A simple dynamic climate cooperation model *

Eugen Kováč† Robert C. Schmidt‡

March 22, 2019

Abstract

We introduce a novel framework for analyzing coalition formation, applied to climate cooperation. Our model allows for multiple rounds of negotiations and is able to explain the formation of large coalitions. The incentive of each coalition member to join and subsequently to sign a long-term contract is to prevent inefficient delay that arises as soon as a single country deviates. This undermines the free-rider incentive that destabilizes large coalitions in static coalition formation games. The equilibrium coalition size is then determined by a "threshold effect" due to which deviations of coalition members become unprofitable for sufficiently large coalitions.

JEL Codes: D62, F53, H23, Q54
Keywords: climate treaty, coalition, dynamic game, coordination, delay

*We are particularly grateful to Larry Karp and Hiroaki Sakamoto for insightful discussions. We would also like to thank Scott Barrett, Renaud Foucart, Bård Harstad, Peter Katuščák, Daniel Krähmer, Andreas Lange, Roland Strausz, and the seminar audiences in Aachen, Bonn, Duisburg, Groningen, U Kent, Montreal, Oslo, Ottawa, Potsdam, and Toronto for additional helpful comments and discussions. Financial support from the German Research Foundation (GRK 1659) is gratefully acknowledged. Part of the research was carried out while Schmidt was visiting the University of Oslo, and he is grateful for the institution's hospitality and to Bård Harstad in particular for the friendly invitation and various discussions during the two research stays.

†Mercator School of Management, University of Duisburg-Essen, Lotharstr. 65, 47057 Duisburg, Germany; E-mail: eugen.kovac@uni-due.de.

‡Institute for Microeconomics, University of Hagen, Universitätsstr. 11, 58097 Hagen, Germany; E-mail: robert.schmidt@fernuni-hagen.de

1 Introduction

Climate cooperation is a prime example where a coalition can help to internalize externalities between players, or to provide a public good that would otherwise be undersupplied, due to the free-rider incentive. It is well-documented that international environmental agreements play a significant role in practice, and many of them involve a substantial number of countries (see Barrett, 2003). Economic theory has, however, struggled to provide a sound explanation for successful cooperation in the light of the free-rider incentive: why would a player join a coalition in the first place, if she could also abstain from contributions to the public good (e.g., climate protection) by remaining an outsider, while benefiting from the efforts of others? An influential strand of literature that analyzes participation in international environmental agreements finds that due to the free-rider incentive, usually only small coalitions form, and especially so when the potential gains from cooperation are large (Barrett 1994). Kolstad and Toman (2005) coined the term "Paradox of International Agreements" in the context of the finding that large coalitions are typically only stable when the potential welfare gains from cooperation (compared to the non-cooperative Nash equilibrium outcome) are modest.

In this paper we present a novel theoretical framework that sheds new light on the issue of coalition formation, and climate cooperation in particular. Departing from the bulk of the literature on open membership games, we present a dynamic model, where countries may suspend the current negotiations and continue negotiating in the next period. Introducing dynamics changes each country's trade-off. If no long-term climate contract is signed today, then there is a delay and a new round of negotiations starts in the next period. Such a delay is costly in the short-run, but may be profitable in the long-run if the countries anticipate that a *better* agreement can be signed in the future.

Surprisingly, we find that this simple modification (i.e., allowing countries to restart negotiations tomorrow if no agreement is reached today) of an otherwise standard coalition formation game leads to fundamentally different results. As the main result, we show that large coalitions that achieve substantial welfare gains can be stable under mild conditions. At the heart of our analysis lies an endogenous threshold effect: coalition members only sign an agreement today if the resulting welfare is at least as large as their expected welfare under delay. This requires a sufficiently large number of participants. The corresponding *threshold equilibria* have the property that if a single country deviates, no agreement is signed in the current period and negotiations are delayed.

In line with much of the existing literature, we assume that there is only one (long-term) agreement that can be signed, and each country is free to join the agreement. Once an agreement is signed, the game effectively ends. However, most climate coalition formation models assume that there is only a single participation stage, so that countries

can decide only once and for all if they join the coalition or not.[1] In such a case, countries always sign an agreement, but generally with few members.

