Schmidt, Robert; Kovac, Eugen

Conference Paper

A simple dynamic climate cooperation model

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: International Climate Policy, No. D15-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Schmidt, Robert; Kovac, Eugen (2016) : A simple dynamic climate cooperation model, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: International Climate Policy, No. D15-V3, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/145481

A simple dynamic climate cooperation model

Robert C. Schmidt* Eugen Kováč†

November 20, 2015

Abstract

A standard result from the game theoretic literature on international environmental agreements is that coalitions are either 'broad but shallow' or 'narrow but deep'. Hence, the stable coalition size is small when the potential welfare gains are large. We modify a standard climate coalition game by adding a – seemingly – small but realistic feature: we allow countries to delay climate negotiations until the next 'round' if a coalition forms but decides to remain inactive. It turns out that results are surprisingly different under this modification. In particular, a large coalition with deep emissions cuts forms if countries are sufficiently patient. Our results also indicate that countries should try hard to overcome coordination problems in the formation of a coalition. A more cooperative outcome may then be reached, and it may be reached more quickly.

*School of Business and Economics, Humboldt University, Spandauer Str. 1, 10178 Berlin, Germany; E-mail: robert.schmidt.1@wiwi.hu-berlin.de; URL: http://u.hu-berlin.de/schmidt.

†Mercator School of Management, University of Duisburg-Essen, Lotharstr. 65, 47057 Duisburg, Germany; E-mail: eugen.kovac@uni-due.de; URL: https://sites.google.com/site/eugenkovac.

Keywords: climate treaty, coalition, dynamic game, coordination, delay

JEL classification: D62, F53, H23, Q54

1 Introduction

We use a "standard" climate cooperation model, and introduce a – seemingly – minor extension, that is plausible and captures an important aspect of the problem. Surprisingly, this small adjustment in the model leads to markedly different results. In particular, the central conclusion of a significant part of the literature on international environmental agreements (IEAs) turns out *not* to be robust under this modification. In contrast to previous findings from the literature, large coalitions that achieve significant welfare gains for their members *are* stable for reasonable parameter values.

The extension of the model is simple. Instead of modeling climate negotiations as a one-shot game, we assume that if no satisfactory agreement (from the perspective of the members of the coalition) is reached in a period, then the coalition can decide to remain inactive, and new negotiations take place in the next period. Of course there is a welfare-cost of such delay, but delaying has the advantage that a better outcome may be achieved in the next round of the negotiations. In particular, a coalition only decides to become active if it achieves sufficiently high welfare gains, i.e., if it is large. Otherwise, the coalition remains inactive, in expectation of a better outcome in the next round.

Apart from this extension, we maintain the following standard assumptions from the literature:

- only one IEA can be negotiated (that lasts for the entire time horizon of the model once it becomes active)
- countries are ex-ante symmetric
- the identity of the countries that become members of the coalition is randomly determined (in each round of the negotiations)[1]
- countries can coordinate at the participation stage; hence, if the stable coalition size is k^* then exactly k^* countries join the coalition.

The "standard" climate cooperation model (e.g., Barrett 1994) produces the following predictions, which were seen as fairly robust by most authors until today, in the sense that they are obtained unless major changes in the setup or in the equilibrium concept are introduced:[2]

[1] The central result of this paper, i.e., that large coalitions that achieve significant welfare gains for their members are stable, can be obtained also when this assumption is dropped. See the main text, in particular Section 3.2 for details.

[2] Among the changes that have been considered in the literature are the introduction of penalties (in particular in the form of trade sanctions) or farsightedness in the context of climate contracts. See the related literature part (below) for further details.

1. the stable coalition is either "broad but shallow" or "narrow but deep"; hence, large coalitions are stable precisely when they can only achieve modest welfare gains for their members

2. the stronger the free-rider incentive (as measured e.g. by the ratio of welfare of a non-signatory over that of a signatory), the smaller is the stable coalition size.

Although our model differs only in one aspect from the "standard" model – an aspect that we believe makes the model more realistic – it leads to surprisingly different results. In particular, both of the above predictions are reversed (under mild restrictions on the parameters of the model). We find that:

1'. the stable coalition can be large *and* achieve significant welfare gains for its members

2'. the stronger the free-rider incentives, the *larger* is often the stable coalition size.

These surprising results can be explained intuitively by identifying two central effects. First, with the possibility to reach an agreement also in the next period, the coalition members in period t become more demanding: the coalition has to achieve a lot (i.e., it has to be sufficiently large) in order for its members to be willing to sign an agreement. Otherwise (i.e., if the coalition is too small), not much is "sacrificed" by postponing the negotiations, as a better outcome may be achieved in the next round. Furthermore, each coalition member from period t then has a chance to become a non-signatory in the next period, which is beneficial for this country if an agreement is negotiated successfully in that period (free-rider incentive). And second, although it is more tempting to become a non-signatory when the coalition size is large (because abatement per signatory is then high), this does not destabilize the agreement in period t. The reason is as follows. Since the (expected) coalition size in the next period is then also large, the *probability* that a country that drops out today (which induces a delay in the climate negotiations) becomes a non-signatory in the next period is small. This effect undermines the free-rider incentive, and allows for the formation of a large stable coalition even when the welfare of a non-signatory of the agreement is significantly higher than that of a signatory.

Although the model we present in this paper is a dynamic one, our results are not based on countries' usage of punishment (so-called "grim-trigger") strategies that are well-known from repeated games. In our model, we assume that if a long-term coalition forms at some stage, then signatories' abatement targets are fixed for the remaining time horizon of the model. Hence, the game then effectively ends. As long as no such agreement has been reached, the game always has the same payoff-structure in each period. Hence, we focus on stationary (Markov) strategies that are not conditioned on countries' actions in previous periods as long as no agreement has been signed yet. It is remarkable that

large coalitions are nevertheless sustainable in the model. This is due to an *endogenous* threshold effect regarding the minimum size of an active coalition.

In an extension, we also consider the possibility of a coordination failure at the participation stage. We assume that if countries try to coordinate on a given coalition size, then with a certain probability, only a smaller coalition forms. Due to the threshold effect, this coalition, then, decides to remain inactive so that a delay arises although countries play their equilibrium strategies. We demonstrate that with a higher probability of a coordination failure, the stable coalition size becomes smaller. This is due to a reduced continuation value, which implies that countries are more eager to sign an agreement in a given period. Eagerness reduces the endogenous threshold for the stable coalition size, which leads to lower participation in equilibrium. Our results, thus, indicate that a lower probability of coordination failure is beneficial in two ways: the equilibrium outcome then becomes more cooperative, and furthermore, cooperation may be achieved more quickly. Countries should, thus, structure the negotiation process in a simple and transparent way in order to reduce the chances of a coordination failure.

In a second extension, we allow countries to sign also a short-term climate agreement in a period where the negotiations about a long-term agreement have failed. In the base case that we consider, we assume that all countries negotiate about a short-term agreement when it is common knowledge that negotiations about a long-term agreement have failed in that period. We show that the possibility to sign a short-term agreement has a stabilizing effect upon long-term cooperation. This is because deviations at the participation stage are now less costly, since welfare in a period where no long-term agreement is implemented, is increased. Hence, countries are less eager to reach a long-term agreement, so that the stable coalition size is higher.

We also consider alternative ways of modeling negotiations about short-term agreements, and demonstrate that the results in the overall model are sensitive to the exact way in which this is done. In some cases, the incentives of countries to free-ride on a short-term agreement in a period where negotiations about a long-term agreement have failed, can drive a wedge between the payoff of a deviator, and the incentives of a coalition as a whole whether or not to sign a long-term agreement. This can destabilize the equilibrium with high participation in a long-term agreement. Hence, great care should be taken to model negotiations about short-term and long-term agreements in a realistic way, as the overall outcome of a game can be quite sensitive to the details of the modeling setup.

Related literature

Our model is closely related to Battaglini and Harstad (2015). The main differences are twofold. On the one hand, we do not include countries' investments in R&D in our model.

This is the main simplification. On the other hand, we depart from the assumption that – if a coalition forms in a period – it can endogenously determine the length of the commitment period, without a possibility for individual coalition members to drop out in case no long-term agreement is signed. Intuitively, in the model of Battaglini and Harstad (2015) it is the incentive to free-ride on a *short-term* climate agreement that destabilizes the new equilibrium type we identify in our model. However, the assumption that countries must stay in a coalition that was formed in a period with the goal to sign a long-term agreement, even when no such agreement is signed, seems restrictive. It violates the idea that participation in a coalition is voluntary. Hence, in our extension where we allow for the possibility of countries to sign a short-term agreement when no long-term agreement is signed, we assume that *new* negotiations start about a short-term agreement. This is important because the stable coalition size in such an agreement is typically much smaller than for a long-term agreement. This explains why we can identify a new equilibrium type, that Battaglini and Harstad (2015) could not find.[3] The possibility to negotiate a short-term agreement even has a stabilizing effect upon long-term cooperation in our model.

In comparison with static approaches, our model is closest to Barrett (1994), Carraro and Siniscalco (1993), and various papers that followed in this strand of literature.[4] Similar to Karp and Simon (2013), we also adopt a non-parametric modeling approach, that does not rely on specific functional forms.

A paper that inspired our approach is Hong and Karp (2012). These authors consider the possibility that countries randomize over their participation decisions. Hence, they consider mixed-strategy equilibria at the participation stage. This allows for the possibility of a coordination failure. Namely, in Hong and Karp (2012), coalitions that fall short of a critical minimum size do not implement positive abatement efforts. This effect is related with the assumption of binary abatement decisions in their model (respectively, abatement decisions at the boundary of a continuous interval, given linear benefits and costs of abatement). In our model, abatement decisions are continuous. The possibility that a coalition remains inactive (does not sign an agreement and, hence, does not implement additional abatement efforts as compared to the non-cooperative benchmark) is related to the *dynamic* structure of our model and, hence, very different in nature. Nevertheless, Hong and Karp (2012) is one of the first papers in this strand of literature that allows for the possibility of a coordination failure (in equilibrium). In the main part of our paper, such coordination failure does not arise, because we assume that countries *can* coordinate their participation decisions. However, in our second extension we conceptualize a different explanation for a possible coordination failure. It is related to the one in Hong and Karp (2012), as countries play a waiting game and enter the coalition

[3]For a related paper see also Harstad (2014).

[4]For an overview, see Finus (2008).

with a certain (endogenous) hazard rate.

Other climate coalition formation games that are also able to generate larger stable coalition sizes often depart more fundamentally from the basic setup introduced in Barrett (1994). E.g., Helm and Schmidt (2015) assume that countries that unilaterally implement higher carbon prices than others can protect the competitiveness of their domestic industries via border carbon adjustment (BCA) measures. Barrett (1997) considers the possibility of trade sanctions to foster participation in a climate agreement. Finus and Maus (2008) assume that signatories do not fully internalize the environmental externalities between them, by implementing a lower carbon price. This way, the free-rider incentive is reduced, which allows for larger stable coalition sizes. Hoel and Schneider (1997) assume there is a social cost of non-cooperation. Barrett (2006) and Hoel and de Zeeuw (2010) consider the possibility of a technological breakthrough in low-carbon technologies and how it affects the incentives to participate in a climate treaty.

2 Model

There are N ex-ante symmetric countries that negotiate about an international environmental agreement (IEA). The negotiations start in period 0, and as long a no agreement has been signed in the previous period, a new round of negotiations starts in each period $t = 1, 2, ...$ (the time horizon is infinite). If an agreement is reached in period t, an IEA is implemented from that period onwards that covers the remaining time horizon of the model. Hence, the game then (effectively) ends.

Let k_t be the number of countries that become members of the coalition in period t. We restrict our attention to the case where only one coalition is formed. If the coalition signs a long-term agreement, the abatement targets of the signatories are chosen such that their aggregated welfare is maximized, whereas each of the remaining countries ("non-signatories") chooses its emissions individually in this and all future periods so as to maximize its welfare. However, the k_t members of the coalition in period t can collectively decide not to sign an agreement. In this case, the coalition dissolves, and all N countries choose their emissions non-cooperatively in that period. A new round of negotiations then starts in the next period. Figure 1 illustrates the timing of actions in period t (the details of the negotiation process are specified further below).

