Int Environ Agreements (2022) 22:659–672
https://doi.org/10.1007/s10784-022-09574-7

Designed to be stable: international environmental agreements revisited

Nahid Masoudi[1]

Accepted: 24 February 2022 / Published online: 15 April 2022

Abstract
In a three-stage game, we revisit the non-cooperative coalition approaches into international environmental agreements by tackling a fundamental design flaw in these approaches. We show how a treaty can effectively remove the free-riding problem from its roots by farsightedly choosing its members' emissions. We prove that under this approach, the grand coalition is a self-enforcing equilibrium. We will argue how the modified timing of the coalition game suggested in this article is more realistic and consistent with real-world practices. Another advantage of the farsighted rule is its simplicity and applicability to all coalition game settings, regardless of whether agents are homogeneous or heterogeneous.

Keywords International agreements · Transboundary pollution · Strategic behaviors · Farsighted Stackelberg · Farsighted rule

JEL Classifications F53 · Q54 · C72

1 Introduction

The standard non-cooperative coalition theory that is often applied in the context of international environmental agreements (IEAs) comes to the grim conclusion—known as the paradox of cooperation—that self-enforcing IEAs may only have a small number of members, or if they sustain large numbers, then the gains from cooperation are minimal. This paper focuses on the former case, where the free-riding incentives make larger coalitions (including the grand coalition) internally unstable. That means the cost and benefit functions are such that, under the standard approaches, a member of the large coalition will find leaving the treaty more rewarding than staying (see, d'Aspremont et al. 1983; Hoel 1992; Carraro and Siniscalco 1993; Barrett 1994 among many others, also for a literature

I am grateful to the editor, Dr. Joyeeta Gupta, and two anonymous reviewers for their helpful guidance and constructive comments.

✉ Nahid Masoudi
nmasoudi@mun.ca

[1] Department of Economics, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada

review on the stability of IEAs see, e.g., Finus 2003; Chander and Tulkens 2006; Finus 2008; Marrouch and Chaudhuri 2016; Eyckmans and Finus 2006; Finus et al. 2021). In this paper, we argue that the instability encountered by large coalitions is due to the failure of the standard approaches to contemplate the impact of the free-riding incentives. Note that when we are dealing with a situation where the standard non-cooperative approach leads to unstable–unachievable–large IEAs, this approach is strategically failing to maximize the members' welfare by trapping them in a prisoners'-dilemma-like situation. Hence focusing on unstable, non-self-enforcing mechanisms and dismissing the fact that the end result of the suggested solution will not achieve the intended goal is impractical, if not irrational.

We redefine the coalition game to overcome this drawback by allowing the members to carefully weigh the free-riding incentives into their choices. That means they choose their joint-welfare-maximizing emissions such that free-riding becomes ineffectual and the incentive of a member to leave the coalition is eliminated. To that end, in a three-stage non-cooperative coalition game, we use a specially designed constraint optimization mechanism, where the cost of achieving stability is minimal for the members. Moreover, we argue that since a treaty only sustains if it is stable, complying with the suggested rule is in members' self-interest. In other words, our solution is an equilibrium.

To define the members' emissions, we begin from the grand coalition and then solve all partial coalitions by working our way back to the smaller ones.[1] For simplicity, we call the members' complete emissions profile the "farsighted rule." Without loss of generality, we assume that an international climate agency, e.g., the United Nations Climate Change (UNCC), is responsible for announcing the farsighted rule in stage one. However, there is no assumption that this entity has any supranational authority to implement/enforce emissions levels. We only assume that the agency coordinates with countries and provides advice and information during the pre-agreement negotiations, i.e., during stage one. This assumption is consistent with real-world practices, e.g., the organizing committee of the UN Climate Change Conference of Parties (COP) 21 (which took place from 30 November to 11 December 2015) provided information, suggested commitments, and negotiated the Paris Agreement, which later got adopted by 196 Parties on 12 December 2015. This article proposes that the agency informs parties about the free-riding incentives during such pre-agreement period (e.g., COP XX) and presents the members' emissions plan for all possible treaty sizes, i.e., the farsighted rule. Consequently, we will modify the timing of the game from the standard approaches.

The key result of our approach is to remove an inherent flaw from the non-coalition approaches in the formation of IAES, and consequently, to ensure the stability of the grand coalition in a simple and easy-to-implement fashion. Furthermore, we argue that our departure from the standard literature, i.e., our agency assumption and the modified timing of the game, makes the non-cooperative coalition games structure more realistic and better consistent with real-world practices. Moreover, since countries make their membership choices "informed," i.e., after getting all the information and learning about the farsighted rule, the formation of the grand coalition is expected to be immediate. Hence, we have a strong rationale for our departure from the standard coalition games in this respect. However, for the those readers who are more interested in the standard settings, later in the article, we will present two alternative approaches that can be followed to achieve the outcome that corresponds to the farsighted rule in a setting that reinstates the standard Stackelberg

[1] Beginning from the grand coalition is not unusual in the literature, e.g., in studying the equilibrium binding agreements Ray and Vohra (1997) also begin from the grand coalition.

timing. Nevertheless, we argue that the farsighted rule and the revisited timing have notable advantages over the standard frameworks.