Our dynamic model shows that the pessimistic predictions of many static models depend heavily on the one-shot nature of the negotiation process: while a unilateral deviation in a static model leads to the signature of a smaller agreement (with one member less), in our dynamic model a deviation by a country supposed to be a coalition member leads to a period of delay. The incentives to join are, thus, significantly different from those in a static model. Paying more attention to the dynamics of reaching an agreement is, therefore, crucial for a deeper understanding of the trade-offs involved in countries' decisions whether or not to cooperate.

In order to determine a country's welfare in case of delay, it needs to anticipate whether it will become a signatory or not, if an agreement is signed in the future. This creates a coordination problem: For any coalition size, the countries need to be able to determine which of them become coalition members. In order to resolve this coordination problem, we analyze two variants of our model, one where the identity of coalition members is only determined during each round of negotiations (*random membership approach*), and one where there exists some pre-defined ordering of countries (*deterministic membership approach*). Under the random membership approach, countries overcome the coordination problem with the help of an external public randomization device, which (for a given coalition size) selects the identities of the coalition members in each period. The random membership approach may be justified in particular when countries are ex-ante symmetric (as in our model). But even with asymmetric countries, no country would have an incentive to build up a reputation of being *more cooperative* than others (to avoid becoming locked in a coalition, while it is preferable to be a non-signatory).

By contrast, under the alternative deterministic membership approach, for any given coalition size, it is ex-ante known which countries should (in equilibrium) be in the coalition and which should be outside.[2] This approach is more suitable to analyze situations where countries have already built up some reputation for being more or less cooperative than others. In static models, the two approaches are isomorphic and lead to identical predictions about the size of stable coalitions. Most authors have, therefore, paid little attention to how countries overcome the coordination problem at the participation stage, and simply assumed that countries can coordinate to reach the equilibrium coalition size (or one of these coalition sizes in case of multiple equilibria at the participation stage). In a dynamic coalition formation model such as ours, by contrast, it matters for the equilibrium outcome how countries coordinate at the participation stage.

Under the random membership approach, in our model the equilibrium coalition size

[1] See Hoel (1992), Carraro and Siniscalco (1993), Barrett (1994), Dixit and Olson (2000), Karp and Simon (2013).

[2] This latter approach has also been adopted by Battaglini and Harstad (2016).

is determined by two basic motives. On the one hand, by signing an agreement today, coalition members give up the chance to free-ride by becoming non-signatories in the future, when new negotiations start (in case of delay). Therefore, coalition members are willing to sign an agreement today, only if the agreement is sufficiently attractive (relative to the expected outcome of future negotiations) to compensate them for the forgone benefits of free-riding in the future. This makes coalition members demanding, and explains why only large coalitions are stable if countries are sufficiently patient. On the other hand, there is a countervailing force: Countries may be willing to sign a weaker agreement today than what would be expected tomorrow in case of delay, in order to avoid inefficient delay. This undermines the stability of cooperation when countries are impatient, because it makes them less demanding. Overall, we find that an equilibrium coalition size (for a given discount factor) is such that these two motives are balanced.

Our central result (that large coalitions that achieve high welfare gains are stable when countries are sufficiently patient) is preserved also under the deterministic membership approach. However, the underlying intuition is different. If countries are optimistic and anticipate the formation of a large coalition in the next period if current negotiations fail, then countries become *demanding* already in the current period: rather than locking-in an inefficient agreement, they prefer to wait a period until a larger coalition forms in the next period. However, if the same coalition that is expected to form next period, already forms today, then countries are better off signing a long-term agreement already today. Hence, the formation of a large coalition already in the first period is, then, an equilibrium (self-fulfilling expectations).

Related literature

In a closely related paper, Karp and Sakamoto (2018) introduce another dynamic framework that is based on the idea of randomization. While under our random membership approach, the *identities* of countries that are supposed to join a coalition (for some equilibrium coalition size) are determined randomly during each round of negotiation (as long as no long-term agreement has been signed yet), in their model, the randomization device selects among the set of *equilibria* that exist at the participation stage. Unlike in our model, where the game ends when coalition members sign a long-term agreement, in Karp and Sakamoto (2018), the members of today's coalition can decide at the beginning of the next period whether they maintain the coalition, or dissolve it, in which case the randomization device is again used to coordinate on a new stable coalition (without delay). Hence, in their model, an agreement that lasts indefinitely can arise following a number of short-term agreements that last for only one period each. Our model is simpler, as we assume that countries can coordinate immediately on a long-term agreement. Yet, commitment to a long-term agreement is not strictly required even in our framework.