Let $\pi_0 \geq 0$ denote the per-period payoff for a country in a period where no agreement has been reached yet. We assume this payoff is the same for all countries and constant over time (i.e., independent of the index of the period, t). Let $\pi_s(k)$ denote the per-period payoff of a signatory of a long-term agreement with k members and $\pi_n(k)$ denote the per-period payoff that a non-signatory obtains in this case. Although it is sufficient to consider $\pi_s(k)$ and $\pi_n(k)$ only for $k \in \{1, 2, \ldots, N\}$, it will turn out to be convenient

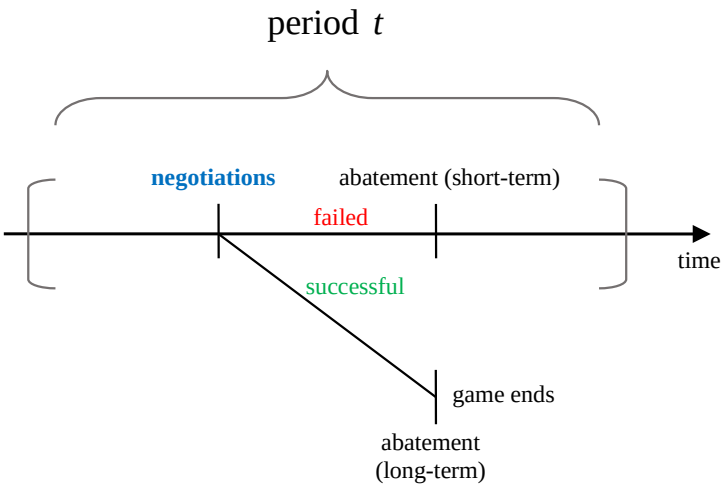

Figure 1: Timing of actions in period t

to define them over the whole interval $[0, N]$. We assume that these payoffs are constant over time. Then

$$\Pi_s(k) = \frac{\pi_s(k)}{1-\delta} \qquad \text{and} \qquad \Pi_n(k) = \frac{\pi_n(k)}{1-\delta} \tag{1}$$

are the discounted payoffs of a signatory and of a non-signatory if an agreement with k members has been signed, where $\delta \in (0, 1)$ is the discount factor.[5] We assume that the functions $\pi_s(\cdot)$ and $\pi_n(\cdot)$ are continuously differentiable and strictly increasing.[6] Moreover, we assume that

$$\pi_n(k_0) = \pi_s(k_0) \text{ , and } \pi_n(k) > \pi_s(k) \text{ if } k > k_0 \tag{2}$$

for some $k_0 \in \{0, 1\}$. Intuitively, k_0 is the coalition size where each coalition member (if any) behaves in the same way as each of the non-signatories. This is generally the case when $k = 1$, hence, $k_0 = 1$. However, when we introduce a specific example (see Section 3), we will sometimes impose the additional assumption that non-signatories do not regulate their emissions. In that case we have $k_0 = 0$, because a singleton coalition then behaves differently than each of the non-signatories. With the above assumptions, we can easily accommodate both cases in our general, non-parametric formalization of the model without having to specify the value of k_0 at this stage. The condition $\pi_n(k) > \pi_s(k)$ if $k > k_0$ is intuitive in the context of climate cooperation, because the non-signatories enjoy the same benefits of abatement as the signatories (with pollution being a global

[5] Throughout the paper, capital letters usually indicate discounted (long-term) values, whereas small type letters reflect values in a single period.

[6] RS: I would delete this footnote!!! It is easy to see that an increasing function defined over $\{1, 2, \ldots, N\}$ can be extended to the interval $[0, N]$ so that it is increasing and continuously differentiable. While there are many possibilities for such an extension, the results of the paper do not depend on the particular extension.

public bad), but incur lower costs of abatement.

In addition, we assume that

$$\pi_0 \geq \pi_n(k_0). \tag{3}$$

The most reasonable assumption about the per-period payoff of a country in a period in which no long-term agreement has been signed yet is $\pi_0 = \pi_n(k_0)$. This reflects the case where all countries choose their abatement targets non-cooperatively in such a period. However, we will later extend the model and allow for the possibility that a short-term agreement is signed in a period where the negotiations about a long-term agreement have failed (see Section 4). This case can also be accommodated in our general framework by allowing for the possibility that $\pi_0 > \pi_n(k_0)$. Clearly, as long as such a short-term agreement does not reach full participation, it must also hold that $\pi_0 < \pi_s(N)$, and we assume that this condition is always satisfied.[7]

For later reference, let us also define the following function (see also Karp and Simon, 2013):

$$G(k) = \Pi_n(k) - \Pi_s(k+1), \tag{4}$$

where $k \in [k_0, N-1]$. Intuitively, $G(k)$ measures the incentives to free-ride, because it is the additional payoff incurred by a country that stays outside of a climate coalition (of size k), instead of joining it. We assume that the function $G(k)$ is strictly increasing in k. This means that the free-rider incentives are rising in the coalition size. This is intuitive, as larger coalitions internalize more of the environmental externalities.[8]

Clearly, if $1 < k_t < N$ there is a coordination problem at the participation stage of the negotiations where each of the N countries simultaneously and non-cooperatively decides whether to become a coalition member in that period. In the following we thus discuss how the *identity* of the countries that become members of the coalition in period t is determined (for some given coalition size k_t that will later be determined endogenously). It turns out that this can significantly affect the equilibrium outcome. There are two different approaches that are used in the literature, and there seems to be no general agreement about which of the approaches is more suitable. There are good arguments in favor of both approaches, and our model allows us to use either one of them (see below). Let us briefly sketch the two approaches and summarize some of their merits.

The first approach is to assume that the identity of the coalition members (for some given coalition size k_t) is pre-determined and commonly known. Furthermore, these iden-

[7] The limit case where $\pi_0 = \pi_s(N)$ would imply that full cooperation is reached in any period where no long-term agreement has been implemented yet. In this case, long-term cooperation cannot improve upon the outcome any further, so there is no scope for it.

[8] This restricts the set of possible specifications of the functions Π_s and Π_n. The condition $G'(k) > 0$ is similar to the condition $\Pi_n'(k) > \Pi_s'(k)$ used by Helm and Schmidt (2015) who drop the integer constraint on the number of signatories. If $\Pi_s(\cdot)$ is convex then this latter condition is implied by $G'(k) > 0$ since then $\Pi_s'(k+1) \geq \Pi_s'(k)$.

tities are time-invariant. As an example, a country like Germany may have a reputation for being "cooperative" so that even if the equilibrium coalition size is small, this county would be expected to become a member of the coalition. Conversely, a country like India may have a reputation to be reluctant to accept any binding target for greenhouse gas emissions, and only in a very large coalition other countries would expect this country to join in. In line with such observations, some scholars favor the assumption that there exists some natural "ordering" of countries, so that for any given coalition size k_t, it is always clear which countries (in equilibrium) will be part of the coalition and which countries will be the outsiders.

Although the above approach has its merits, there are also some doubts about its validity from a theoretical perspective. The reason for this is the following. Especially in a setting with ex-ante symmetric countries (such as this one), but also under asymmetries, countries may not have an *incentive* to be predictable. If a country is known to become a coalition member, then the incentives to free-ride in any period are low because all countries then expect this country to become a member in the next period in case no agreement is signed in period t. Hence, the highest welfare that such a country can expect is that of a signatory of an active agreement. A country with a reputation for being (sufficiently) non-cooperative, by contrast, will in equilibrium achieve the welfare of a non-signatory, that is strictly larger than that of a signatory by our earlier assumptions. Therefore, no country has an incentive to build up a reputation for being cooperative in the first place, so we should assume that the identity of the countries that become coalition members in period t is determined only during the negotiations in period t (and not before). In line with this, some researchers assume that the identity of the coalition members is (from an ex-ante perspective) *random*. To fix ideas, one may assume that in this case, some randomization device ('nature') selects an "assignment" of countries, so that when countries try to coordinate on an outcome where k_t countries join the coalition in period t, then each country knows whether it is supposed to become a signatory or a non-signatory. Of course, the actual participation decision of country i can deviate from the assigned role.

Figure 2 illustrates the timing of decisions in the negotiations in period t when the identities of the countries that become members of the coalition in period t are randomly determined. If the approach with a pre-determined role of countries as coalition members and outsiders is used, the randomization stage is simply skipped. With reference to Figure 1, we say that the negotiations in period t are "successful" if the coalition signs a long-term agreement. Otherwise, we say that the negotiations "have failed".

Formally, suppose k^* is the equilibrium coalition size in the overall game (in case the stable coalition size is not unique, we assume that countries try to coordinate on a coalition size of k^* in each period where the integer k^* is in the set of equilibrium coalition

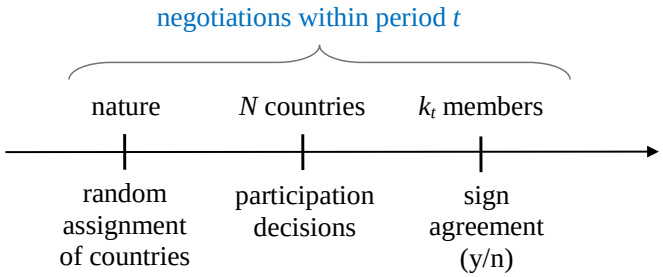

Figure 2: Timing of climate negotiations, with randomly assigned identities

sizes).[9] Then if no agreement is signed in period t, any country that is "assigned" to become a coalition member in period t (but possibly deviated at the participation stage) is assumed to become a coalition member in the following period with a probability of $p_s(k^*)$ in case of a delay. Note, that this probability is by assumption independent of whether this country deviated at the participation stage in the current period (or some other country). Under the approach with a pre-determined identity of (equilibrium) coalition members, we have $p_s(k^*) = 1$, whereas a country which was *not* assigned to become a coalition member in the current period will become a coalition member in the following period with a probability of zero. Under the approach with random identities of coalition members we obtain the specification $p_s(k^*) = k^*/N$. In this case, the probability for a country to become a coalition member in the following period is independent of whether this country was assigned to become a coalition member in the current period or not. We can accommodate both cases (random / non-random assignment of coalition members) with general notation by assuming that the function $p_s(\cdot)$ is continuously differentiable over the interval $[1, N]$, and $p_s'(k) > 0$ whenever $p_s(k) \in (0, 1)$.[10]

Following a deviation in period t that induces a delay in the negotiations, countries expect equilibrium behavior from the following period onwards so that a coalition is expected to form in the next period with probability 1.[11] Hence, when no agreement is signed in period t, a country that was assigned to become a coalition member in period t achieves an expected welfare in the next period of

$$V_s(k^*) = p_s(k^*)\Pi_s(k^*) + (1 - p_s(k^*))\Pi_n(k^*). \tag{5}$$

Because a larger coalition size implies that a larger share of the externalities between

[9] Because the game has the same structure in all periods (as long as no agreement has been signed), we assume that the expected coalition size in the next period is the same whenever a delay occurs, and equal to k^*.

[10] Recall that in the case with a non-random assignment we have $p_s(k) = 1$ for all $k \in [1, N]$. This function also satisfies these criteria.

[11] Hence, we restrict attention to single-period deviations. Due to the time-invariant payoff-structure in the model, this is without loss of generality.

countries are internalized, we require the function $V_s(\cdot)$ to be strictly increasing.[12] If the coalition decides not to sign an agreement in period t, all N countries choose their emissions non-cooperatively in that period and obtain the payoff π_0.

Let us now determine under what condition the coalition in period t signs an agreement. To this end, for $k \in [k_0, N]$ we define a critical coalition size $f(k)$ that satisfies the following condition:

$$\Pi_s(f(k)) = \pi_0 + \delta V_s(k), \tag{6}$$

or equivalently

$$f(k) = \Pi_s^{-1}\left(\pi_0 + \delta V_s(k)\right). \tag{7}$$

We show in the proof of Lemma 1 that the function f is well defined.

A (hypothetical) coalition size of $f(k^*)$ thus leaves the members of the coalition indifferent between signing an agreement in period t, and delaying the negotiations until the next period. Of course, $f(k^*)$ will not be an integer in general. Hence, let $\hat{f}(k^*) = \lceil f(k^*) \rceil$ be the smallest integer at least as large as $f(k^*)$.[13]

Lemma 1. *Function f is well defined and strictly increasing. If countries expect the formation of a coalition of size k^* in the following period (with probability 1) if no agreement is signed in period t, then $\hat{f}(k^*)$ is a cutoff-value for the coalition size, such that the coalition signs an agreement in period t if and only if $k_t \geq \hat{f}(k^*)$.*

We can now write down equilibrium conditions for k^*. If k^* is a stable coalition size, then it must hold that either

$$\Pi_n(k^*) \geq \Pi_s(k^*+1), \quad \Pi_s(k^*) \geq \pi_0 + \delta V_s(k^*), \quad \text{and} \quad k^* = \hat{f}(k^*), \tag{8}$$

or

$$\Pi_n(k^*) \geq \Pi_s(k^*+1), \quad \Pi_s(k^*) \geq \Pi_n(k^*-1), \quad \text{and} \quad k^* > \hat{f}(k^*). \tag{9}$$

The second condition in (8) is an internal stability condition for the case where (on the equilibrium path) the coalition is just large enough to become active (i.e., willing to sign a long-term climate agreement: $k^* = \hat{f}(k^*)$). The condition requires that the equilibrium payoff of an individual signatory, $\Pi_s(k^*)$, is at least as large as the payoff that this country expects if it unilaterally deviates at the participation stage in period t and does not become a member of the coalition although it was "assigned" to participate. In that case, $k_t = k^* - 1 < \hat{f}(k^*)$, so that (by Lemma 1) the coalition decides not to sign an agreement, and the deviating country obtains an expected (discounted) payoff

[12] Again, this restricts the set of possible specifications of the functions Π_s and Π_n. When these are derived from the primitives of a (plausible) emissions game, this assumption is automatically satisfied. An example is given in Section 3.

