\infty} z^{\min} \to \frac{nD''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})}{D''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})} = \frac{nD''(NE_{nc})-B''(e_{nc})}{D''(NE_{nc})-B''(e_{nc})} = \frac{nD''\left(\widetilde{E}_{nc}\right)-B''(\widetilde{e}_{nc})}{D''\left(\widetilde{E}_{nc}\right)-B''(\widetilde{e}_{nc})} = \widetilde{z}_{x=0}$, .

7.3 Appendix III: Equilibrium choices in the two benchmark cases, assuming quadratic functions.

In the pure non-cooperative case, countries act individually and no coalition is formed. Since all countries are symmetric, at the equilibrium, they all choose the same level of emissions, denoted by e_{nc}, and abatement, denoted by x_{nc}. The system of reaction functions (8) and (9) yields the non-cooperative level of emission and abatement, $e_{nc} = \frac{(n+\delta)a}{\delta+(1+\gamma\delta)n}$ and $x_{nc} = \frac{na}{\delta+(1+\gamma\delta)n}$. Therefore, each country's net emissions are, $\varepsilon_{nc} = e_{nc} - x_{nc} = \frac{\delta a}{\delta+(1+\gamma\delta)n}$.[18] Aggregate net emissions are, $NE_{nc} = E_{nc} - X_{nc} = n\frac{a\delta}{\delta+(1+\gamma\delta)n}$.

In the case of full cooperation, we assume that in the first stage, the grand coalition is formed and is stable. In the second stage emission and abatement decisions are taken collectively, i.e., countries choose e_i and x_i so as to maximize aggregate welfare, $\sum_{i=1}^{n} W_i$, where W_i is given in (7).

We use the notation defined above and we denote equilibrium values of emission and abatement levels by a subscript c, $e_c = \frac{a(n^2+\delta)}{n^2+\delta+n^2\gamma\delta}$, $x_c = \frac{an^2}{n^2+\delta+n^2\gamma\delta}$. Thus, each country's net emissions ε_c are, $\varepsilon_c = e_c - x_c = \frac{\delta a}{n^2+\delta+n^2\gamma\delta}$.[19] Aggregate net emissions are, $NE_c = E_c - X_c = n\frac{a\delta}{n^2+\delta+n^2\gamma\delta}$.

Country i's net emissions are lower in the full cooperation case, i.e. $e_c - x_c < e_{nc} - x_{nc}$. In the full cooperation case, each country emits less ($e_c < e_{nc}$) and abates more ($x_c > x_{nc}$) relative to the non-cooperative case. That is,

[18] In the absence of net environmental damages, $c = 0$, emissions take their highest value, $e_{nc|c=0} = a$, and abatement effort approaches zero, $x_{nc|c=0} = 0$. Also, if abatement becomes costless, $d = 0$, countries choose the highest level of emissions, $e_{nc|d=0} = a$, since they can costlessly abate the total amount of emissions, $x_{nc|d=0} = a$.

[19] In the two extreme cases of $c = 0$ and $d = 0$, emissions and abatement take the same values as in the case of non-cooperation.

aggregate net emissions are lower when all countries cooperate. It can also be shown that aggregate welfare is higher under full cooperation.

7.4 Appendix IV: Equilibrium values in coalition formation, assuming quadratic functions

The first-order conditions yield signatories emission and abatement effort levels,

$$e_s = a\left(1 - ns\frac{\gamma\delta^2}{\Psi}\right), \qquad \text{(A IV. 1)}$$

$$x_s = ans\frac{\delta}{\Psi}, \qquad \text{(A IV. 2)}$$

where $\Psi = \Omega^2 + s^2\delta\left(1+\gamma\delta\right) > 0$ and $\Omega = \delta + (n-s)\left(1+\gamma\delta\right) > 0$. Note that, in the absence of environmental damages, $c = 0$, emissions take their highest value $e_s|_{c=0} = a$, and abatement effort is zero, $x_s|_{c=0} = 0$. Therefore, the signatories' net emissions are,

$$\varepsilon_s = a\left(1 - ns\frac{\delta(1+\gamma\delta)}{\Psi}\right) . \qquad \text{(A IV. 3)}$$

Thus, the coalition's total net emissions are, $NE_s = E_s - X_s = sa\left(1 - ns\frac{\delta(1+\gamma\delta)}{\Psi}\right)$.

Substituting e_s and x_s into the non-signatories' reaction functions, we derive the non-signatories' emission and abatement level,

$$e_{ns} = a\left(1 - n\frac{\gamma\delta\Omega}{\Psi}\right) = e_s + \frac{an\gamma\delta\left(\delta s - \Omega\right)}{\Psi} , \qquad \text{(A IV. 4)}$$

$$x_{ns} = \frac{an\Omega}{\Psi} = x_s - \frac{an\left(\delta s - \Omega\right)}{\Psi} . \qquad \text{(A IV. 5)}$$

From (A IV. 4) and (A IV. 5), the net emission level of the non-signatories is,

$$\varepsilon_{ns} = a\left(1 - n\frac{\Omega(1+\gamma\delta)}{\Psi}\right) = \varepsilon_s + \frac{an(1+\gamma\delta)\left(\delta s - \Omega\right)}{\Psi}. \qquad \text{(A IV. 6)}$$

The above imply that $e_{ns} \lessgtr e_s \Leftrightarrow \delta s \lessgtr \Omega$, $x_{ns} \gtrless x_s \Leftrightarrow \delta s \lessgtr \Omega$ and $\varepsilon_{ns} \lessgtr \varepsilon_s \Leftrightarrow \delta s \lessgtr \Omega$. We will explore this condition further when we compare the welfare of signatories to that of non-signatories. At this point, it is

noteworthy that since for small coalition sizes $\Omega > \delta s$ regardless of the value of the parameters, when coalition size is small non-signatories emit less and abate more than the signatories. For small coalition sizes, the leadership effect dominates the damage reduction effect.