Even if coalition members that signed an agreement today, could decide to dissolve the coalition again at the beginning of the next period, they would have no incentive to do so, as long as the *same* countries would subsequently be asked again if they wish to form a new coalition or not (in the latter case there would be a delay of one period). Hence, in our model there is no incentive to opt out of a coalition for a country that volunteered to join in a previous period.[3] It is also worth mentioning that it is not only the dynamic nature of our model that explains large stable coalition sizes. Simply allowing repeated negotiations (without delay) about short-term agreements would yield the same equilibrium conditions as in the static case. It is rather the opportunity cost of signing a weak agreement today (thereby giving up a possibly better agreement tomorrow) that makes small agreements unstable in our framework.[4]

In another related paper, Battaglini and Harstad (2016) analyze a dynamic climate coalition formation model that, similar to our model as well as Karp and Sakamoto (2018), can help to resolve the Paradox of International Agreements, by showing that large coalitions that achieve substantial welfare gains can be stable in equilibrium. They demonstrate how an endogenous length of the commitment period in conjunction with incomplete contracts that exclude countries' technology investments (leading to a hold-up problem) can explain larger coalition sizes than predicted by static models. While Battaglini and Harstad (2016) find large coalitions only to be stable under those specific circumstances (requiring an endogenous length of the commitment period *and* a hold-up problem at the same time), we demonstrate that large coalitions can emerge under much more general conditions. This neither requires countries to determine the contract duration endogenously, nor the presence of a hold-up problem related with (non-contractible) technology investments. Indeed, to highlight this point, our paper abstracts from technology investments altogether, and focuses entirely on countries' emission choices. In our model, it is the sheer *possibility* to negotiate again in the future that makes large agreements stable, while such agreements would not be stable in a comparable static framework where countries can negotiate only once.

In order to facilitate a comparison of our model to Battaglini and Harstad (2016), we also consider an extension where the countries may sign short-term agreements under the deterministic membership approach. However, we depart in one specific detail from the way in which these authors model short-term agreements. It turns out, that the results of dynamic climate coalition formation games such as our's and their's, are rather sensitive to such modeling details (i.e., if countries are able to sign short-term agreements or not, and if so, the details of how the negotiations about these are modeled). We explain in Section 5 how we model short-term agreements, and why we believe that our assumptions

[3]To simplify the exposition, we do not allow countries to opt out. Hence, we assume that a long-term agreement *is* binding once it has been signed.

[4]We are grateful to Hiroaki Sakamoto for pointing this out.

are plausible. A small change in those details explains why Battaglini and Harstad (2016) only find large coalitions to be stable in a much more complex framework that involves also countries' R&D investments and a hold-up problem related with those. The sensitivity of our models to such seemingly small modeling differences, should however not be seen as a failure of our models to adequately describe possible trade-offs in real world negotiations. Instead, it may well be, that negotiations in the real world are equally sensitive to such details in the mode of negotiating. This may explain why some negotiations led to a success (Paris Agreement) while others failed spectacularly (Copenhagen climate summit), and why negotiators have paid so much attention to the mode of negotiating. For example, while in previous negotiations countries tried to commit to emissions targets, later they switched to the so-called "pledge-and-review" process.[5]

The open membership approach to study climate negotiations builds on earlier papers by d'Aspremont et al. (1983) and Palfrey and Rosenthal (1984). Their concepts have later been adopted by Barrett (1994), Carraro and Siniscalco (1993), and other papers that followed in this strand of literature, to analyze the formation of International Environmental Agreements (IEAs) using game-theoretic tools.[6] Similar to Karp and Simon (2013), we also adopt a non-parametric modeling approach, that does not rely on specific functional forms. While these authors demonstrate that under very specific conditions, a large coalition that achieves substantial welfare gains can form even in a static model, we show that a similar result can be obtained in a dynamic coalition formation model under much more general conditions.[7]

Hong and Karp (2012) consider mixed-strategy equilibria at the participation stage, which allows for the possibility of a coordination failure. In their model, there is a critical coalition size below which no positive abatement efforts are implemented.[8] In our model, a coalition may form that fails to sign an agreement (off the equilibrium path).[9] However, this is not because an agreement would not be welfare-enhancing, but because coalition members anticipate an even better outcome (from their perspective) in the future. In an extension, Hong and Karp (2014) study the interaction of endogenous risk (stemming from mixed strategies) and exogenous risk (stemming from uncertainty about costs and benefits), when countries are risk averse.