[13] Given some $x \in \mathbb{R}$, $\lceil x \rceil$ is defined as the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$ or equivalently $x \leq \lceil x \rceil < x + 1$.

of $\pi_0 + \delta V_s(k^*)$. This captures a "threshold effect" regarding the coalition size that arises *endogenously* in this model. In particular, if a coalition forms in a period t that is perceived as "too small" by its members, then this coalition decides to dissolve. Hence, in this type of equilibrium coalitional stability is driven by an endogenous threshold effect regarding the minimum size of an active coalition. The "last" country that enters the coalition assures that the coalition signs an agreement. By contrast, in the "standard" model where countries can negotiate only once, coalitional stability is usually driven by the fact that a country that enters a coalition induces the other signatories to raise their abatement efforts. However, the remaining coalition members would sign an agreement also if a country dropped out of the coalition.

Note, that (6) (together with the definition of $\hat{f}$) implies that the second condition in (8) is automatically satisfied whenever $k^* = \hat{f}(k^*)$ holds, which is the third condition in (8). However, this latter condition relates to the *collective* choice of the coalition members whether or not to sign an agreement, whereas the second condition in (8) concerns the participation decision of an *individual* country. The observation that fulfillment of the second condition in (8) is implied by the condition $k^* = \hat{f}(k^*)$ indicates that (on the equilibrium path) the incentives of an individual coalition member whether or not to stay in the coalition, are aligned with the incentives of the coalition as a whole whether or not to sign an agreement. We will later show that this property is crucial for the existence of an equilibrium that satisfies conditions (8).

Conditions (9) characterize a second equilibrium type that can exist in this model. Note, that the first two conditions in (9) coincide with the conditions of external and internal stability in a "standard" IEA model (where countries can negotiate only once).[14] In our model, however, existence of such an equilibrium requires that, in addition, also the condition $k^* > \hat{f}(k^*)$ (third condition in (9)) is satisfied. This assures that even if an individual coalition member deviates at the participation stage and "drops out of the coalition" (i.e., fails to join), then the remaining $k^* - 1$ members still sign an agreement. The payoff of the deviator is, thus, $\Pi_n(k^* - 1)$ (second condition in (9)). Because the functions Π_s and Π_n are discounted payoffs, existence of an equilibrium of the first or the second type will depend on the size of the discount factor δ. Also note that – if an equilibrium of the second type exists (i.e., an equilibrium that satisfies conditions (9)), then the stable coalition size in this equilibrium coincides with the one in the corresponding static version of the model (i.e., the "standard model"), where countries can negotiate only once about a climate agreement. This follows immediately from the fact that the first two conditions in (9) coincide with the conditions of external and internal stability in the static model.

The case $k^* < \hat{f}(k^*)$ can be neglected, because this would imply that in each period

[14] Using the function G, the external stability condition (first condition in (9)) is simply $G(k) \geq 0$. The internal stability condition (second condition in (9)) is $G(k-1) \leq 0$.

a coalition forms that remains inactive. The outcome would, therefore, be equivalent to the fully non-cooperative benchmark case where no coalition forms in any period.

We will be particularly interested in the situation where the stable coalition size k^* is determined by conditions (8), rather than (9). If k^* is not too small, the external stability condition (first condition in (8)) will in general be satisfied. This always holds if the standard conditions of external and internal stability (first two conditions in (9)) yield a participation level that is smaller than the lowest value of k that satisfies the condition $k = \hat{f}(k)$ and, hence, is a fixed point of the function $\hat{f}$.

Let us analyze the fixed points of $\hat{f}$. If (3) holds with equality, a trivial fixed point is $k = k_0$.[15] This outcome, however, coincides with the fully non-cooperative one. The following proposition claims that there is always a fixed point above k_0.

Proposition 1. *Function $\hat{f}$ has a fixed point in the interval $(k_0, N]$.*

As follows from the above discussion, such a fixed point represents an equilibrium coalition size, if it also satisfies external stability. In general, the equilibrium in the overall game need not be unique. In particular, $\hat{f}$ may have several fixed points if f has several fixed points. However, even if f has only one fixed point, besides $k = k_0$, $\hat{f}$ may have several fixed points (see the discussion further below).

If f has indeed only *one* fixed point above k_0, we can provide an additional characterization of the set of fixed points of $\hat{f}$. Before we come to this, let us first state sufficient conditions under which f can have *at most* one fixed point above k_0. A sufficient (albeit not necessary) condition for this is clearly that f is concave.[16] The assumption that f is concave restricts the set of possible specifications of the underlying functions Π_s and Π_n.

The following result allows us to impose a restriction directly on the functions Π_s and Π_n, rather than on the derived function f.[17] Similarly as the restriction 'f concave', it is a *sufficient* condition for f to have at most one fixed point above k_0, not a necessary one.

Lemma 2. *Assume that $\Pi_n(k)/\Pi_s(k)$ is weakly decreasing on $(k_0, N]$. Then f has at most one fixed point in the interval $(k_0, N]$.*

Although the condition '$\Pi_n(k)/\Pi_s(k)$ weakly decreasing' is not fulfilled for all conceivable emissions games that give rise to the functions Π_s and Π_n, it does not seem to be overly restrictive. Intuitively, the benefits from free-riding, when measured by the *ratio* $\Pi_n(k)/\Pi_s(k)$, must not be rising too sharply[18] with the coalition size. E.g., in the simpler

[15] In general, (3) implies that $\Pi_s(k_0) = V_s(k_0) \leq \pi_0/(1-\delta)$ and thus $\Pi_s(k_0) \leq \pi_0 + \delta V_s(k_0)$, or equivalently $k_0 \leq f(k_0)$. On the one hand, if (3) holds with equality, then all above inequalities become equalities and k_0 is indeed a fixed point of f and, thus, of $\hat{f}$ (since k_0 is an integer). On the other hand, if (3) holds with a strict inequality, all above inequalities are strict and we have $\hat{f}(k_0) = f(k_0) > k_0$.

[16] Concavity of the function f is, e.g., satisfied in our example presented in Section 3.

[17] Note, however, that f is derived directly from the primitives of the model.

[18] Since the condition is sufficient but not necessary, the result still holds when the ratio is increasing, but not too sharply.

version of our example (see Section 3), this ratio is always a constant, thus satisfying also this criterion.

Now we are ready to provide the additional characterization of the set of fixed points of $\hat{f}$, for the case where f has at most one fixed point above k_0. If there is indeed such a fixed point, let us denote it $\underline{k}$; otherwise let $\underline{k} = k_0$. Similarly, since $f(k_0) + 1 > k_0$, there can be at most one fixed point of the function $f(\cdot) + 1$. If there is one, let us denote it $\bar{k}$; otherwise, let $\bar{k} = N$.

Proposition 2. *If f has at most one fixed point above k_0, then any integer $k \in (k_0, N]$ is a fixed point of $\hat{f}$ if and only if $k \in [\underline{k}, \bar{k})$.*

This result is illustrated in Figure 3, that shows the functions f and $\hat{f}$, based on a specification of payoff functions Π_s and Π_n from the example introduced in the following section (the details of which are not relevant for our discussion here). As can be seen from the figure, f has (besides k_0) a single fixed point $\underline{k}$ equal to approximately 5.4. Moreover, function $f(\cdot) + 1$ has a fixed point $\bar{k}$ equal to approximately 8.2. On the other hand, $\hat{f}$ has three fixed points: $k = 6$, $k = 7$, and $k = 8$. These are exactly all integers from the interval $[\underline{k}, \bar{k}) = [5.4,\ 8.7)$, as claimed by Proposition 2.

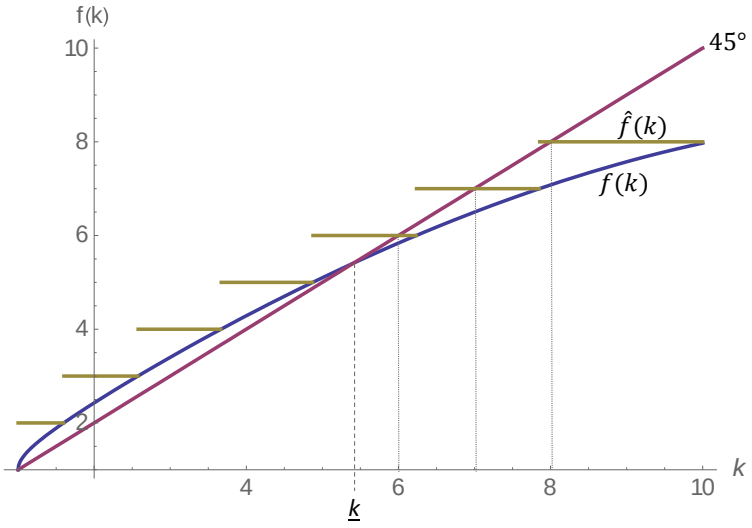

Figure 3: $f(k)$ and $\hat{f}(k)$, for $\delta = 0.6$ and $N = 10$ (example of Section 3 with $a_n > 0$)

Intuitively, if coalition members in period t are optimistic and believe that a larger coalition k^* will form in the following period if no agreement is signed in period t, then they also become more demanding in the current period. The threshold level $\hat{f}(k^*)$ is, then, larger. This is an example of self-fulfilling expectations, because under these circumstances, of course, a larger coalition forms immediately and signs an agreement. If countries are less optimistic and anticipate a smaller coalition size k^* in the future in case

of a delay, then also the critical coalition size $\hat{f}(k^*)$ is smaller and an agreement is signed immediately by fewer countries. Hence, there can be multiple stable coalition sizes.

Now consider an equilibrium of the dynamic game and compare it to an equilibrium in the corresponding static game where countries can negotiate only once about a climate agreement (referred to as the "standard" model). Recall that an equilibrium in the "standard" model is characterized by the first two conditions in (9) (external and internal stability condition), that can be expressed more conveniently with the help of the function G (see (4)). Let us denote $\tilde{k}$ the solution of the equation $G(\tilde{k}) = 0$ if $G(N-1) \geq 0$, and let us set $\tilde{k} = N$ otherwise.[19] Intuitively, in the "standard" model, the stable coalition size is given by the external stability condition $G(k) \geq 0$ and the internal stability condition $G(k-1) \leq 0$. If $\tilde{k}$ is not an integer, since G is assumed to be increasing, the stable coalition size is unique and is equal to the smallest integer at least as large as $\tilde{k}$, denoted $\lceil \tilde{k} \rceil$. In the special case where $\tilde{k}$ is an integer, both $\tilde{k}$ and $\tilde{k}+1$ are stable coalition sizes and the "standard" model then has two equilibria. To avoid some tedious case distinctions resulting from this knife-edge case, let us assume that $\tilde{k}$ is *not* an integer when $G(N-1) \geq 0$. Then in the "standard" model the stable coalition size is indeed unique and is equal to $\lceil \tilde{k} \rceil$.

In the case when $G(N-1) < 0$, the monotonicity of G implies that $G(k) < 0$ for all $k \in [k_0, N-1]$. In this case, countries have always incentives to join the coalition. Thus, the only stable coalition includes all countries, and we set $\tilde{k} = N$. Moreover, in such a case internal stability becomes irrelevant.

Proposition 3. *In any equilibrium of the dynamic game, the coalition is at least as large as in the equilibrium of the static game.*

Clearly, whenever there exists an equilibrium in the dynamic game with a stable coalition size that is strictly larger than in the "standard" (static) model, then this equilibrium is of the first type (i.e., satisfies conditions (8)). Furthermore, any equilibrium in the dynamic game must *either* be of the first *or* of the second type, because the condition $k^* = \hat{f}(k^*)$ (third condition in (8)) cannot be fulfilled simultaneously with the condition $k^* > \hat{f}(k^*)$ (third condition in (9)). In this sense, the two equilibrium types are mutually exclusive. In addition, whenever the condition in Proposition 2 is satisfied, and $\underline{k} > \lceil \tilde{k} \rceil$, then in the dynamic game an equilibrium of the second type fails to exist. In other words, an equilibrium with participation as in the static model cannot co-exist with an equilibrium with higher participation. Equilibrium participation is then *strictly* higher in the dynamic than in the static model.

[19] If $G(N-1) \geq 0$, such a solution indeed exists and is unique. To see this, observe that $G(k_0) = \Pi_n(k_0) - \Pi_s(k_0+1) = \Pi_s(k_0) - \Pi_s(k_0+1) < 0$. Existence follows from continuity of G and uniqueness from monotonicity.