Another noteworthy aspect of our work is that while we begin with a symmetric setting, we show that the symmetry assumption is only used for simplicity and can be relaxed. Indeed, our mechanism is applicable to all cases regardless of the shape of the benefit and cost functions or the presence of any asymmetry among agents.

We must state at the outset that the focus of this article is on solving the free-riding issue as the force behind breaking large coalitions, and therefore, we always assume that the benefit and cost functions are such that the grand coalition is unstable under the standard non-cooperative coalition settings. Hence, all the statements, results and proofs are based on this assumption.

The rest of the article is organized as follows. The model is presented in Sect. 2. Section 3 describes the solution and sets up the process to find the treaty's farsighted rule. In Sect. 4, we comment on the comparison of the welfares and emissions under the farsighted rule and standard Stackelberg and Cournot-Nash solutions. Section 5 is designated to illustrate how we can guarantee full cooperation as the only equilibrium solution of the coalition game. We explain the insignificance of the symmetry assumption for the treaty's farsighted rule in Sect. 6. Section 7 will discuss the merits behind our assumption regarding the agency and the timing of the game. To complete the analysis, an illustrative example is provided in Sect. 8. Section 9 concludes the article.

2 Setup

2.1 Notation

Consider a set $K = \{1, \dots, k\}$ of k countries sharing a common environment. To simplify the notation and presentation of the model, at first we assume that countries are symmetric. However, later we will explain that this assumption has no qualitative impact on the main results of the article and can be relaxed. Suppose a subset of countries $S \subset K$ have formed a treaty, and let s denote the size of the treaty, $s = |S|$. Denote a representative member's emission and welfare by e_s^m and w_s^m, and a representative non-member's by e_s^n and w_s^n, where the subscript s refers to the size of the treaty. Assuming a one-to-one relationship between emission and production, we can define a country's benefit as a concave and increasing function of its emission, denoted by $B\left(e_s^x\right)$, where $x = m$ if this country is a member and $x = n$ otherwise. Suppose that the environmental damage can be presented by a convex and increasing function of the global emissions and denoted by $D(se_s^m + (k-s)e_s^n)$. Therefore, an individual's net welfare w_s^x is: $w_s^x = B(e_s^x) - D(se_s^m + (k-s)e_s^n)$.

To complete our notation, denote the treaty's farsighted emission rule by a $(k-1) \times 1$ vector $E^m = \left(e_2^m, \dots, e_k^m\right)$, defining the members' emissions for all possible treaty sizes of $s = 2, \dots, k$.[2]

[2] We are ruling out the trivial case of a treaty of one member, clearly such an inclusion will not change any of the results.

2.2 Timing

Consider a three-stage game as follows:

- Stage one: The agency prepares a plan (the farsighted rule) defining the members' emissions for all treaty sizes.
- Stage two: Assuming an open-membership framework and given the treaty's farsighted rule from stage one, countries independently and simultaneously decide whether to sign the treaty and become a member or not.
- Stage three: Non-members act as singletons and choose their emission level by maximizing their individual welfare, taking the treaty's size and its members emissions as given, while each member emits according to the farsighted rule.

It is essential to point out that what we are suggesting as stage one (and the agency) mainly refers to the pre-negotiations and discussions ahead of the signing of international agreements, similar to what we observe in all real-world cases. There are no implications that such an agency has any supranational power. Later, in Sect. 7, we will discuss alternative settings in the absence of such agency.

Stability Definition:

We use the internal-external stability notion introduced by d'Aspremont et al. (1983), where the treaty S is (weakly) internally stable if no member benefits from deviating unilaterally and leaving the treaty, i.e.:

$$w_s^m \geq w_{s-1}^n, \tag{1}$$

and it is (weakly) externally stable if no non-member benefits from joining the treaty, i.e.:

$$w_{s+1}^m \leq w_s^n. \tag{2}$$

3 Solution

To find the sub-game-perfect Nash Equilibria of the game, we follow a backward induction procedure beginning from stage three as explained in the following.

Stage three: Given the size of the treaty formed in stage two, each member adheres to the farsighted rule and emits e_s^m units. Each non-member acts as a singleton choosing its emission by maximizing its welfare, taking the treaty's size and emissions as given. Given our assumptions about the shapes of the benefit and damage functions, and using a non-member's first-order condition, we can find a nonmember's emission as a function of the members' emissions, i.e., we have $e_s^n = g(e_s^m)$, where $\frac{dg}{de_s^m} < 0$ (due to the negative externality nature of the emissions).

Stage two: Countries, independently and simultaneously, decide upon membership, i.e., they choose whether to become a member of the treaty or not. Following the literature (e.g., Long 1992; Barrett 1994 among others) and the real-world practices used for treaties such as the Paris Agreement, we assume that membership is open, meaning countries can freely choose to join.