[5]Inspired by the Paris climate-change agreement, pledge-and-review bargaining is analyzed formally by Harstad (2019).

[6]For an overview, see Barrett (2005) and Finus (2008).

[7]Alternative approaches to model coalition formation are analyzed by, among others, Bloch (1996) and Ray and Vohra (1997, 2001), and applied to analyze climate treaties by de Zeeuw (2008) and Diamantoudi and Sartzetakis (2015). A mechanism design approach to climate agreements is presented by Martimort and Sand-Zantman (2016).

[8]This effect stems from the assumption of linear benefits and costs of abatement (up to a maximum level), that effectively leads to binary abatement decisions.

[9]Relaxing the Markov restriction, delay can also occur on the path, as we demonstrate in the Supplementary Appendix B.3.

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). Helm and Schmidt (2015) analyze the size of stable coalitions under trade with border carbon adjustment. 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. Hoel and Schneider (1997) assume there is a social cost of non-cooperation. The role of side-payments in fostering participation is investigated by Barrett (2001) and Carraro, Eyckmans, and Finus (2006), among others. Karp (2010) analyzes the role of safety valves in permit trading. Barrett (2006) and Hoel and de Zeeuw (2010) consider the possibility of a technological breakthrough in low-carbon technologies and how it affects the stable coalition size. Cooperation in the light of an approaching climate catastrophe is analyzed by Barrett (2013) and Schmidt (2017).

Finally, there is a literature analyzing self-enforcing agreements in repeated emissions games (see Barrett 1994; Harstad, Lancia, and Russo 2019 among others). Barrett (1994) shows that even for discount factors arbitrarily close to 1, full cooperation may not be sustainable if countries' strategies must be renegotiation-proof. By contrast, we assume that countries can sign a *binding* long-term climate contract (so compliance is not an issue). Breitmeier, Young, and Zürn (2006) provide empirical evidence for their finding that most international environmental rules are complied with, most of the time. See also Young (2011).

2 "Toy model"

Before introducing our full dynamic model, we first present a simple "toy" version of our model, to illustrate the basic ideas that are underlying our approach. Our full model involves considerable technical detail and therefore requires more heavy notation. Analyzing first a drastically simplified version of our model will foster intuition and thereby facilitate the understanding of our general setup that is analyzed in the following sections. Below, we will also provide a brief discussion and motivate different extensions of our full model that we analyze in later sections (Outlook).

The (presumably) simplest abatement game among N countries involves a binary abatement choice of each country: abate or not abate. Assuming a constant marginal benefit of $b > 0$, each country's total benefit from abatement (that is a pure public good) is bk, when k countries decide to abate. The cost incurred by a country that chooses to abate is c, while there are no costs if the country does not abate. In order to make this an interesting problem, we assume that $b < c$ (otherwise, 'abate' is a dominant strategy for each country). A coalition of k countries, by contrast, chooses to abate if and only if $bk \geq c$, or equivalently $k \geq c/b$. For simplicity of the exposition, let us assume that the

coalition members abate if indifferent.

In a *static* climate coalition formation model, the abatement game is embedded in a three-stage game. In the first stage, each of the N countries can decide to join or stay outside of the coalition. In stage 2, the coalition members collectively determine their abatement activities, so as to maximize their joint welfare. And in stage 3, the remaining outsiders determine their efforts non-cooperatively (see Barrett, 1994). Then the above abatement game yields the following equilibria in pure strategies: (i) a set of trivial equilibria in which a sufficiently small number of countries enters the coalition such that nobody abates, even if one additional country were to join, so that $k < c/b - 1$, and (ii) a *threshold equilibrium*, where k is the (unique) integer k that satisfies

$$c/b \leq k < c/b + 1. \tag{1}$$

In this case, k is just large enough, so that a coalition of this size decides to abate (left inequality), but it is not so large, that the coalition would still decide to abate if one country were removed from it (right inequality). Hence, the "last" country to join induces the coalition to become active. No additional country would volunteer to join, as the private costs (equal to c) of doing so outweigh the additional private benefits (equal to $b < c$).