The stable coalition size k^* in an equilibrium that satisfies conditions (8) crucially depends on the discount factor δ, whereas (as is easy to verify) the coalition size that satisfies the standard conditions of external and internal stability (first two conditions in (9)) does not depend on δ. In conjunction with Proposition 3, the next result thus indicates that in this dynamic climate cooperation model, the stable coalition size is (weakly) increasing in the size of the discount factor.

Proposition 4. *Assume that* (3) *holds with equality. Then the value of* $\underline{k}$*, and thus, the smallest stable coalition size that satisfies conditions* (8)*, is increasing in* δ*.*

To see the intuition, consider an equilibrium of the first type, i.e., that satisfies conditions (8). A change in the parameters that makes countries *less* eager to sign an agreement in period t leads to a larger stable coalition size. This is because the outside option (i.e., when coalition members in period t do not sign an agreement) is, then, relatively more profitable, which leads to a larger endogenous threshold for the minimum size of an active coalition. Clearly, when the discount factor is higher, a delay in climate negotiations is relatively less costly because most of the benefits from cooperation are incurred in the future. Hence, countries are less eager to sign an agreement in period t. This explains why the stable coalition size increases in δ.

The best way to sharpen our intuition for this model is to look at a specific example. In the following section, we focus on the simple case with linear benefits and quadratic costs of abatement, that has often been considered in the literature.

3 Example

Suppose, in each period there is a constant marginal benefit of abatement, $b > 0$. A country that abates an amount of $a \geq 0$ of its emissions in a period incurs an abatement cost of $ca^2/2$. For simplicity, let us first assume that non-signatories do not regulate their emissions ($a_n = 0$), hence, $\pi_0 = 0$. Although this assumption is inconsistent with non-signatories' maximizing behavior, it does not qualitatively affect the main results. We will later relax this assumption, and verify that similar results are obtained also in this case. The simplification to assume that $a_n = 0$ makes the algebra particularly simple, and the results very clear and transparent. It does not qualitatively affect any of our main results. Furthermore, it provides the basis for a generalization (see Subsection 3.1) that allows us to derive another interesting result.

With the above assumptions, we find that the aggregated welfare of a coalition of size k in a period is given by $k(bka - ca^2/2)$. The coalition maximizes this over a, so that the abatement per signatory is given by $a_s(k) = bk/c$ in all periods once an agreement is

signed. We thus obtain for the per-period payoffs of signatories, resp. non-signatories:

$$\pi_s(k) = b^2k^2/(2c) \quad \text{, and} \quad \pi_n(k) = b^2k^2/c. \tag{10}$$

Note, that in this example, it holds for any $k > 0$ that $\pi_n(k)/\pi_s(k) = 2$. This nicely illustrates the free-rider incentives in this model.

Let us first focus on the case with a random "assignment" of coalition members and outsiders, i.e., $p_s(k^*) = k^*/N$. (The case with pre-defined roles, i.e., $p_s(k^*) = 1$, is analyzed in Subsection 3.2.) Inserting the expressions from (10) in (**??**), thereby using $\Pi_s(k) = \pi_s(k)/(1-\delta)$ and $\Pi_n(k) = \pi_n(k)/(1-\delta)$, we obtain the following result.

Proposition 5. *Under a random assignment of countries, linear benefits and quadratic costs of abatement, and the additional assumption that non-signatories do not regulate their emissions ($a_n = 0$), we obtain*

$$\underline{k} = \big(2 - \frac{1}{\delta}\big)N \tag{11}$$

as a lower bound for the stable coalition size.

The simple condition (11) nicely captures the central result of this paper. It indicates that if the discount factor δ is close to 1, then the stable coalition size is (almost) equal to N. In this model, the grand coalition can, therefore, be obtained in equilibrium, even when the gains from cooperation are large. Only if the discount factor is close to or below 1/2, the stable coalition size is small. In this case, the coalition size is determined by conditions (9), rather than (8), so that this model then leads to identical results as obtained in a "standard" climate cooperation game. Given the above simplification that $a_n = 0$ (i.e., non-signatories do not regulate their emissions), the stable coalition size is, then, $k^* = 3$, as can easily be verified by using the conditions of internal and external stability in (9).

Let us now relax the assumption that non-signatories do not regulate their emissions to see how this affects the results. Hence, in equilibrium also non-signatories will choose a positive abatement level ($a_n > 0$), and in a period without cooperation countries achieve a positive welfare ($\pi_0 > 0$). It is straight-forward to verify that with our assumption of linear benefits and quadratic costs of abatement, one obtains (see, e.g., Barrett 2005):

$$\pi_s(k) = \frac{b^2}{2c}(k^2 + 2N - 2k) \quad \text{, and} \quad \pi_n(k) = \frac{b^2}{2c}(2k^2 + 2N - 2k - 1). \tag{12}$$

The payoff per country in a period without cooperation is $\pi_0 = b^2(2N-1)/(2c)$. Using this and (12) in the equilibrium condition (**??**), we can determine again the lower bound ($\underline{k}$) for the stable coalition size in an equilibrium that satisfies (8).

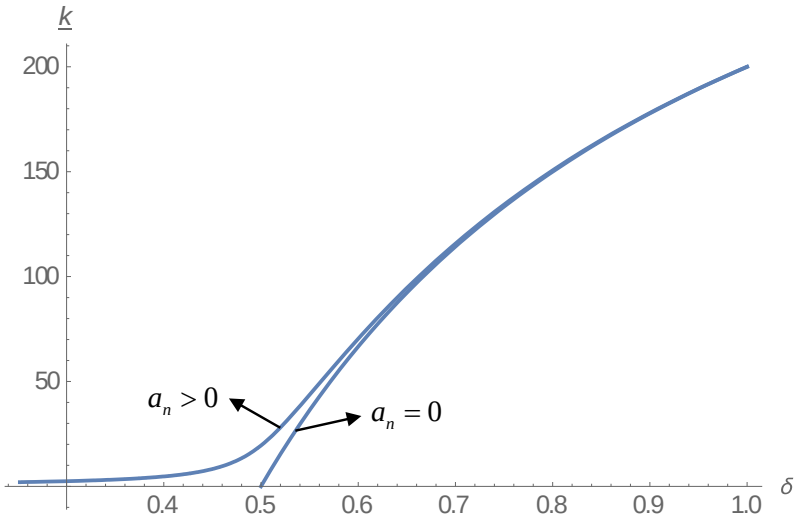

Figure 4: Stable coalition size (lower bound) as function of δ, for $N = 200$

Figure 4 compares the lower bound on the stable coalition size ($\underline{k}$) in the case where $a_n = 0$ with the case where non-signatories also implement climate policies ($a_n > 0$).[20] The figure illustrates that for sufficiently large values of δ, the stable coalition size that satisfies (8) is almost identical in the case where abatement by non-signatories is ruled out by assumption ($a_n = 0$) and in the case where this simplifying assumption is relaxed ($a_n > 0$). Only when δ is close to or below 1/2, the results in the two cases differ significantly. However, these differences also vanish when N is raised further (not shown). The observation that the simple formula for the size of the stable coalition in the case $a_n = 0$, (11), delivers a good approximation for the stable coalition size in the case $a_n > 0$, is not surprising. Consider the following ratio (follows from (12)):

$$\frac{\Pi_n(k)}{\Pi_s(k)} = \frac{2k^2 + 2(N-k) - 1}{k^2 + 2(N-k)}.$$

We observe that it is almost equal to 2 when k is close to N (since $N - k$ is, then, small), and the simple formula for the size of the stable coalition, (11), is obtained precisely when this ratio equals 2 (and when in addition $\pi_0 = 0$, but with k close to N we have $\pi_0 \ll \Pi_s(k)$ so that the impact of π_0 on the stable coalition size is negligible).

[20] If the integer constraint is taken into consideration, then in each case a conservative prediction of the equilibrium coalition size is just the smallest integer greater than the value shown in the figure. However, if N is sufficiently large, the resulting curves that are obtained with the integer constraint are almost indistinguishable from the curves in Figure 4 (not shown).

3.1 Generalized example

Let us generalize the above example. With $\pi_0 = 0$ (hence, when $a_n = 0$), condition (11) is *always* obtained if it holds that $\Pi_n(k)/\Pi_s(k) = 2$. This holds independently of the exact specification of these functions. Hence, consider a more general case where this ratio is fixed, but equal to some arbitrary constant $\alpha > 1$.[21] The parameter α measures the free-rider incentives. If α is close to 1, then it is only slightly more profitable to be a non-signatory rather than a signatory when a long-term agreement is signed. Conversely, if α is large, then being a non-signatory is a lot more profitable than being a member of an active agreement. A fixed ratio $\alpha = \Pi_n(k)/\Pi_s(k)$ is e.g. obtained if the benefit of abatement is linear as in our earlier example, and the abatement cost function is given by $c(a) = ca^z/z$ (with $z > 1$). In this case, we obtain $\alpha = z/(z-1)$. Hence, the free-rider incentives are more intense the closer the parameter z is to 1. The quadratic case that was analyzed above is obtained for $z = 2$.

It is easy to see that in a "standard" climate cooperation model where countries negotiate only once, the stable coalition size becomes very small when α becomes large. In particular, the internal stability condition is $\Pi_s(k^*) \geq \Pi_n(k^* - 1)$. Dividing this condition by $\Pi_n(k^*)$ and using $\Pi_n(k)/\Pi_s(k) = \alpha$, it can be rewritten as follows:

$$1/\alpha - \frac{\Pi_n(k^* - 1)}{\Pi_n(k^*)} \geq 0. \tag{13}$$

Clearly, considering the limit $\alpha \to \infty$, the inequality can only be fulfilled for $k^* \leq 1$ because $\Pi_n(0) = \pi_0 = 0$. Hence, when the free-riding incentives are sufficiently strong, then no cooperation is sustainable in the "standard" model.

The situation is different in the dynamic climate negotiation model that is analyzed in this paper. Following the same steps as before, with $\Pi_n(k)/\Pi_s(k) = \alpha$ our earlier condition (11) now generalizes to

$$\underline{k} = \frac{\alpha - 1/\delta}{\alpha - 1} N. \tag{14}$$

The next result follows immediately from (14).

Proposition 6. *In the generalized example, the lower bound on the stable coalition size ($\underline{k}$) is* increasing *in the size of the free-rider incentive, α.*

This counter-intuitive result reverses our basic intuition and corresponding results from the standard model, where the stable coalition size is declining in the size of the free-rider incentive. Our result can be explained as follows. If the ratio $\alpha = \Pi_n(k)/\Pi_s(k)$

[21] The results that follow remain approximately valid if the ratio $\Pi_n(k)/\Pi_s(k)$ is only (roughly) constant in the relevant range of values for k, and π_0 is sufficiently small.

is large, then it is very profitable to be a non-signatory when an agreement is signed. Hence, in order to be willing to sign an agreement, the signatories of this agreement must be "compensated" for the forgone opportunity to become a non-signatory in the next period in case the negotiations are delayed. Therefore, in order to be willing to sign an agreement, a coalition has to be large. But since the expected coalition size is then large also in the next period (as follows from the stationarity in this model), a deviation by a country that drops out of the coalition in period t (and, hence, causes a delay) is not very profitable. This is because with k^* large, the *probability* to become a non-signatory in the next stage, $1 - p_s(k^*) = (N - k^*)/N$, is small. This undermines the free-rider incentive and explains why a large coalition can indeed form in equilibrium.

3.2 Non-random assignment of coalition members

Formally, the case where countries' roles as coalition members and outsiders are predetermined differs from the case with a random assignment of these roles only in the assumption $p_s(k^*) = 1$ (instead of $p_s(k^*) = k^*/N$). Otherwise, the analysis stays the same. With $V_s(k) = \Pi_s(k)$, condition (6) now simplifies to

$$\Pi_s(f(k)) = \pi_0 + \delta\Pi_s(k). \tag{15}$$

Considering the lower bound ($\underline{k}$) on the stable coalition size satisfying the equilibrium conditions (8), we can apply condition (**??**) to find that

$$\Pi_s(\underline{k}) = \frac{\pi_0}{1-\delta}.$$

Clearly, if π_0 is very small (or zero, as in the simpler version of our example), then also $\underline{k}$ is very small (or zero). This suggests that without the assumption of a random assignment of countries' roles as coalition members and outsiders, the stable coalition size is generally small. It turns out, however, that this conclusion is incorrect.