Stage one: The agency announces the farsighted rule. The fundamental difference between this game and the standard non-cooperative IEAs comes from how members choose their

emissions, i.e., how the farsighted rule is defined to effectively curb the free-riding incentives with minimum impacts on the benefits of forming a treaty. To that end, the agency (and later the treaty members by adhering to the farsighted rule) take into account that the free-riding temptations may break large treaties and, consequently, drop the global welfare to a prisoners' dilemma like situation (i.e., no treaty or a treaty with very few members). We implement a backward procedure to accomplish this purpose, beginning from finding the members' emissions under the grand coalition and then rolling down to solve the emissions for the rest of the treaty sizes.

For the grand coalition, we have the standard full-cooperation solution, where e_k^m maximizes the global welfare, i.e.:

$$e_k^m = \text{argmax}_{e_k^m>0} k\left\{B\left(e_k^m\right) - D(ke_k^m)\right\}. \tag{3}$$

Then, to assure the internal stability of the grand coalition, e_{k-1}^m must be chosen such that no member could be better off by leaving it unilaterally. That is, the welfare of staying as a member must be at least as high as the welfare of leaving, i.e., $w_k^m \geq w_{k-1}^n$. Therefore, e_{k-1}^m is given by:

$$\begin{aligned} e_{k-1}^m = \text{argmax}_{e_{k-1}^m} (k-1)\left\{B\left(e_{k-1}^m\right) - D((k-1)e_{k-1}^m + e_{k-1}^n)\right\}, \\ \text{such that } w_k^m \geq w_{k-1}^n. \end{aligned} \tag{4}$$

Note that since $e_{k-1}^n = g(e_{k-1}^m)$—a representative nonmember's best response—then the right-hand-side of the constraint, w_{k-1}^n, is only a function of e_{k-1}^m, while the left-hand-side, w_k^m, is known from problem (3)'s solution. Moreover, we know that this constraint is binding since the a priori assumption of this article is that the grand coalition is unstable under the standard coalition approaches. Therefore, the solution of problem (4) is the value of e_{k-1}^m that satisfies $w_k^m = w_{k-1}^n$. Now, using this solution, we can calculate the total emissions, $(k-1)e_{k-1}^m + g(e_{k-1}^m)$, and w_{k-1}^m. Then we can follow a similar procedure to find e_{k-2}^m (that is simply the value that solves $w_{k-1}^m = w_{k-2}^n\left((k-2)e_{k-2}^m + g(e_{k-2}^m)\right)$, where w_{k-1}^m is given) and then roll down to find the rest of the elements of the farsighted rule. In general, $\forall s \in \{3..., k\}$, e_{s-1}^m is the solution to the equation $w_s^m = w_{s-1}^n$, where at the time of solving for e_{s-1}^m the values of e_s^m, e_s^n, and w_s^m are known. Upon completion of this process, we have the farsighted rule or the vector $E^m = \left(e_2^m, \ldots, e_k^m\right)$. As it is formally presented in Proposition (3.1), a coalition $S \subset K$, formed using E^m, not only is internally stable but also is (weakly) externally stable as well.

Proposition 3.1 *Any coalition $S \subset K$ formed according to the farsighted rule, including the grand coalition, is (weakly) stable.*

Another noteworthy comment here is that while the treaty's farsighted rule defines the emissions for all the treaty sizes, nevertheless, if the only concern is assuring stability of the grand coalition, then finding the emissions for the treaty of size $k-1$ (i.e., problem 4) suffices without a need to solve for the rest of the treaty sizes and the complete vector.

Also note that we are proposing this mechanism only for the situations where the benefit and cost functions are such that the grand coalition (large coalitions) is (are) unstable under the standard non-cooperative solutions; otherwise, the proposed treatment will be redundant. Moreover, we need to emphasize that if a treaty S is unstable under the standard non-cooperative settings, one cannot solve for the members' emission using the internal stability rule for that same treaty as defined by the problem:

$$\begin{aligned} &\max_{e_s^m} \; s(B(e_s^m) - D\big(se_s^m - (k-s)e_s^n\big) \\ &\text{subject to } w_s^m \geq w_{s-1}^n, \end{aligned} \tag{5}$$

Note the subtle difference between problem (5) and problem (4) that is the one we solve for the farsighted rule. If a treaty is unstable under standard approach, then the solution set for problem (5) is empty, while problem (4) has a solution that can be find easily. Therefore, while a great advantage of the farsighted solution is its simplicity and clarity, it has a notable and significant novelty in how it tackles the free-riding problem and successfully defines a mechanism that makes the socially optimal solution, i.e., the grand coalition stable.

4 Standard IEA approaches versus the treaty's farsighted rule

The non-cooperative IEA literature takes two different behavioral approaches regarding the interaction between the treaty members and the non-members: (i) members act as the collective Stackelberg leader (choose first) in the emission game, while all non-members act as singleton followers (observe the treaty's choice and then make their choices), as, e.g., in Barrett (1994); ii) members act as a collective player and non-members act as singletons in a Nash-Cournot setting choosing their emissions simultaneously, as, e.g., in Hoel and Schneider (1997). For simplicity, let us call the former approach Stackelberg, the latter Nash-Cournot, and the approach proposed in this article the farsighted Stackelberg.