Nonsignatories' aggregate net emissions are, $NE_{ns} = E_{ns} - X_{ns} = (n - s)\left[\varepsilon_s + \frac{an(1+\gamma\delta)(\delta s-\Omega)}{\Psi}\right]$. Therefore, global net emissions are, $NE = E - X = \sum_{i=1}^{n}(e_i - x_i) = an\frac{\delta\Omega}{\Psi}$.

Unlike the non-cooperative and the full cooperation cases, in which the level of emissions are strictly positive, $e_{nc} > 0$ and $e_c > 0$, in the coalition formation case, we have to restrict the parameters of the model in order to guarantee interior solutions. Therefore, we need to restrict the parameters so that both are positive, given that emissions correspond to production and consumption, which cannot be negative. The following Lemma establishes the necessary conditions for interior solutions. The proof is presented in Appendix III.

Lemma 1 *In the case that abatement is not available, that is, $\delta \to \infty$, $e_s > 0$ and $e_{ns} > 0$ if and only if $0 < \gamma < \frac{4}{n(n-4)}$ and $n > 4$. As abatement becomes relatively inexpensive, $e_s > 0$ and $e_{ns} > 0$ hold for higher values of γ, which are increasing as δ decreases.*

Proof. From (A IV. 1) we have that $e_s > 0 \Leftrightarrow \Omega^2 + \delta s^2 > (n-s)s\gamma\delta^2 \Longrightarrow \delta^2 + \delta s^2 + 2\delta(n-s) + (n-s)\left[\left(1 + \gamma^2\delta^2 + 2\gamma\delta\right)(n-s) - \gamma\delta^2(s-2)\right] > 0$. We derive the size of the coalition that minimizes this expression $A(s) = \delta^2 + \delta s^2 + 2\delta(n-s) + (n-s)\left[\left(1 + \gamma^2\delta^2 + 2\gamma\delta\right)(n-s) - \gamma\delta^2(s-2)\right]$, which is $\underline{s} = \frac{2\delta(1+\gamma\delta)+n(2+\gamma\delta(4+\delta+2\gamma\delta))}{2(1+\gamma\delta)(1+\delta+\gamma\delta)}$. Substituting the value $\underline{s}$ into the $A(s)$, we get $A(\underline{s}) = \frac{a\left(4\delta^2(1+\gamma\delta)+4n\delta(1+\gamma\delta)(2+\gamma\delta)+n^2(4+\gamma\delta(8+\gamma(4-\delta)\delta))\right)}{4\delta^2(1+\gamma\delta)+8n\delta(1+\gamma\delta)^2+n^2(4+\gamma\delta(12+\gamma\delta(12+\delta+4\gamma\delta)))}$. Then $A(\underline{s}) > 0$, if $4\delta^2(1+\gamma\delta) + 4n\delta(1+\gamma\delta)(2+\gamma\delta) + n^2(4+\gamma\delta(8+\gamma(4-\delta)\delta)) > 0$, which is definitely true for $\delta < 4$. Clearly this is a sufficient but not necessary condition. Notice that if we divide the last expression with δ^3 and let $\delta \to \infty$, the expression reduces to the condition presented in Proposition 1, in Diamantoudi and Sartzetakis (2006). That is, if abatement is not available (extremely expensive), then the only option is to decrease the economic activity and thus emissions, which restricts the size of the coalition to a maximum of four countries as shown in Diamantoudi and Sartzetakis (2006). However, as the cost of abatement

decreases substantially, a large part of the necessary emission reduction is achieved through abatement, which eases the restriction on γ. That is, we can have a situation with substantial damages relative to benefits from emissions so that countries want to take strong action and, at the same time, have available abatement technologies that can achieve the necessary emission reductions at a reasonable cost.

From (A IV. 4) we have that $e_{ns} > 0 \Longleftrightarrow (\Omega - n\gamma\delta)\,\Omega + s^2\delta(1+\gamma\delta) > 0$. Dividing the expression by δ^2 and denoting by $\tau = \frac{1}{\delta}$, yields, $(n-s)^2(\phi+\gamma)^2 + [s^2 - (n-s)(\gamma n - 2)]\,(\phi+\gamma) - (\gamma n - 1) > 0$. If abatement is not available, that is $\delta \to \infty$, which implies $\tau \to 0$, the expression reduces to the condition presented in Proposition 1, in Diamantoudi and Sartzetakis (2006). However, as the cost of abatement decreases, emission of nonsignatories are positive for higher values of γ. Although the necessary condition can be derived using the above technique, it is very complicated and we omit it. It is important to notice that the condition constraints both the values of γ and δ. ■

The intuition of the above result is clear: as abatement becomes available at low cost, countries both in and outside the coalition are able to decrease their net emissions by engaging in abatement, keeping their emissions and thus their production positive, even when damages from emissions are relatively high. This will prove very important for determining the maximum size of the stable coalition that follows.

Substituting the equilibrium values of the choice variables from (A IV. 1), (A IV. 2), (A IV. 4) and (A IV. 5) we derive the indirect welfare function of signatories (w_s) and non-signatories (w_{ns}),

$$w_s = \frac{ba^2}{2}\left(1 - \frac{\gamma n^2\delta^2}{\Psi}\right), \qquad \text{(A IV. 7)}$$

$$w_{ns} = \frac{ba^2}{2}\left[w_s - \frac{\gamma\delta n^2\left(\delta^2 s^2 - \Omega^2\right)(1+\delta\gamma)}{\Psi^2}\right]. \qquad \text{(A IV. 8)}$$