To see this, consider again the simple version of our example with linear benefits and quadratic costs of abatement, and the additional assumption that non-signatories do not regulate their emissions (i.e., $a_n = \pi_0 = 0$). Inserting the payoff function $\Pi_s(k)$ (using (10)) in (15), we find that

$$f(k) = \sqrt{\delta}k.$$

Hence, if δ is sufficiently large, then the function $f(k)$ is very close to the 45° - line, so that when the integer constraint on the number of signatories is taken into consideration, the resulting function $\hat{f}(k)$ may intersect with the 45° - line at various points, thereby satisfying the equilibrium condition $k^* = \hat{f}(k^*)$ (third condition in (8)).

Figure 5 shows the functions $f(k)$ and $\hat{f}(k)$ for our simple example with $a_n = 0$, for

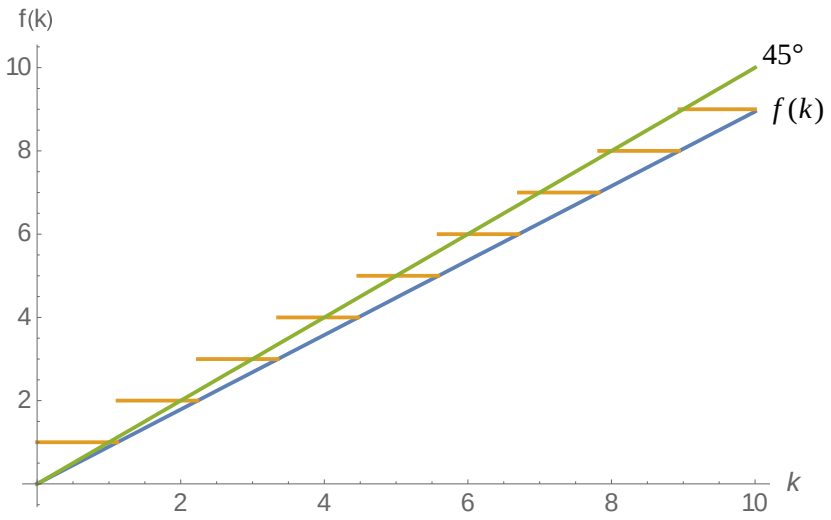

Figure 5: $f(k)$ and $\hat{f}(k)$, for $\delta = 0.8$ and $N = 10$ (example with $a_n = 0$)

$\delta = 0.8$ and $N = 10$ countries. For the given parameter values, we find that *all* integer values between 3 and 10 are stable coalition sizes.[22] Indeed, using the above expression for $f(k)$ in (**??**), we find the following expression for the upper bound on the stable coalition size: $\overline{k} = 1/(1-\sqrt{\delta})$. If δ is close to 1, this expression is large so that large coalition sizes are stable in this model even without a random assignment of countries' roles.[23]

It is easy to verify that this result does not depend on our simplifying assumption $a_n = 0$. Without this assumption, one obtains $f(k) = 1 + \sqrt{\delta}(k-1)$. Again, the function $f(k)$ is linear in this case, and very close to the 45° - line if δ is sufficiently large.[24]

4 Short-term vs. long-term agreements

So far we have assumed that if the coalition in period t decides not to sign a long-term agreement, then all countries choose their abatement efforts individually and non-cooperatively in that period, and new negotiations about a long-term agreement start in the next period. However, even if no long-term agreement is signed in period t, countries could still reach a short-term agreement in that period. In this section, we will allow

[22] $k = 1$ and $k = 2$ are not stable coalition sizes because external stability would be violated (second condition in (8)).

[23] In line with our results from Section 3.1 (see Proposition 6) for the generalized example, also with a non-random assignment of countries' roles as coalition members and outsiders, the stable coalition size is increasing in the size of the free-rider incentive, as measured by $\alpha = \Pi_n(k)/\Pi_s(k)$. Although (in the example with $a_n = 0$) it always holds that $\underline{k} = 0$, the upper boundary for the stable coalition size is given by $\overline{k} = (1-\delta^{1/\alpha})^{-1}$. This is increasing in α.

[24] As a curiosity, when the following functional form is applied: $\pi_s(k) = \alpha^k$, where $\alpha > 1$ is a parameter, then one obtains $f(k) = \frac{\ln \delta}{\ln \alpha} + k$. In this case, when δ is sufficiently large, then *any* coalition size $k \leq N$ is stable, independently of the size of N, because the function $f(k)$ is parallel to the 45° - line.

for this possibility, and consider two alternative ways to model short-term agreements. In the first case, we maintain our earlier assumption that if no long-term agreement is signed in period t, then the coalition dissolves. In this case, new negotiations among all N countries start about a short-term agreement that lasts only until the end of that period. In the second case, we assume instead that even if no long-term agreement is signed in period t, the coalition nevertheless remains intact, and can sign a short-term agreement. In that case, the length of the commitment period (either one or infinitely many periods) is determined endogenously.[25]

4.1 New negotiations about short-term agreements

The assumption that new negotiations about a short-term agreement take place when the coalition in period t decides not to sign a long-term agreement is justified if it is common knowledge of all countries that the negotiations about a long-term agreement have failed in that period. Then it is irrelevant whether a country was part of that coalition (before it dissolved) or not, as all countries have the same probability of becoming a signatory of a short-term agreement. Another way to justify the assumption that new negotiations about a short-term agreement take place when the negotiations about a long-term agreement have failed in period t is by assuming that before countries negotiate about a long-term agreement, some preparations are required. If countries only prepare to negotiate a long-term agreement, but these negotiations unexpectedly fail, then it is possible to negotiate about a short-term agreement, but since countries were not prepared for this, new negotiations have to be organized.

The timing of actions in the overall game is now as follows. Consider again Figure 1. Now the stage "abatement (short-term)" is replaced by a new negotiation stage about a short-term agreement, that has the same structure as shown in Figure 2 (we denote the stable coalition size in a short-term agreement by k_s^*).[26]

Formally, let π_0 be the expected payoff per country in a period where a short-term agreement is implemented. Due to the random assignment of countries' roles as coalition members and outsiders in a short-term agreement, it is given by

$$\pi_0 = \frac{k_s^*}{N}\pi_s(k_s^*) + \frac{N - k_s^*}{N}\pi_n(k_s^*).$$

[25] If it is profitable for the coalition to sign an agreement that lasts for more than one period, then the endogenous length of the commitment period is infinite because of the time-invariant payoff structure in our model. Hence, it suffices to distinguish only between short-term agreements that last for one period, and long-term agreements that cover the remaining time horizon of the model. See Battaglini and Harstad (2015).

[26] The stable coalition size in a short-term agreement coincides with the one in a standard (static) climate coalition formation game. If this coalition size is not unique, then we assume that countries seek to coordinate on a coalition size of k_s^* where this integer value is part of the set of stable coalition sizes in the static game.

By Proposition 3, the stable coalition size in a short-term agreement can never exceed the stable long-term coalition size in the overall game. Hence, our results from Section 2 are qualitatively preserved. However, the above value of π_0 is larger than in the case where no short-term agreements are feasible. This implies that countries are less eager to sign a long-term agreement in a given period, because they achieve a higher welfare in a period without a long-term agreement in place. Since eagerness to reach an agreement reduces the stable coalition size, the possibility to sign a short-term agreement has a *stabilizing* effect upon the equilibrium (long-term) coalition size in the overall game. This is summarized by the following proposition.

Proposition 7. *When new negotiations about a short-term agreement take place in case no long-term agreement is reached in a period, then the possibility to sign a short-term agreement has a positive effect upon the stable long-term coalition size in the overall game.*

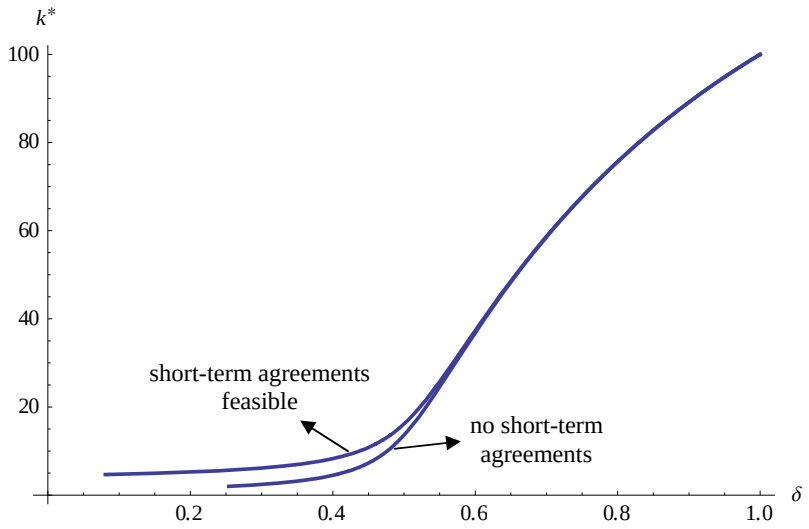

Figure 6: Stable coalition size (lower bound) as function of δ, for $N = 100$ (example of Section 3 with $a_n > 0$)

Figure 6 compares the lower bound ($\underline{k}$) for the stable coalition size in the case where short-term agreements can be implemented with the one in the reference case where this is not possible, using the specification of payoffs from our earlier example (case $a_n > 0$). We observe that for smaller values of δ, the stable coalition size is larger when short-term agreements are feasible, in line with Proposition 7. However, when δ is large, then almost identical results are obtained in the two cases. This is because future payoffs are then relatively more important compared to short-term payoffs (in case of a delay), so that an increase in π_0 has little effect upon the equilibrium outcome.

4.2 Endogenous length of the commitment period

In contrast to the previous subsection, we now assume that the coalition in period t can set the length of the commitment period of the climate contract, without new negotiations taking place if a short-term agreement is signed. For any given coalition size (greater than one), a short-term contract improves the welfare of the signatories relative to the case where no contract is signed at all. Hence, a climate contract will always be signed by a coalition (either short-term or long-term).[27]

Now we have to distinguish between the payoff of a country that is assigned to become a member of a long-term coalition but deviates at the participation stage, and the expected payoff of the k^* coalition members when no country deviated but the coalition as a whole decides not to sign a long-term contract. A country that unilaterally deviates at the participation stage is sure to become a non-signatory if a climate contract is signed (short-term or long-term). Hence, the first condition in (8) becomes:

$$\Pi_s(k^*) \geq \pi_n(k^* - 1) + \delta V_s(k^*), \tag{16}$$

because if an individual country drops out of the negotiations, then $k^* - 1$ countries sign a short-term agreement, and the country that dropped out enjoys the benefits of free-riding in that period. The third condition in (8) implies that

$$\Pi_s(k^*) \approx \pi_s(k^*) + \delta V_s(k^*).$$

Since k^* is larger than the stable coalition size in a short-term agreement (k_s^*), the internal stability condition from the "standard" model: $\pi_n(k^* - 1) \leq \pi_s(k^*)$ is usually violated for k^*. But then the new condition (16) will generally be violated, too. Therefore, in this version of the model, in the dynamic climate cooperation game an equilibrium with a participation greater than the level reached in the "standard" climate coalition model (e.g., $k_s^* = 3$ for our earlier example) generally fails to exist.

However, the assumption that countries cannot drop out of the coalition in period t when this coalition decides not to sign a long-term climate contract is not very compelling, because participation in IEAs is voluntary (e.g., Barrett 2005). Implicitly, this requires that once countries have decided to join a coalition, they are "locked-in", and cannot drop out anymore even when no contract has been signed yet. If the coalition decides to negotiate a short-term (rather than a long-term) contract, however, countries clearly have an incentive to "leave the room" (i.e., to drop out of the negotiations) because the stable coalition size in a short-term agreement is smaller than in a long-term agreement. Hence,

[27] If the identity of the countries that (in equilibrium) join a long-term climate agreement is predetermined, then the setup is essentially the same as in Battaglini and Harstad (2015), with the only difference that we do not allow for technology investments.

this version of the model may be less plausible than the one analyzed in the previous subsection.

Comparing these two variants, there is a general insight that we can gain about this model. Namely, for an equilibrium with a high participation to exist (i.e., an equilibrium of the first type, that satisfies conditions (8) or variants of these conditions), it is crucial that the incentives of the coalition as a whole whether or not to sign a long-term agreement in period t, are aligned with the incentives of each individual coalition member whether or not to stay in the coalition. In the variant of the model that was analyzed in this subsection, these incentives are not aligned. Therefore, it is the incentive of each coalition member to free-ride on a short-term agreement that drives a wedge between the individual incentives to participate in the coalition, and the incentives of the coalition as a whole. This destabilizes the equilibrium with high participation. By contrast, in the variant of the model that we introduced in the previous subsection, or when no short-term climate agreements can be implemented at all, these incentives are perfectly aligned. Intuitively, if a country drops out of the coalition (or, more precisely, does not join the coalition in the first place although it was "assigned" to become a member), thereby inducing a delay in the negotiations about a long-term agreement, then this country expects the same payoff as in the case where the coalition as a whole decides not to sign a long-term agreement. Whenever this holds, a stable coalition with high participation may exist.