Denote the emission and welfare of a representative country under Nash-Cournot by $e_{s,C}^x$ and $w_{s,C}^x$, and under Stackelberg by $e_{s,St}^x$ and $w_{s,St}^x$, respectively, where the superscript $x = \{m, n\}$ refers to member and nonmember. This article assumes that the grand coalition is unstable under the Nash-Cournot and the Stackelberg assumption. This means a deviating country has higher welfare as a free-rider than as a cooperator in the grand coalition given that the remaining coalition of $k-1$ countries react either as a collective Stackelberg or as a collective Nash-Cournot player. To eliminate the free-riding incentive–to attain the outcome that corresponds to the farsighted rule–the remaining $k-1$ countries must increase their emissions so that the welfare of the defector drops back to its level if it has remained a member of the grand coalition. Hence, compared to the Stackelberg case, the welfares of the countries in the remaining coalition of size $k-1$ are clearly reduced. In this context, note that the deviating country will reduce its emission in response to the increased emissions of the shrunk coalition. In the Nash-Cournot case, however, the welfare effect for the remaining coalition members is not clear a priori: Starting from the Nash-Cournot equilibrium, when the coalition members increase their emissions, their welfares will increase until they reach their emission levels in the standard Stackelberg equilibrium. From that level on, their welfares will fall through the further increase in the coalitions' emissions. Proposition (4.1) presents these results formally.

Proposition 4.1 $\forall s \in \{2, \ldots, k-1\}$:

- $w_s^m \leq w_{s,St}^m$ and $w_s^n \leq w_{s,St}^n \leq w_{s,C}^n$;
- $e_s^m \geq e_{s,St}^m \geq e_{s,c}^m$ and $e_s^n \leq e_{s,St}^n \leq e_{s,C}^n$.

Proof See Appendix (A). □

As presented by Corollary (A), due to the internal instability assumption for the grand (a large) coalition, the farsighted rule will not improve the global welfare compared to the standard Stackelberg assumption for a treaty of size s. The impact on global emissions is not straightforward. It depends on how much of the reduction in the non-members' emissions are compensated with the members' emissions increase.

Corollary 4.1 $\forall s \in \{2, \ldots, k-1\}$:

- $sw_s^m + (k-s)w_s^n \leq sw_{s,St}^m + (k-s)w_{s,St}^n$;

According to Proposition (3.1), any treaty formed in stage two, regardless of its size, is stable; therefore, the game's outcome is going to be an equilibrium. However, Proposition (4.1) reveals that the global welfare for partial coalitions under standard Stackelberg will be at least as high as the global welfare under the farsighted rule. Nevertheless, it is worth emphasizing that, as long as self-enforcing is a necessary characteristic for a treaty to be implemented, the results reported in Proposition (3.1) will not undermine the value and importance of the farsighted approach because we are proposing this mechanism for the cases that the paradox of cooperation prevails by making the grand coalition unstable and, hence, unattainable. In other words, a priori, we know that a large coalition will not form under the standard approach. Therefore, the farsighted rule is indeed *welfare improving* by making cooperation viable and the formation of a large treaty attainable. More importantly, the now stable grand coalition solution under the farsighted rule coincides with the global optimal, i.e., the standard fully cooperative solution, i.e., $w_k^m = w_{k,St}^m = w_{k,C}^m$ and $e_k^m = e_{k,St}^m = e_{k,C}^m$.

5 Grand coalition as the unique solution

Up to here, we show that the grand coalition will be self-enforcing under the farsighted rule. However, this rule also makes all partial coalitions (weakly) stable. Therefore, one may suggest that there is no guarantee to end up with the grand coalition. A straightforward argument against this suggestion is that the stability mechanism as designed here renders participation a (weakly) dominant strategy for all countries. By definition, the grand coalition is the socially optimum outcome (or the Pareto-dominance among all the stable agreements); therefore, outsiders have an incentive to join the agreement irrespective of how large the already existing coalition is. Moreover, in Sect. 3, we used the stability condition in its weak sense as in, e.g., Barrett (1994). However, imposing the stability in its strong version (strict inequality) similar to, e.g., in Carraro and Siniscalco (1993) will make the grand coalition the only stable solution of the coalition game. The strong internal stability constraint makes all the partial coalitions strictly externally unstable, and, as a result, the grand coalition becomes the only stable coalition and, therefore, the only equilibrium solution to the coalition game. This result is formally presented in Corollary 5.1.

Corollary 5.1 *If the strong version of the coalition stability condition is used to define the farsighted rule, then the grand coalition will be the unique equilibrium solution of the coalition game.*

We know that imposing the strong inequality as the constraint will result in equal or lower welfares (for coalition members and globally) and higher global emissions for any partial coalition S, where $S \subset N$ and $S \neq N$, compared to the original rule. Nevertheless, similar to what we argued in the previous section, this is not a concern because the adjusted rule guarantees the formation of the grand coalition, i.e., the globally optimal solution, as the only equilibrium of the coalition game.