Now our *dynamic* climate cooperation model is based on the simple idea, that abatement efforts are chosen not just in a single period, but in infinitely many periods. Countries can negotiate about a long-term climate agreement in every period. However, once a long-term agreement has been signed by a number of countries, the game ends. In other words, there can be only one climate change agreement, and once an agreement is reached, countries are committed to stick with it in all future periods. However, as long as no agreement has been signed yet, in every new period, countries can start to negotiate again (in the same way as in the static game). This is presumably the simplest possible extension of the static coalition formation game to a dynamic framework.

Let us now analyze how the equilibrium outcomes change for our simple emissions game introduced above, as compared to the static approach. As in most of our main analysis, we restrict attention to Markov perfect equilibria.[10] Hence, if along the equilibrium path, a coalition of size k is supposed to form and sign a long-term climate agreement in period 1, but (say, following some deviation) the coalition that actually forms signs no agreement, then all countries expect that in the following period 2, again a coalition of size k forms. Furthermore, under a *deterministic membership approach* which we assume here,[11] the identities of these k countries in the next period will be the same as in the current period (if countries had chosen to stick with their equilibrium strategies today).

[10] We thereby rule out tacit collusion or *grim trigger strategies*, which may sustain abatement efforts even in the absence of a binding climate contract.

[11] In the following sections, we also study an alternative *random membership approach*.

Then clearly, the set of trivial equilibria remains the same as in the static game. However, the set of threshold equilibria is now changed.

Formally, a threshold equilibrium in the dynamic game must simultaneously fulfill the following two conditions (with δ being the discount factor):

$$\frac{1}{1-\delta}(bk-c) \geq \frac{\delta}{1-\delta}(bk-c), \tag{2}$$

$$\frac{1}{1-\delta}[b(k-1)-c] < \frac{\delta}{1-\delta}(bk-c). \tag{3}$$

The first condition states that if a coalition of size k forms in some period t, then coalition members sign a long-term agreement in that period, thereby ending the game, because the discounted payoff per signatory (left-hand side) is at least as large as the payoff if the coalition does not sign an agreement, in which case new negotiations start in the next period. In the latter case, the payoff in the current period of all countries is zero, as nobody abates. It is easy to see, that condition (2) reduces to $k \geq c/b$, the same condition as in the static game for a coalition to become *active*. Intuitively, if it is profitable for coalition members to sign a long-term agreement of size k in the next period, assuming that no agreement is signed today, then it is profitable for the same k countries to already sign such an agreement today.[12]

The second condition, however, has changed in the dynamic game, as compared to the static one. Now (3) requires that if one country that should be in the coalition in period t along the equilibrium path (payoff on the left-hand side of the inequality) does not join, the remaining coalition members decide not to sign a long-term agreement today to avoid locking-in an inefficient agreement. They prefer to wait a period for the remaining country to join the coalition (right-hand side).

Rewriting conditions (2)–(3), we obtain

$$\frac{c}{b} \leq k < \frac{c}{b} + \frac{1}{1-\delta}. \tag{4}$$

Now there might be several equilibrium coalition sizes, while the static equilibrium also remains an equilibrium in the dynamic model.[13] Moreover, if δ is sufficiently large (close to 1), the set of threshold equilibria in the dynamic game can be substantially larger than in the static game, and even the grand coalition (i.e., $k = N$) can be sustained if the right-hand side is greater than N. This points to a significant degree of multiplicity of threshold equilibria in the dynamic game, as compared to the unique one in the static

[12]The incentive of each coalition member to join the coalition in the first place (so-called *internal stability condition* in static coalition formation games) is captured by the same conditions ((2) and (3)). External stability is then automatically satisfied, as no additional country would join a coalition that is willing to sign an agreement and abate also without this country.