5 Coordination failure and delay

The assumption that countries can always coordinate at the participation stage may not be entirely realistic. E.g., during the negotiations, a country would clearly have an incentive to "pretend" to be a non-cooperative player, in the hope that some other country that was previously assumed to become a non-signatory will then revise its participation decision in order to avoid a delay (which arises if the number of coalition members k_t falls short of the threshold $\hat{f}(k^*)$). But since all coalition members have this incentive, one can easily imagine that countries may not always succeed in coordinating.

Let us incorporate the idea of a possible coordination failure into the model. We have earlier assumed that countries overcome the coordination problem with the help of a randomization device ("nature") that assigns to countries their roles as "signatories" and "non-signatories". Now suppose, that this process is not fully reliable, and with some probability assigns to more countries the role of a "non-signatory" than would be consistent with an equilibrium participation of k^*. In order to keep things simple, we assume that if countries try to coordinate on an equilibrium participation of k^*, then with a probability of λ a coordination failure occurs, and only $k^* - 1$ countries receive a signal "assigning" them to become a signatory. Then even if all countries choose their

participation decisions $p_{i,t}$ in line with their signal, then $k_t < k^*$ so that no coalition forms in period t if the equilibrium participation k^* satisfies conditions (8) (see below). With the remaining probability, $1-\lambda$, countries can successfully coordinate their participation decisions, so that if no country deviates unilaterally, then $k_t = k^*$ and an active coalition forms in period t.

In this modified model, the expected discounted equilibrium payoff per country at the beginning of a period t (before countries' roles as "signatories" and "non-signatories" are assigned) is

$$V_s(k^*) = (1-\lambda)\left(\frac{k^*}{N}\Pi_s(k^*) + \frac{N-k^*}{N}\Pi_n(k^*)\right) + \lambda(\pi_0 + \delta V_s(k^*)).$$

Solving for $V_s(k^*)$, this yields:

$$V_s(k^*) = \frac{1}{1-\lambda\delta}\left[(1-\lambda)\left(\frac{k^*}{N}\Pi_s(k^*) + \frac{N-k^*}{N}\Pi_n(k^*)\right) + \lambda\pi_0\right]. \tag{17}$$

With this adjustment, the definition of the function $f(\cdot)$ in (6), and the result of Lemma 1 carry over to this setting. Let us now rewrite the equilibrium conditions in (8) for this more general setup. The first (internal stability) condition now reads:

$$(1-\lambda)\Pi_s(k^*) + \lambda(\pi_0 + \delta V_s(k^*)) \geq \pi_0 + \delta V_s(k^*). \tag{18}$$

The left-hand side of the condition reflects the expected (discounted) payoff of a country that has received a signal assigning this country to the group of signatories, and assuming that the country indeed chooses $p_{i,t} = 1$. Even then, with a probability of λ, a coordination failure will occur in this period, causing a delay of the negotiations until the next period. The right-hand side of the condition is the country's expected payoff if it deviates at the participation stage, and chooses $p_{i,t} = 0$ although it is assigned to become a member of the coalition in that period. In this case, the coalition remains inactive with probability 1 (assuming that all other countries stick with their equilibrium strategies). Rearranging the condition, it becomes (with $\lambda < 1$)

$$\Pi_s(k^*) \geq \pi_0 + \delta V_s(k^*). \tag{19}$$

Hence, it coincides with the first condition in (8), and its fulfillment follows again from the third condition (by rewriting it using (6)), as in the simpler version of the model that was introduced in Section 2. The reason for this is that, again, the incentives of an individual country that is "assigned" to become a signatory whether or not to stay in the coalition are aligned with the incentives of the coalition as a whole whether or not to become active. Intuitively, the occurrence of a coordination failure is perceived like

an exogenous shock, that does not affect the relation between the incentives of coalition members to stay in the coalition in period t and the incentives of the coalition as a whole whether or not to activate an agreement.

Now consider the external stability condition (second condition in (8)). Rewriting it for this setup, it becomes

$$(1-\lambda)\Pi_n(k^*) + \lambda(\pi_0 + \delta V_s(k^*)) \geq (1-\lambda)\Pi_s(k^*+1) + \lambda\Pi_s(k^*). \tag{20}$$

The left-hand side reflects the expected payoff of a country that was assigned to become a non-signatory and indeed chooses $p_{i,t} = 0$. In this case, with a probability of $1-\lambda$ a coalition of size k^* becomes active, so the country obtains the payoff $\Pi_n(k^*)$. With the remaining probability, a coordination failure occurs, so that no coalition forms in that period. On the right-hand side we see the expected payoff of this country if it joins the coalition even though it was assigned to become a non-signatory. This can be beneficial, because in that case, the coalition will become active in this period even when a coordination failure occurs. This is reflected by the term $\lambda\Pi_s(k^*)$. If no coordination failure occurs, then the coalition size becomes k^*+1 and again, the coalition becomes active. Joining the coalition although the signal indicates to become a non-signatory can, thus, be seen as an "insurance" against a coordination failure.

The condition can be violated if λ is large and δ is small. Let us focus, however, on the case where the discount factor is large. This is the case that led to a significant amount of cooperation in the model as introduced in Section 2. Also observe that the right-hand side of the above condition is strictly larger than $\Pi_s(k^*)$ because $\Pi_s(\cdot)$ is an increasing function. However, the right-hand side will in general be only slightly larger than $\Pi_s(k^*)$ so that this replacement will deliver a good approximation of the condition. After replacing the right-hand side by $\Pi_s(k^*)$, using (17) to replace $V_s(k^*)$, and taking the limit $\delta \to 1$, the condition simplifies to:

$$(1-\lambda)\Pi_n(k^*) + \lambda\left(\frac{k^*}{N}\Pi_s(k^*) + \frac{N-k^*}{N}\Pi_n(k^*)\right) \geq \Pi_s(k^*).$$

Clearly, this condition is always fulfilled since $\Pi_n(k) > \Pi_s(k)$ for all $k > 1$. We conclude that if δ is sufficiently large, the external stability condition will continue to be satisfied, so that an equilibrium with (potentially) high participation is again characterized by conditions (7) and (**??**), in conjunction with the modified expression for $V_s(k^*)$ in (17).

Figure 7 compares the resulting stable coalition size ($\underline{k}$) in the extended model (neglecting the integer constraint on the number of signatories) for $\lambda = 0.5$ using the example from Section 3 with $a_n > 0$, and the respective results from Section 3 (that are also obtained as a special case of the above model for $\lambda = 0$). We observe that if a coordination failure occurs with a positive probability, then the stable coalition size becomes smaller

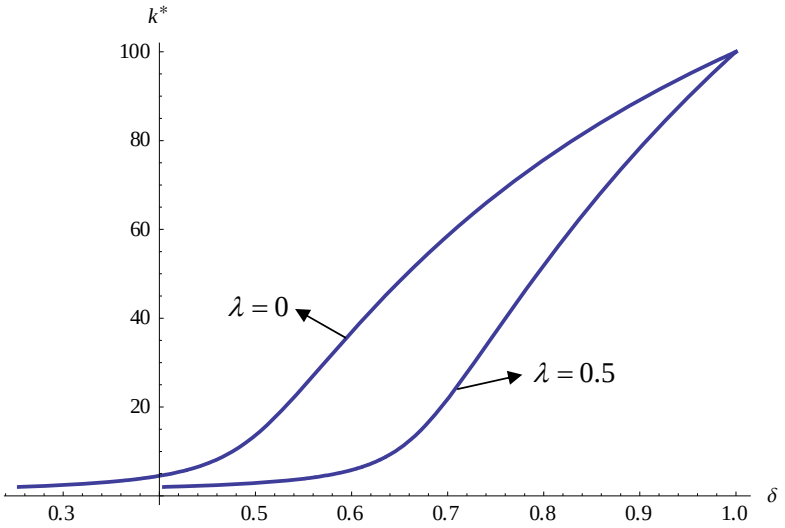

Figure 7: Stable coalition size as function of δ, for $N = 100$

for any given value of δ (and assuming that δ is sufficiently large so that an equilibrium of the first type exists). This result is not very surprising, because by (17), the possibility of a coordination failure has the effect that $V_s(k)$ is reduced (for any given value of k). Hence, a delay becomes more costly, so that countries are more eager to reach an agreement in period t. The endogenous threshold for the minimum amount of participation in order for a coalition to become activate is, thus, reduced.

Introducing the possibility of a coordination failure, thus, adds another realistic aspect to the model. Namely, the model can now capture the idea that a delay occurs in climate negotiations – something that we have often observed in the real world. Furthermore, if the probability of a coordination failure increases, the predictions of the model regarding the size of a stable coalition become more pessimistic. However, there is also a more optimistic insight from this modeling exercise. Namely, even if climate cooperation fails for a certain period of time, this does not imply that it will necessarily fail also in the future. A significant amount of cooperation may still be sustainable, so that the prospects for the future may not be as gloomy as one might think, even when past climate summits (e.g., in Copenhagen, 2009) have failed to produce any binding agreement. Our results, therefore, indicate that countries should try hard to overcome coordination problems. A more cooperative outcome may then be reached, and it may be reached more quickly.

6 Conclusion

Allowing for the possibility that new negotiations about a long-term climate treaty start in the next period, when no agreement is signed in the current period, appears to be a minor departure from the static approach of Barrett (1994) and other authors from

this literature. It turns out, however, that this simple modification drastically changes the strategic incentives of countries in the model. Having the possibility to negotiate also in the future, countries become more demanding in the current period. Coalition members are, then, only willing to sign a long-term agreement if the agreement achieves a lot, i.e., if the coalition size is sufficiently large (from their perspective). This paper, thus, shows, that the pessimistic insights from Barrett (1994) and other authors from this literature are not robust to a simple modification that captures an important aspect of the problem from the real world. Namely, countries can negotiate again in the future when negotiations have failed in the current period. This allows for some optimism regarding the negotiations in Paris this year, even after the failure of the Copenhagen climate summit in 2009.

Our results indicate that large coalitions that achieve significant welfare gains for their members can be stable. Furthermore, our results point towards a transparent design of the negotiation process in order to raise its chances of success. This lowers the risk of a coordination failure, and leads to a larger coalition in equilibrium. Furthermore, it increases the chances that a long-term agreement will be negotiated successfully in the near future.

Appendix A: Proofs

Proof of Lemma 1. First we show that the function f is well defined. For the extreme values $k = k_0$ and $k = N$ we easily obtain[28]

$$V_s(k_0) = \Pi_s(k_0) \leq \frac{\pi_0}{1-\delta} \qquad \text{and} \qquad V_s(N) = \Pi_s(N). \tag{21}$$

As V_s is assumed to be increasing, then it follows from (21) that for all $k \in [k_0, N]$:

$$\Pi_s(k_0) \leq \pi_0 + \delta V_s(k_0) \leq \pi_0 + \delta V_s(k) \leq \pi_0 + \delta V_s(N) < \Pi_s(N). \tag{22}$$

Continuity and monotonicity of Π_s then imply that there is a unique $k' \in [k_0, N)$ such that $\Pi_s(k') = \pi_0 + \delta V_s(k)$. Then we set $f(k) = k'$.

Second, we show that f is strictly increasing. Recall that due to (7), we have $f(k) = \Pi_s^{-1}(\pi_0 + \delta V_s(k))$. By assumption, Π_s and V_s are strictly increasing. Hence, also the inverse function Π_s^{-1} is increasing, which shows that f is increasing.

Now, it follows that $\hat{f}$ is (weakly) increasing. By the definition of the function f, and since k_t is an integer, it follows that: $\Pi_s(k_t) \geq \pi_0 + \delta V_s(k^*)$ if and only if $k_t \geq \hat{f}(k^*)$. The signatories then find it optimal to sign a long-term climate agreement. Otherwise, they prefer to delay the negotiations until the next period. □

Proof of Proposition 1. Recall from the proof of Lemma 1 that

$$f(k) \in [k_0, N) \qquad \text{for all} \qquad k \in [k_0, N]. \tag{23}$$

Let $\eta = \lfloor k_0 \rfloor + 1$ be the smallest integer (strictly) larger than k_0.[29] Below we show that

$$\hat{f}(k) \in [\eta, N] \qquad \text{for all} \qquad k \in [\eta, N]. \tag{24}$$

Since $\hat{f}$ is weakly increasing, we can apply the *Tarski's Fixed Point Theorem*, which implies that $\hat{f}$ has a fixed point in the interval $[\eta, N]$.