6 Easing the symmetry assumption

This section discusses another significant feature of the farsighted rule: its applicability to both *symmetric and asymmetric* settings. We began our analysis by a model under the symmetry assumption to save in notation and simplify the presentation of the model. Here, we show that our results are not dependent on the symmetry assumption. To that end, suppose that countries are asymmetric, yet by assumption, the grand coalition is unstable under the standard Stackelberg setting. To accommodate the asymmetry assumption, let us modify the notation by denoting the member i's emission and welfare when all the countries except for country j sign the treaty by $e^m_{i,K\backslash j}$ and $w^m_{i,K\backslash j}$, and emission and welfare of the fringe country j by $e^n_{j,K\backslash j}$ and $w^n_{j,K\backslash j}$, respectively. To eliminate j's incentive for leaving the grand coalition, the farsighted rule determines the emissions of the remaining members of the treaty $\{K\backslash j\}$ in such a way that the defector j's welfare drops back to its level as it has stayed a member of the grand coalition, i.e., $w^m_{j,K}$. Therefore, the farsighted rule solves the following problem for all $j \in K$:

$$\begin{aligned} &\max_{e^m_{i,K\backslash j}, i\in K \& i\neq j} \sum_{i\neq j} \left(B_i(e^m_{i,K\backslash j}) - D(e^n_{j,K\backslash j} + \sum e^m_{i,K\backslash j}) \right), \\ &\text{such that, } w^n_{j,K\backslash j} \leq w^m_{j,K}. \end{aligned} \tag{6}$$

Upon finding the solution to the above set of problems, the treaty's farsighted rule will have a complete plan of action for all the treaties of size $k-1$ with all possible members combinations. Obviously, for some of such treaties, the farsighted emission rule can coincide with the standard solution since not all members necessarily have free-riding incentives in the presence of heterogeneity. If interested to have the complete farsighted rule profile, similar to what we did in the case of symmetric countries, we can repeat the procedure for all the treaties $\{K\backslash j, z\}$, where, $j \in K$, $z \in K$, and $j \neq z$. However, one may stop here since the $e^m_{i,K\backslash j}$, defined by problem (6) for all $j \in K$, suffices to ensure the stability of the grand coalition.

7 Easing the agency's assumption and modifying the timing of the game

Another simplifying assumption that we have made in this article is introducing an international agency that announces the farsighted rule ahead of the agreement formation. This entity possesses no supranational authority, and its role is solely coordination and information provision. In this section, we show how one can relax this assumption, but first, we explain our rationale for this departure from the literature by highlighting the key advantages of our approach. Firstly (and the least importantly), this approach makes our arguments and presentation of the farsighted rule and the emissions vector E^m more intuitive and easier to motivate. Secondly, the pre-negotiations and discussions through a coordinating agency are the standard, real-world practices of any international agreement formation, e.g., during the UN Climate Change Conference of Parties (COP) summits. Thirdly, by redefining the timing of the game, this approach resolves another inherent flaw in the standard approaches used in the IEAs literature. In fact, the mainstream assumption that countries just independently (and randomly) make their membership choices prior to having any discussions or information, not only is oddly unconvincing and unrealistic; it also renders itself subject to many unanswered questions. How and why do we expect countries to make their decisions in advance and individually, without consulting with each other and knowing about other countries' choices? What if s countries signed the agreement and it turned out a treaty of z, $z \neq s$, members is stable? Why and how do we expect that a treaty of z will form? In other words, at what stage does that happen? Do we go back to stage one? Do we abandon the game right here? If we have to go back to stage one, that is the stage countries choose whether to be a member or not, why did we not begin the game from a coordination stage in the first place? The standard framework is unable to provide an apparent response to such questions. In contrast, in our setting, countries come to the signing table (stage two) informed–after learning about the farsighted rule during the pre-signature period (stage one). Then, given that by design joining the treaty is the (weakly) dominant strategy for all, the formation of the grand coalition is expected to be immediate. Indeed, the agency and the pre-choice-making stage provide information that puts matters into perspective and allows countries to make better choices in the first place.

We want to add that if one insists on following the standard timing and abandoning the agency from the game, it is possible and can be done without having any qualitative impacts on the results and the key messages. And there are two alternative methods to accomplish this. One is to present a three-stage game as follows. Stage one: the membership game as in the standard Stackelberg. Stage two: members act as the collective "farsighted" Stackelberg leader, meaning not only do they agree on the emissions for the treaty formed in stage one, they solve and agree on the entire farsighted rule for all the possible treaty sizes. Finally, nonmembers choose their emissions acting as singleton followers in stage three.