[13]In the case $\delta = 0$, the second condition collapses to $k < c/b+1$ and we obtain only a single equilibrium coalition size, which is the same as in the static game.

game. The reason is self-fulfilling expectations: if countries are optimistic and believe that a large coalition will sign an agreement in the next period, then the same coalition may as well form already today and sign an agreement.

Outlook

In the remainder of this paper, we generalize the above ideas by considering a dynamic coalition formation game. We take a reduced form approach and consider general welfare functions that depend only on the number of signatories. These welfare functions may be considered as equilibium outcome of an underlying abatement game. We present examples of such abatement games with continuous abatement choices (as opposed to just binary ones in our "toy model"). Furthermore, we introduce an alternative assumption of how the identities of coalition members in the next period are determined, in case the current coalition decides not to sign a long-term agreement (random membership approach). However, under both deterministic and random membership approaches, large coalitions can form in equilibrium if countries are sufficiently patient. We illustrate our results for the random membership approach in Section 4 with the help of specific functional forms known from the literature. In Section 5, we extend our model to allow countries to sign a short-term agreement in a period in which no long-term agreement is (or has been) signed yet. We demonstrate that the possibility to sign short-term agreements can foster participation in a long-term agreement. Section 6 analyzes the (simpler) deterministic membership approach (which we assumed also in the toy model) within our general modeling framework. Other extensions include a finite number of periods in which countries can negotiate (Section B.2 in the Supplementary Appendix), and a departure from the Markov restriction towards non-stationary equilibria (Section B.3).

3 Full dynamic model

There are N ex-ante symmetric countries that negotiate about an international environmental agreement (IEA). The time horizon is infinite. The negotiations start in period $t = 1$, and as long as no agreement has been signed in any previous period, a new round of negotiations starts in each period. If an agreement is reached in period t, an IEA is implemented from that period onwards and the game (effectively) ends. As usual in this strand of literature, we restrict our attention to the case where only one coalition (and not multiple coalitions) can sign a binding long-term climate contract.

If the coalition signs an agreement, the abatement targets of the signatories are chosen so 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. Let $\Pi_s(k)$ denote the present value of payoffs (welfare) of a

signatory of a long-term agreement with k members and $\Pi_n(k)$ denote the present value of payoffs that a non-signatory obtains when k other countries sign the agreement. The welfare functions Π_s and Π_n can be derived from an underlying emission game. We do not model such a game specifically here, but rather take a reduced form approach without imposing specific functional forms. We provide several examples of emission games and corresponding welfare functions in Section 4.

We assume that the welfare $\Pi_s(k)$ is the same for all signatories and the welfare $\Pi_n(k)$ is the same for all non-signatories. Moreover, we assume that the welfare is independent of the time when the agreement is signed.[14] Although it is sufficient to consider $\Pi_s(k)$ and $\Pi_n(k)$ only for integer values of k, it will turn out to be convenient to define them over the whole interval $[0, N]$ and to assume that the functions $\Pi_s(\cdot)$ and $\Pi_n(\cdot)$ are continuously differentiable.