Now it remains show (24). It follows from the definition of η that $\eta > k_0$ and we have $\hat{f}(\eta) \geq f(\eta) > f(k_0) \geq k_0$, where the first inequality follows from the definition of $\hat{f}$, the second one from f being strictly increasing (Lemma 1), and the third one from (23). Since $\hat{f}(\eta)$ is an integer and it is larger than k_0, we obtain $\hat{f}(\eta) \geq \eta$. Moreover, $f(N) < N$ due to (23). Because N is an integer, $\hat{f}(N) \leq N$. As $\hat{f}$ is weakly increasing, it indeed maps the interval $[\eta, N]$ into the interval $[\eta, N]$, which completes the proof. □

[28] The former follows from (3). To see the latter, observe that for $k = N$ the coalition includes all N countries and, thus, $p_s(N) = 1$, which then yields $V_s(N) = \Pi_s(N)$.

[29] Given some $x \in \mathbb{R}$, $\lfloor x \rfloor$ is defined as the unique integer such that $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$.

Proof of Lemma 2. Clearly it is sufficient to show that $f(k) \leq k$ implies that $f'(k) < 1$ for $k > k_0$. We consider two cases, that correspond to a random assignment of coalition members and to a non-random assignment, respectively.

Case 1. Let $p(k) \in (0,1)$. Observe that since Π_s is increasing, the inequality $f(k) \leq k$ is equivalent to $\pi_0 + \delta V_s(k) \leq \Pi_s(k)$. Since $\pi_0 \geq 0$, this yields

$$\frac{\delta V_s(k)}{\Pi_s(k)} \leq 1. \tag{25}$$

Moreover, taking the derivative with respect to k in the equation $\Pi_s(f(k)) = \pi_0 + \delta V_s(k)$, we obtain $\Pi_s'(k)f'(k) = \delta V_s'(k)$. The inequality $f'(k) < 1$ is then equivalent to

$$\frac{\delta V_s'(k)}{\Pi_s'(k)} < 1. \tag{26}$$

We now show that (25) implies (26). Recall from the definition of V_s that

$$\frac{\delta V_s(p)}{\Pi_s(k)} = \delta p_s(k) + \delta(1 - p_s(k))\frac{\Pi_n(k)}{\Pi_s(k)}. \tag{27}$$

Moreover, we obtain

$$V_s'(p) = p_s(k)\Pi_s'(k) + (1 - p_s(k))\Pi_n'(k) + p_s'(k)[\Pi_s(k) - \Pi_n(k)], \tag{28}$$

$$\frac{\delta V_s'(p)}{\Pi_s'(k)} = \delta p_s(k) + \delta(1 - p_s(k))\frac{\Pi_n'(k)}{\Pi_s'(k)} + \frac{p_s'(k)}{\Pi_s'(k)}[\Pi_s(k) - \Pi_n(k)]. \tag{29}$$

By assumption, $p_s'(k) > 0$, $\Pi_s'(k) > 0$, and $\Pi_n(k) > \Pi_s(k)$ for $k \in (k_0, N]$. Thus, the last term in (29) is negative.

Now, since $\Pi_n(k)/\Pi_s(k)$ is weakly decreasing, we have

$$\frac{\Pi_n'(k)}{\Pi_s'(k)} \leq \frac{\Pi_n(k)}{\Pi_s(k)}. \tag{30}$$

This follows directly from the derivative, $[\Pi_n'(k)\Pi_s(k) - \Pi_n(k)\Pi_s'(k)]/[\Pi_s(k)]^2 \leq 0$.

Then we obtain from (29), (30), and (27) that

$$\frac{\delta V_s'(p)}{\Pi_s'(k)} < \frac{\delta V_s(k)}{\Pi_s(k)}. \tag{31}$$

This indeed shows that (25) implies (26) and completes the proof for Case 1.

Case 2. Let $p_s(k) = 1$. The proof is almost identical to the previous case. The only difference is that the last term in (29) is now non-positive. Thus, both (30) as well as (31) only hold with weak inequalities. However, now $V_s(k) = \Pi_s(k)$ and thus (25) holds with a strict inequality, which then indeed implies (26) with strict inequality. □

Proof of Proposition 2. Let $k \in (k_0, N]$ be an integer. By definition (see footnote 13), $\hat{f}(k) = \lceil f(k) \rceil = k$ is equivalent to $f(k) \leq k < f(k) + 1$. Due to concavity of f the first inequality is equivalent to $\underline{k} \leq k$, while the second inequality is equivalent to $k < \bar{k}$. Summing up, we obtain that indeed $\hat{f}(k) = k$ if and only if $\underline{k} \leq k < \bar{k}$. □

Proof of Proposition 3. As argued in the main text, the equilibrium coalition size is unique and is equal to $\lceil \tilde{k} \rceil$, the smallest integer at least as large as $\tilde{k}$, where $G(\tilde{k}) = 0$. Now consider an equilibrium of the dynamic game with coalition size k^*. In both cases, (8) and (9), the first condition can be rewritten as $G(k^*) \geq 0$. Since G is assumed to be increasing, this is equivalent to $k^* \geq \tilde{k}$. Finally, as k^* is an integer, then $k^* \geq \lceil \tilde{k} \rceil$. □

Proof of Proposition 4. Because f is monotonically increasing (Lemma 1), it is sufficient to show that $f(k)$ is increasing in δ for any given value of k. To this end, we rewrite $f(k)$ (as given in (7)) by using the functions $\pi_s(k) = (1-\delta)\Pi_s(k)$ and $v_s(k) = (1-\delta)V_s(k)$ and obtain

$$f(k) = \pi_s^{-1}\left((1-\delta)\pi_0 + \delta v_s(k)\right).$$

Note, that $\pi_s(\cdot)$ and $v_s(\cdot)$ do not depend on δ. Since (by assumption) it holds that $v_s(k) > \pi_0$ for any $k > k_0$, it follows immediately that $f(k)$ is increasing in δ. □

Appendix B: Alternative negotiation process

Here we present a micro-foundation for the type of coordination failure analyzed in Section 5. To this end, we introduce an alternative formalization of the negotiation process in period t, that is embedded in the overall game as illustrated in Figure 1. This process determines endogenously whether a climate agreement is successfully negotiated in period t, and if so, which countries are part of the long-term agreement. Hence, the alternative negotiation process replaces all three negotiation stages illustrated in Figure 2, and endogenizes also the probability of a coordination failure λ introduced in Section 5, where it was treated as exogenous.[30] Apart from these changes, the analysis conducted in Section 5 and the results remain valid.

Suppose, the negotiation process within period t has the following structure. The process is in continuous time, and the clock is set to $\tau = 0$ at the beginning of period t where the process starts. The number of coalition members at interim time τ (within period t) is denoted s_τ. Negotiations stop as soon as the critical coalition size k^* is reached

[30]The negotiation process that we present in the following endogenously leads to a random determination of the identities of coalition members in period t, as a result of equilibrium strategies of countries. It cannot replicate a non-random assignment of coalition members as analyzed in Section 3.2.

($s_\tau = k^*$). In this case, a long-term agreement is signed (immediately). We assume that the time span of the negotiation process within period t is short, as compared to the time that elapses between two periods in the overall game. Therefore, there is no discounting of costs or benefits within the negotiation process (the discount rate is zero).

We introduce two types of frictions into the negotiation process. First of all, as long as the critical coalition size has not been reached yet (i.e., $s_\tau < k^*$), the negotiation process is terminated with a constant exogenous hazard rate of stopping: $p^s > 0$.[31] In this case, the negotiations have failed in this period, and new negotiations start in period $t+1$ (see Figure 1). In addition, we assume that there is a negotiating cost. In particular, the representatives of the involved countries who carry out the negotiations incur a flow cost of negotiating, denoted by r, as long as the respective country has not joined the coalition yet.[32]

Let $W_s(s_\tau)$ be the expected discounted payoff of a country that has already joined the coalition, when the current state in the negotiation process (i.e., the coalition size) is s_τ. Similarly, $W_n(s_\tau)$ is the expected discounted payoff of an outsider (at time τ), as evaluated from the perspective of the representative of this country (note that negotiating costs only affect the utility of these individuals, but not the welfare of the country they represent).

Lemma 3. *In equilibrium it must hold that $W_n(k^*-1) = \Pi_s(k^*)$, hence, when $s_\tau = k^* - 1$, each of the remaining non-members is indifferent between joining the coalition and continuing to wait.*

Proof of Lemma 3. We show that 1. $W_n(k^*-1) < \Pi_s(k^*)$, and 2. $W_n(k^*-1) > \Pi_s(k^*)$ *cannot* hold in equilibrium. The result follows from these two observations.

1. Suppose to the contrary that $W_n(k^*-1) < \Pi_s(k^*)$ holds in equilibrium. Then each of the non-members has a strict preference for joining immediately when $s_\tau = k^*-1$ countries are already in the coalition. But then countries would be better off joining already at $\tau = 0$ to avoid becoming one of the non-members when the state s_τ reaches the value k^*-1. Hence, a coalition of size k^* would form immediately at $\tau = 0$. This, however, cannot be an equilibrium because (by $\Pi_n(k) > \Pi_s(k)$ for all $k > 1$) non-members would be strictly better off than each of the signatories, so that each of the k^* countries would benefit from revising its participation decision, anticipating that another country is willing to join the coalition in its place.

[31] The idea behind this assumption is that negotiations can be extended in the real world, but not indefinitely. Hence, the exact time when negotiations terminate (irrespective of whether an agreement has been reached or not) is random.

[32] The flow cost of negotiating can also be interpreted as "distress" or "guilt" experienced by the representatives of countries involved in the negotiations. Representatives of countries that are not yet members of the coalition in period t may be under (social) pressure because they are responsible for a potential coordination failure and, as a result, delay, which is costly for all countries.

2. Suppose that $W_n(k^*-1) > \Pi_s(k^*)$ holds in equilibrium. Then all remaining non-members stay outside of the coalition with certainty, so that a coalition of size k^* never forms in equilibrium (not in this and in no other period). The negotiations, thus, fail in each period t so that no agreement is ever signed. This cannot be an equilibrium because (as $\Pi_s(\cdot)$ is increasing) any country would be better off joining an agreement of an arbitrary size $k > 1$ that signs a long-term contract. □

Before we come to the formal characterization of the negotiation process, we can state the following result which describes a central feature of this process. As a simplification, let us assume that $k^* = f(k^*)$ holds *exactly*, hence, $\hat{f}(k^*) = f(k^*)$. This simplifies the analysis considerably because the integer constraint on the stable coalition size is then, automatically satisfied.[33] In case of a multiplicity of coalition sizes k^* that satisfy the condition $k^* = \hat{f}(k^*)$ (third condition in (8)), this approach always selects the *smallest* equilibrium coalition size k^* (conservative approach).[34]

Proposition 8. *When the negotiation process starts in period t, the state jumps to* $s_{\tau=0} = k^* - 1$ *immediately. The remaining negotiation process is, thus, a waiting game played by the* $N - (k^* - 1)$ *countries that do not join the coalition at* $\tau = 0$.

Proof of Proposition 8. Consider a situation where $k^* - 1$ countries join the coalition immediately. Then if a country is in the group of outsiders its expected payoff is $W_n(k^* - 1) = \Pi_s(k^*)$ (by Lemma 3). If a country is in the coalition then its expected payoff is

$$W_s(k^* - 1) = (1 - \lambda)\Pi_s(k^*) + \lambda(\pi_0 + \delta V_s(k^*)),$$

because this country incurs no negotiating costs even when negotiations continue after $\tau = 0$, so that if the negotiations in period t eventually succeed (which happens with a probability of $1 - \lambda$) then the country obtains a discounted payoff of $\Pi_s(k^*)$. Otherwise, there is a delay and new negotiations start in the next period. Because negotiation costs of future negotiators are not included in the payoff of current negotiators nor in the welfare of the country they represent, the continuation value after a delay is $V_s(k^*)$, as given by (17). But with $k^* = f(k^*)$ and the definition of $f(\cdot)$, (6), we get

$$\Pi_s(k^*) = \pi_0 + \delta V_s(k^*). \tag{32}$$

Hence: $W_s(k^* - 1) = \Pi_s(k^*)$. With Lemma 3, this implies that $W_s(k^* - 1) = W_n(k^* - 1)$. Hence, each country is indifferent towards becoming a coalition member or outsider at

[33] We comment further on this issue at the end of this section.

[34] In the context of our example from Section 3, the multiplicity is usually very modest: The range of stable coalition sizes is generally small as compared to the number of countries N, when N is large (say, $N = 100$). Hence, it makes little difference whether one picks the lowest (say, $k^* = 53$) or the highest (say, $k^* = 55$) of these values as the focal point in the selection of an equilibrium.