The other alternative is to introduce a three-stage game with the same order as the first alternative but with different assumptions for the events in stage two as follows. In this sateg, members choose their emission as per standard Stackelberg given the treaty size formed in stage one. However, they also agree on what would be the level of their

emissions if one party chooses to deviate unilaterally, i.e., they agree on e^m_{s-1} as well. And e^m_{s-1} will be chosen such that it eliminates the incentive to leave the treaty, i.e., it comes from the solution to the following problem:

$$e^m_{s-1} = \text{argmax}_{e^m_{s-1}} (s-1)\{B(e^m_{s-1}) - D((s-1)e^m_{s-1} + e^n_{s-1})\}, \quad \text{such that } w^m_{s,St} \geq w^n_{s-1}. \tag{7}$$

This latter alternative is a straightforward one-step extension of the standard Stackelberg that effectively neutralizes the free-riding efforts.[3] Moreover, the e^m_{s-1} can be seen as a punishment strategy with huge advantages compared to some of the punishment strategies suggested in the literature (e.g., Chander and Tulkens (1997)): It ensures stability through a punishment profile that minimizes the losses for the punishing parties–the countries in the remaining coalition after some country has left. Note that this punishment is credible and self-fulfilling as the remaining parties find adhering to it in their own self-interest. Also note that their alternative is to follow the standard Stackelberg strategy, where the free-riding incentives will prevail against the benefits of cooperation. Consequently, adhering to e^m_{s-1} is rational. Another way of putting this is to say signatories come to the negotiation regarding their emissions with the mindset that "if I negotiate a contract about emission reduction, then I only negotiate a contract which is stable."[4]

Both of the above alternatives are interesting and effectively curb the free-riding issue. However, we choose to stay with the agency assumption and the redesigned timing of the game as our main framework because of the said advantages and the fact that all the criticism listed above applies to these two alternatives as well.

8 An illustrative example

In this section we present an illustrative example using the widely used quadratic model. Suppose the benefit function can be presented by $B(e^x_s) = \alpha e^x_s - \frac{1}{2}(e^x_s)^2$, $x \in \{m, n\}$, and the environmental damage is given by $D = \frac{\beta}{2}(se^m_s + (k-s)e^n_s)^2$. Therefore, a representative country's welfare is:

$$w^x_s = \alpha e^x_s - \frac{1}{2}(e^x_s)^2 - \frac{\beta}{2}(se^m_s + (k-s)e^n_s)^2. \tag{8}$$

Stage three:

Assuming s countries have signed the treaty in stage two, each member will emit according to the farsighted rule, while each non-member takes choices made in stages one and two as given and decides upon its emission by maximizing its own welfare. From the representative non-member's first-order-condition we have:

$$e^n_s(e^m_s) = g(e^m_s) = \frac{\alpha - \beta s e^m_s}{1 + \beta(k-s)}. \tag{9}$$

Stage two:

[3] Note that if the formed treaty turns out to be internally stable under standard Stackelberg, then problem (7)'s solution also coincides with the Stackelberg solution.

[4] I thank an anonymous referee for suggesting this phrase.

Table 1 Cournot results

s	$e^m_{s,C}$	$e^n_{s,C}$	$w^m_{s,C}$	$w^n_{s,C}$	Total emissions	Total welfare
1	–	6.67302	–	16.8098	46.71114	117.6686
2*	6.35466	6.67733	16.8548	17.011	46.09597	118.7646
3	6.05683	6.68561	16.9951	17.3905	44.91293	120.5473
4	5.78906	6.69726	17.2204	17.9078	43.24802	122.605
5	5.55761	6.71152	17.5155	18.5141	41.21109	124.6057
6	5.36537	6.72756	17.8624	19.1613	38.91978	126.3357
7	5.21221	–	18.2427	–	36.48547	127.6989

Table 2 Stackelberg results

s	$e^m_{s,St}$	$e^n_{s,St}$	$w^m_{s,St}$	$w^n_{s,St}$	Total emissions	Total welfare
1	–	6.67302	–	16.8098	46.71114	117.6686
2*	6.37592	6.67704	16.8551	16.9977	46.13704	118.6987
3	6.08107	6.68511	16.9954	17.3681	44.98365	120.4586
4	6.83022	6.69665	17.2207	17.8809	43.33615	122.5255
5	5.57463	6.71093	17.5157	18.4897	41.29501	124.5579
6	5.37447	6.72718	17.8624	19.1464	38.9740	126.3208
7	5.21221	–	18.2427	–	36.48547	127.6989

Table 3 Farsighted Stackelberg results

s	e^m_s	e^n_s	w^m_s	w^n_s	Total emissions	Total welfare
1	–	6.67302	–	16.8098	46.71114	117.6686
2	6.44205	6.67615	16.8528	16.9560	46.26485	118.4856
3	6.35375	6.67954	16.9560	17.1135	45.77941	119.322
4	6.25156	6.68458	17.1135	17.3438	45.05998	120.4854
5	6.11656	6.69223	17.3438	17.6867	43.96726	122.0924
6	5.9050	6.70505	17.6867	18.2427	42.13505	124.3629
7	5.21221	–	18.2427	–	36.48547	127.6989

In this stage, countries simultaneously choose whether to become members of the treaty or not. The collective decisions made in this stage define the treaty's size.