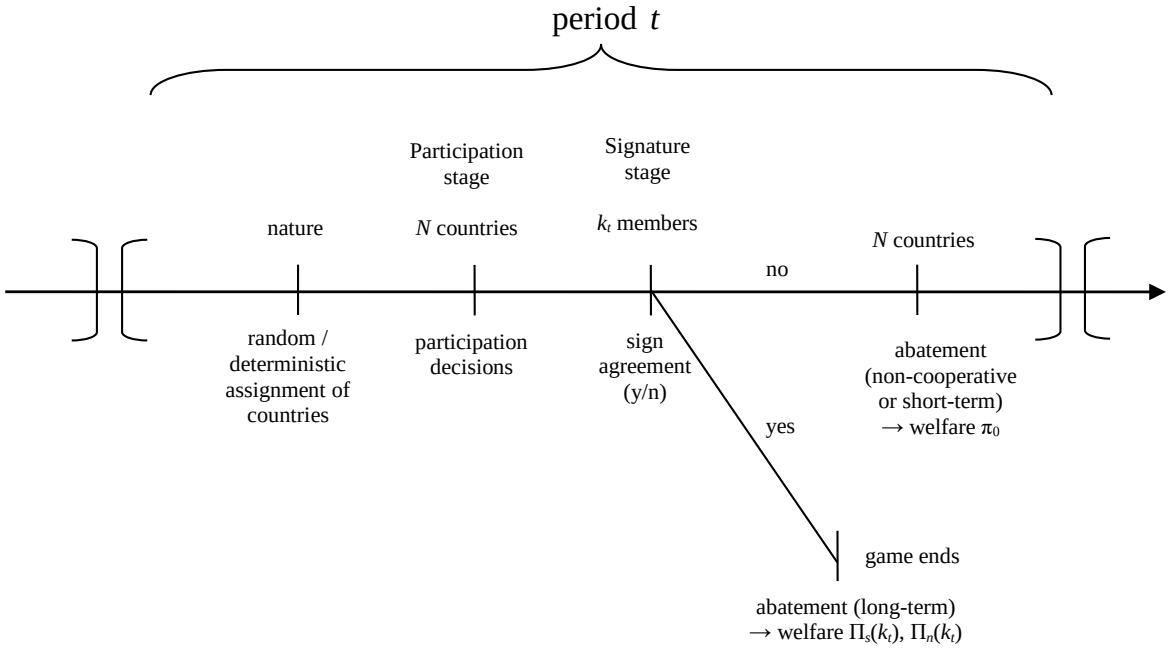

Figure 1: Timing of actions in period t

In our full dynamic game, each round of negotiations involves two stages (see Figure 1). In the *participation stage*, each country decides individually whether to join the coalition or not; let k_t denote the number of countries who join in period t (provided that period is reached). In the *signature stage* stage, the k_t coalition members decide whether to sign an agreement in this period or not. If they indeed sign the agreement, the game (effectively) ends, the IEA is implemented and the resulting payoffs from subsequent abatements,

[14]Later, we consider a special case, where these payoffs are outcomes of time-independent per period interactions. For the time being, the dynamics of the interaction after an agreement is signed, are irrelevant. We are interested in the dynamics of reaching an agreement. The setup resembles stopping games, where stopping corresponds to a coalition signing a binding long-term climate agreement.

namely $\Pi_s(k_t)$ for a signatory and $\Pi_n(k_t)$ for a non-signatory, are realized. If they do not sign, the coalition dissolves, and all N countries choose their emissions non-cooperatively in that period.[15] A new round of negotiations then starts in the next period. Let $\pi_0 \geq 0$ denote the (constant) per-period payoff for a country in a period where no agreement has been reached yet. Let $\delta \in (0,1)$ be the common discount factor.

As a benchmark, let k^{st} be the stable coalition size in the static case, based on the welfare functions $\Pi_s(k)$ and $\Pi_n(k)$, that is obtained when countries can negotiate only in period 1, or if future payoffs are fully discounted away (i.e., $\delta = 0$). The static model has been studied thoroughly in the literature (e.g., Barrett 1994; Karp and Simon 2013). The following conditions of *external* and *internal stability* characterize a Nash equilibrium at the participation stage in the static case:

$$\Pi_n(k) \geq \Pi_s(k+1), \tag{ES}$$

$$\Pi_s(k) \geq \Pi_n(k-1). \tag{IS}$$

External stability (ES) requires that outsiders (non-signatories) have no incentives to join the coalition, while internal stability (IS) requires that insiders (signatories) have no incentives to deviate by staying outside the coalition.

In order to analyze our dynamic model, we need to impose additional structure on the payoff functions $\Pi_s(k)$ and $\Pi_n(k)$.

Assumption 1 (Welfare comparison)**.** There is $k_0 \in [0, N)$ such that $\Pi_n(k_0) = \Pi_s(k_0)$ and $\Pi_n(k) \lesseqgtr \Pi_s(k)$ for $k \lesseqgtr k_0$. In addition, $\Pi_n(k_0) = \Pi_n(0)$.

Assumption 2 (Monotonicity)**.** $\Pi_n(k)$, $\Pi_s(k)$, and $k\Pi_s(k) + (N-k)\Pi_n(k)$ are strictly increasing for $k > k_0$.