$\tau = 0$, a necessary condition for such an equilibrium outcome to exist. For sufficiency, note that under any other outcome where the coalition size $s_\tau = k^* - 1$ is reached later (at some $\tau > 0$), at least one country achieves a strictly lower payoff because it incurs additional negotiating costs, so that it would prefer to join the coalition immediately. □

Equipped with the above results, we can now analyze the waiting game played by the $N - (k^* - 1)$ outsiders formally. As long as the state remains at $s_\tau = k^* - 1$, this game continues and has a stationary structure: irrespective of the time τ that has already elapsed in the negotiations, the probability of each possible event that can happen in the next small time interval (between τ and $\tau + d\tau$) stays the same. In particular, with a hazard rate of p^s, the waiting game is terminated unsuccessfully. Furthermore, focusing on symmetric strategies, denote the probability that an individual non-member joins the coalition in the short time interval by p^j. The probability that any other non-member (except one country of interest) joins the coalition is denoted by $p^+ = (N-(k^*-1)-1)p^j = (N - k^*)p^j$. Finally denote the probability that *some* event takes place that leads to the end of the negotiation process by $p = p^s + p^j + p^+$. Then we obtain for the expected payoff of a non-member, evaluated from the perspective of the representative of this country:

$$W_n(k^* - 1) = p^s(\pi_0 + \delta V_s(k^*)) + p^j \Pi_s(k^*) + p^+ \Pi_n(k^*) + (1 - p)(W_n(k^* - 1) - r).$$

By Lemma 3, we can apply the following equilibrium condition: $W_n(k^* - 1) = \Pi_s(k^*)$. This yields:

$$(p^s + p^+)\Pi_s(k^*) = p^s(\pi_0 + \delta V_s(k^*)) + p^+ \Pi_n(k^*) - (1 - p)r.$$

Using (32), this condition simplifies to

$$\Pi_n(k^*) - \Pi_s(k^*) = \frac{1 - p}{p^+} r.$$

After substituting for p and p^+ this becomes

$$\Pi_n(k^*) - \Pi_s(k^*) = \frac{(1 - p^s)/p^j - (N + 1) + k^*}{N - k^*} r. \tag{33}$$

This implicitly defines a relation $k^* = k^*(p^j)$.

Furthermore, with (17), the condition $\Pi_s(k^*) = \pi_0 + \delta V_s(k^*)$ yields

$$\Pi_s(k^*) = \pi_0 + \frac{\delta}{1 - \lambda\delta} \left[(1 - \lambda) \left(\frac{k^*}{N} \Pi_s(k^*) + \frac{N - k^*}{N} \Pi_n(k^*) \right) + \lambda \pi_0 \right].$$

This implicitly defines a relation $k^* = k^*(\lambda)$. After some rearranging, this condition can

be rewritten more conveniently as follows:

$$\Pi_s(k^*) = \frac{1}{1-\delta}\left[\pi_0 + \delta(1-\lambda)\frac{N-k^*}{N}\left(\Pi_n(k^*) - \Pi_s(k^*)\right)\right]. \tag{34}$$

The condition shows that if λ increases then k^* decreases. In particular, for $\lambda = 1$, the payoff of a signatory is as low as in the case without any cooperation in any period: $\pi_0/(1-\delta)$. Of course, λ is not a parameter but (like k^*) also an endogenous variable.

Finally, in this continuous time process, the probability that the negotiations fail is given by $\lambda = p^s/p$, hence,

$$\lambda = \frac{p^s}{p^s + (N - k^* + 1)p^j}. \tag{35}$$

The system (33), (34), and (35) pins down the three variables k^*, λ, and p^j. Via elimination of λ and p^j it is possible to write down a single equation that determines k^*. This equation is, however, rather lengthy and inconvenient for comparative statics (not shown).

A simpler approach is to combine conditions (33) and (35) by eliminating p^j. This yields a relation $\lambda = \lambda(k^*)$ that can be plotted in a simple λ - k^* - diagram, along with condition (34) that can also be solved for λ in a straight-forward way. Note, that (33) and (35) do not depend on the discount factor δ, whereas condition (34) is independent of the parameters r and p^s that quantify the frictions in the negotiation process.

Figure 8 illustrates the shape of the relation $\lambda(k^*)$, as defined by conditions (33) and (35) (red curve), and by condition (34) (blue curve), for the example of Section 3 (with $a_n = 0$).[35] The equilibrium of the negotiation process is given by the intersection point, that pins down the stable coalition size in the overall model, k^*, along with the endogenous probability that negotiations within a single period fail, λ.

It can be shown that when the stopping rate p^s increases or when the negotiating cost r decreases, then the red curve shifts upwards, while the blue curve is not affected. This leads to a reduction in the equilibrium coalition size k^*, and to an increase in the probability of a coordination failure, λ. Hence, a longer negotiation process (reflected by a lower value of p^s – hence, longer in expectation) is unambiguously good for welfare: the stable coalition size then increases, and (in expectation) a coalition forms more quickly (in an earlier period, since the probability λ that negotiations within a period fail is then smaller). Interestingly, the same results are obtained also when the second friction in the negotiation process, namely the flow cost of negotiating r, *increases*. Intuitively, this cost drives a wedge between the expected utility of negotiators who represent countries that have not joined the coalition yet, and those who have already joined. Raising this cost implies that representatives of the latter are less willing to engage in a waiting game.

[35] The parameter values underlying the figure are $N = 100$, $p^s = 0.001$, $r = 1$, and $\delta = 0.9$.

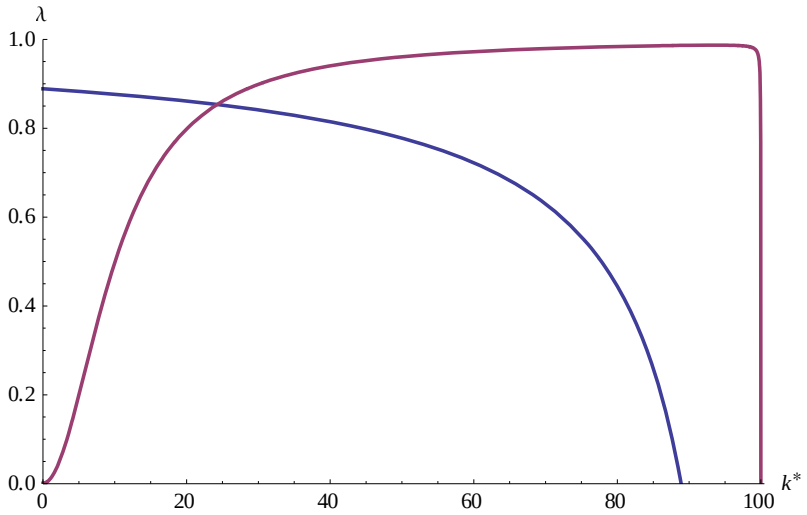

Figure 8: λ as a function of k^*; red: conditions (33) and (35), blue: condition (34)

This lowers the probability of a coordination failure, and raises the stable coalition size in the overall game.

Conversely, a raise in the discount factor δ shifts the blue curve upwards, while the red curve is not affected. This leads to the formation of a larger coalition in equilibrium (k^* increases). However, it also leads to a rise in the probability λ that the negotiations within a period fail. Hence, unlike in the basic version of the model that was analyzed in Section 2 (based on the simpler negotiation process illustrated in Figure 2 that does not lead to a coordination failure in equilibrium), a raise in the discount factor is no longer unambiguously good for welfare. While the stable coalition size rises, it can also lead to more delay in the formation of an active coalition (that signs a long-term climate agreement).

As indicated earlier, the results of Section 5 are not affected by the endogenization of the parameter λ via the above negotiation process. In particular, the conditions that determine the stable coalition size k^* that were applied in Section 5, were used also in the above derivations. This concerns the expression for the function $V_s(\cdot)$ in (17), as well as the definition of the function $f(\cdot)$ in (6) that (in conjunction with the function $V_s(\cdot)$) pins down the stable coalition size k^* via the relation $k^* = f(k^*)$, assuming that this leads to an integer value of k^* (see below). This allows us to approach the analysis of the negotiation process also from a different angle.

Formally, we can treat the probability λ also like a *parameter* (similarly as in Section 5). The corresponding value of k^* then follows directly from the relation $k^* = f(k^*)$ (see, e.g., Figure 7). Note, that this relation does not depend on the frictions in the negotiation process, captured by the parameters r and p^s. Then, we can determine combinations of these parameters that are *consistent* with the given values of k^* and λ. In

effect, we are then treating the probability of stopping p^s (or alternatively r) as an endogenous *variable*, while holding λ fixed. This approach leads to a particularly simple formal characterization of the negotiation process, as we will show in the following.

Starting point is the expression for $V_s(\cdot)$ in (17). After rearranging, this condition can be rewritten as follows:

$$\Pi_n(k^*) - \Pi_s(k^*) = \frac{N}{N-k^*} \frac{(1-\lambda\delta)V_s(k^*) - (1-\lambda)\Pi_s(k^*) - \lambda\pi_0}{1-\lambda}.$$

This is inserted in (33) to obtain:

$$N\left[(1-\lambda\delta)V_s(k^*) - (1-\lambda)\Pi_s(k^*) - \lambda\pi_0\right] = (1-\lambda)r\left[(1-p^s)/p^j - (N+1) + k^*\right]. \quad (36)$$

Conditions (36), (32), and (35) are now the system that determines the variables k^*, p^j, and p^s (treating λ as a parameter).

Solving (35) for p^j and inserting in (36), we obtain

$$N\left[(1-\lambda\delta)V_s(k^*) - (1-\lambda)\Pi_s(k^*) - \lambda\pi_0\right] = \frac{r}{p^s}(\lambda - p^s)(N - k^* + 1). \quad (37)$$

Solve (32) for $V_s(k^*)$ and insert in (37) to obtain after rearranging:

$$r\frac{\lambda - p^s}{p^s} = \frac{N}{N-k^*+1}\left(\frac{1-\delta}{\delta}\Pi_s(k^*) - \frac{\pi_0}{\delta}\right). \quad (38)$$

This condition permits a very simple formal characterization of the negotiation process. Fix some value for λ. Then condition (32) (with $V_s(\cdot)$ as given in (17)) yields the equilibrium coalition size k^*. To assure that this is indeed an integer, λ should be chosen from a restricted set (or grid).[36] Now for a given value of r, condition (38) delivers a *unique* value of p^s that is consistent with the equilibrium outcome (i.e., an outcome (k^*, λ) that results endogenously from the above negotiation process). Note, that for a given value of k^*, the right-hand side of (38) is just a constant. Hence, (38) can be solved for p^s which yields a simple relation $p^s = p^s(r)$.

Figure 9 shows combinations of r and p^s that are consistent with equilibrium values λ and k^*. Along the curve $p^s(r)$ that is shown in the figure, the *equilibrium* values for k^* and λ that arise endogenously from the negotiation process (embedded in the overall game) stay the same. The figure illustrates that for any positive value of the parameter r that captures the negotiating costs, there exists a unique value of p^s that is consistent with the given values of k^* and λ.

[36] This restriction allows us to treat k^* as an integer, without using the function $\hat{f}(\cdot)$.

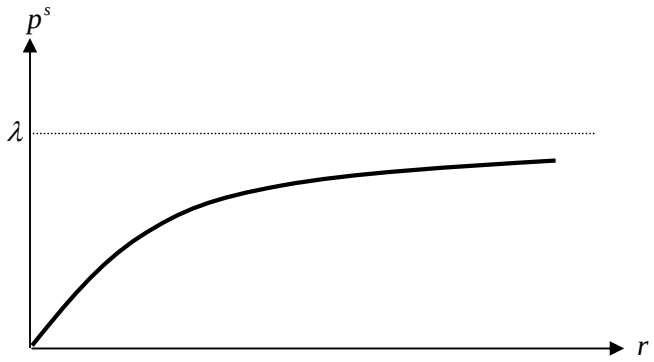

Figure 9: Combinations of r and p^s consistent with equilibrium values λ and k^*

References

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

Barrett, S. (1997) The strategy of trade sanctions in international environmental agreements. *Resource and Energy Economics*, 19, p. 345-361

Barrett, S. (2006) Climate Treaties and "Breakthrough" Technologies. *AEA Papers and Proceedings*, 96, p. 22-25

Battaglini, M. and B. Harstad (2015) Participation and Duration of Environmental Agreements. Forthcoming in *Journal of Political Economy*

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

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, p. 29-67

Finus, M. and S. Maus (2008) Modesty May Pay! *Journal of Public Economic Theory*, 10, p. 801-826

Harstad, B. (2014) The Dynamics of Climate Agreements. Forthcoming in *Journal of the European Economic Association*

Helm, C. and R.C. Schmidt (2015) Climate cooperation with technology investments and border carbon adjustment. *European Economic Review*, 75, 112–130.

Hoel, M. and K. Schneider (1997) Incentives to Participate in an International Environmental Agreement. *Environmental and Resource Economics*, 9, p. 153-170

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

Hong, F. and L. Karp (2012) International Environmental Agreements with mixed strategies and investment. *Journal of Public Economics*, 96, p. 685-697

Karp, L.S. and L. Simon (2013) Participation games and international environmental agreements: A non-parametric model. *Journal of Environmental Economics and Management*, 65, p. 326-344