Stage one:

In this stage, the farsighted Stackelberg leader chooses the emission rule for every possible treaty size. The grand coalition's solution coincides with the standard fully cooperative solution, i.e., $e^m_k = \frac{\alpha}{1+\beta k^2}$ and $w^m_k = \frac{\alpha^2}{2(1+k^2\beta)}$. The e^m_{k-1} is the solution to

$w_k^m = w_{k-1}^n\left(e_{k-1}^n\left(e_{k-1}^m\right)\right)$, where simple algebra yields: $e_{k-1}^m = \frac{\alpha(k(1+\beta)-\sqrt{(1+\beta)(1+\beta k^2)})}{(k-1)\sqrt{(1+\beta)(1+\beta k^2)}}$ and $e_{k-1}^n = \alpha - \frac{k\alpha\beta}{\sqrt{(1+\beta)(1+2k\beta)}}$.[5] Using this solution, we can find the e_{k-2}^m which is the value that satisfies $w_{k-1}^m = w_{k-2}^n(e_{k-2}^n\left(e_{k-2}^m\right))$. Continuing this procedure for the rest of the treaty sizes completes the farsighted emission profile. However, the equations for the rest of E^s elements are not reported because they have long terms and do not convey extra information.

Let us look at a simple numerical example to make the results easier to see, where we have $k = 7$, $\alpha = 7$ and $\beta = 0.007$. Tables 1 and 2 present the emissions and welfares for different treaty sizes under Cournot and Stackelberg approaches, respectively. The only self-enforcing coalition is very small, with only two members in both cases. Table 3 presents the farsighted rule and the corresponding welfares and emissions for the members and non-members. We can see that the global welfare for partial treaties is lower under the agency than under Cournot and Stackelberg approaches, however, unlike those cases, all treaties, including the grand coalition, are stable.

9 Concluding remarks

The theoretical literature on the IEAs has not been that optimistic regarding the success of such agreements since these agreements, e.g., the Paris Agreement and the Kyoto Protocol, pursue a strategy of voluntary reduction in emissions. Therefore, strong free-riding incentives always make large and effective agreements unstable. Here we propose a simple solution that tackles the free-riding problem from its roots, i.e., by changing the design of the IEAs to make free-riding inherently ineffective while imposing a minimum cost on the countries that remain in the agreement. In this new setup, similar to the current literature, we assume that member countries coordinate their emissions; however, they do this farsightedly by taking the free-riding incentives into account.

We show how treaty members can maximize their joint welfare such that no country could be better off by leaving the treaty unilaterally. Given that the coalition members are aware of the strong free-riding temptations that destabilize their coalition and force a prisoners' dilemma-like situation on them, consideration of these temptations in their choice of emission is, in fact, rational. Consequently, members will find complying with the suggested rule in their self-interest since it makes free-riding ineffectual. We also show that the grand coalition is a self-enforcing equilibrium of the coalition game under the farsighted rule.

The internal-external stability is not the only concept of stability in the IEA literature. For example, the γ-core is another competing concept (introduced in this literature by Chander and Tulkens 1995, 1997) that solves the free-riding problem by assuming that the remaining members of the treaty threaten to break apart into singletons. Hence, the free-rider will prefer to stay in the coalition rather than end up with the non-cooperative Nash equilibrium. Generally speaking, the internal-external stability assumption is more conservative than the extreme punishment of breaking the treaty altogether, as suggested in the γ-core stability. Hence, achieving stability using internal-external stability is more challenging than in the latter case. In this article, while we adhere to the conservative assumption of internal-external stability, we successfully attain stability for the fully cooperative solution. Consequently, a substantial advantage of the farsighted rule over the γ-core is that in the former, members do not need to take the drastic measure of breaking the treaty altogether as a punishment for free-riding to sustain cooperation, instead what they do is that from the beginning they choose their

[5] A sufficient condition for $e_{k-1}^n > 0$ is $\beta \leq \frac{2}{k-2}$. As for rest of the article, I am assuming the model parameters can be calibrated such that an interior solution exists for all treaty sizes.

strategies in a way that deems free-riding inherently ineffectual. At the same time, it minimizes the losses of the countries in the remaining coalition that are trying to deter free-riding.

Another advantage of the farsighted rule over the standard approaches and the γ-core concept is its simplicity and applicability to all coalition game settings, regardless of whether agents are homogeneous or heterogeneous.

Finally, we would like to emphasize that our departure from the standard literature by introducing an information-coordination pre-agreement stage and modifying the timing of the game has critical advantages and is a departure toward making coalition models closer to reality. However, we provide two alternative settings that can be used to find the farsighted emissions and yet reinstate the standard timing.

A Proof of Proposition (4.1)

Since the premise of this article is that the grand coalition is unstable under the Stackelberg solution, that means $w^n_{k-1,St} > w^m_{k,St} = w^m_k$ ($w^n_{k-1,C} > w^m_{k,C} = w^m_k$). So, to curb the free-riding incentives the agency should choose e^m_{k-1} to lower the welfare of the free-rider, which also means we readily have $w^n_{k-1} \leq w^n_{k-1,St}$ ($w^n_{k-1} \leq w^n_{k-1,C}$). Moreover, by the negative externality assumption, we have $\frac{\partial w^n_s}{\partial e^m_s} < 0$, therefore, $e^m_{k-1} \geq e^m_{k-1,St}$ (and given the fact that $e^m_{k-1,St} \geq e^m_{k-1,C}$, then $e^m_{k-1} \geq e^m_{k-1,C}$). In addition, by the nature of our assumptions, the members' and non-members' emissions are strategically substitute, therefore, $e^n_{k-1} \leq e^n_{k-1,St}$ (and $e^n_{k-1} \leq e^n_{k-1,C}$). Consequently, $w^m_{k-1} \leq w^m_{k-1,St}$. The welfare comparison for the farsighted and Cournot members depends on how much the increase in members' emissions compensates for the decrease in non-members' emissions.

Mathematically, for the farsighted rule and a treaty of size $k-1$ we solve the following Lagrangian:

$$\mathscr{L} = (k-1)\left\{B\left(e^m_{k-1}\right) - D((k-1)e^m_{k-1} + e^n_{k-1})\right\} + \lambda(w^m_k - w^n_{k-1}), \tag{10}$$

with $B' - (k-1)D'[1+g'] - \lambda\frac{\partial w^n_{k-1}}{\partial e^m_{k-1}} = 0$ as the first order condition, where the Lagrange multiplier λ is strictly positive given the binding constraint. Therefore, at the solution: (i) automatically, we must have $w^n_{k-1} \leq w^n_{k-1,St}$ and $w^m_{k-1} \leq w^m_{k-1,St}$ (adding a constraint to an optimization problem cannot be welfare improving); and (ii) the treaty's net marginal benefit of a member's emission must be negative at the constraint optimum, i.e., $B' - (k-1)D'[1+g'] < 0$. The latter condition, paired with the premise of eliminating the free-riding incentives in a negative externality context, i.e., $\frac{\partial w^n_s}{\partial e^m_s} \leq 0$, yields in $e^m_{k-1} \geq e^m_{k-1,St}$, and since $e^m_{k-1,St} \geq e^m_{k-1,C}$ (for a formal proof see Finus et al. 2021), then $e^m_{k-1} \geq e^m_{k-1,C}$, then we also readily have $e^n_{k-1} \leq e^n_{k-1,St}$, and $e^n_{k-1} \leq e^n_{k-1,C}$.

A similar argument is applied to other coalition sizes.

References

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

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

Chander, P. & Tulkens, H. (2006). *Cooperation, stability and self-enforcement in international environmental agreements: a conceptual discussion.* Working Paper 2006/03.
Chander, P., & Tulkens, H. (1995). A core-theoretic solution for the design of cooperative agreements on transfrontier pollution. *International Tax and Public Finance, 2*(2), 279–294.
Chander, P., & Tulkens, H. (1997). The core of an economy with multilateral environmental externalities. *International Journal of Game Theory, 26,* 379–401.
d'Aspremont, C., Jacquemin, A., Gabszewicz, J. J., & Weymark, J. A. (1983). On the stability of collusive price leadership. *The Canadian Journal of Economics/Revue canadienne d'Economique,16*(1), 17–25.
Eyckmans, J., & Finus, M. (2006). Coalition formation in a global warming game: how the design of protocols affects the success of environmental treaty-making. *Natural Resource Modeling, 19*(3), 323–358.
Finus, M. (2003). Stability and design of international environmental agreements: the case of transboundary pollution. In H. Folmer & T. Tietenberg (Eds.), *The international yearbook of environmental and resource economics 2003/2004: a survey of current issues* (pp. 82–158). Cheltenham: Edward Elgar.
Finus, M. (2008). Game theoretic research on the design of international environmental agreements: Insights, critical remarks, and future challenges. *International Review of Environmental and Resource Economics, 2*(1), 29–67.
Finus, M., Furini, F., & Rohrer, A. V. (2021). International environmental agreements and the paradox of cooperation: revisiting and generalizing some previous results. *Graz Economics Papers 2021-05*.
Hoel, M. (1992). International environment conventions: the case of uniform reductions of emissions. *Environmental and Resource Economics, 2*(2), 141–159.
Hoel, M., & Schneider, K. (1997). Incentives to participate in an international environmental agreement. *Environmental and Resource Economics, 9*(2), 153–170.
Long, N. V. (1992). Pollution control: a differential game approach. *Annals of Operations Research, 37,* 283–296.
Marrouch, W., & Chaudhuri, A. (2016). International environmental agreements: doomed to fail or destined to succeed? A review of the literature. *Environmental and Resource Economics, 9*(3–4), 245–319.
Ray, D., & Vohra, R. (1997). Equilibrium binding agreements. *Journal of Economic Theory, 73,* 30–78.