Assumption 3 (Free-rider incentives)**.** There is a unique threshold $\tilde{k} \in (k_0, N-1)$ such that $\Pi_n(\tilde{k}) = \Pi_s(\tilde{k}+1)$. In addition,

(a) $\Pi_n(k) - \Pi_s(k+1)$ is strictly increasing for $k > \tilde{k}$.

(b) $\Pi_n(k) < \Pi_s(k+1)$ for $k < \tilde{k}$.

Assumption 4 (Non-cooperative payoff)**.** $\Pi_n(0) \leq \dfrac{\pi_0}{1-\delta} \leq \Pi_s(k^{st})$.

According to Assumption 1, the welfare of a non-signatory is larger than that of a signatory (for a sufficiently large coalition size), because non-signatories enjoy the same benefits of abatement as the signatories (with pollution being a global public bad), but incur lower costs of abatement. Only for small coalition sizes, this relation may be

[15] In Section 5, we assume instead that countries negotiate about a *short-term agreement* in a period in which no long-term agreement is reached. Our general modeling framework developed here allows us to accomodate also this case.

reversed. The value k_0 represents the smallest coalition size above which non-signatories are better off than signatories. This is generally also the critical coalition size above which signatories reduce their emissions more than the non-signatories. The second part of Assumption 1 postulates that both for coalition sizes $k = 0$ as well as $k = k_0$, the countries attain an identical welfare, since in both cases there are N countries that behave in the same way. Note also that for a given emissions game, the value of k_0 is determined by the underlying welfare functions $\Pi_s(k)$ and $\Pi_n(k)$. In our examples in Section 4, we obtain $k_0 = 0$ when assuming that the non-signatories do not reduce their emissions, but also larger values of k_0, depending on the shape of the underlying benefits from abatement.

Assumption 2 reflects the property that in a sufficiently large coalition, with increasing coalition size, the signatories lower their emissions more and more to internalize environmental externalities between them. This consequently increases the welfare of each individual country, as well as the total welfare.

Assumption 3 is a single-crossing assumption for the free-rider incentives represented by the expression $\Pi_n(k)-\Pi_s(k+1)$. This expression corresponds to the gain from leaving a coalition of $k+1$ countries and it plays a central role in determining the equilibrium of the static model. It is assumed that the free-rider incentives are negative for small coalitions and increasing for large coalitions, as larger coalitions internalize more of the externalities.

Finally, Assumption 4 implies that welfare from non-cooperation in all periods (equal to $\pi_0/(1-\delta)$ per country) is bounded from below by the welfare when no country signs an agreement and bounded from above by a signatory's welfare in the equilibrium of the benchmark static case. It follows from Assumption 1 that $\Pi_s(k_0) \leq \pi_0/(1-\delta)$. Although we treat π_0 as an independent parameter (bounded only by Assumption 4), in specific models (see Section 4), its value can be derived as an equilibrium payoff from the same underlying interaction as the welfare functions Π_s and Π_n.

Before proceeding with the analysis of the dynamic model, let us point out that the above assumptions also provide enough structure to characterize the equilibrium coalition size in the *static* model. With threshold $\tilde{k}$ defined in Assumption 3, $k^{st} = \lceil \tilde{k} \rceil$ is an equilibrium coalition size in the static model.[16] In order to avoid duplicity and a tedious discussion of a knife edge case, we will assume that $\tilde{k}$ is *not* an integer (otherwise, both $\tilde{k}$ as well as $\tilde{k}-1$ satisfy internal and external stability). Moreover, due to Assumption 4, the equilibrium coalition of k^{st} countries prefers to sign an agreement compared to no agreement at all.[17]

[16]Given an arbitrary $x \in \mathbb{R}$, $\lceil x \rceil$ is defined as the unique integer such that $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. It is the smallest integer at least as large as x.

[17]It is worthwhile to point out that our "toy model" from Section 2 is not nested in the general model, since there the free-rider incentive is equal to $bk - [b(k+1) - c] = c - b$, which is constant. Moreover, the "toy model" features a threshold equilibrium also in the static benchmark case. By contrast, due to

Now consider some coalition of size k_t that would form in period t, provided that period is reached. If $k_t \in [1, N)$, 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. Clearly, the incentives to become a coalition member today and to sign the agreement depend on whether a country expects to become a coalition member in the *next* period (if no agreement is reached today).

In the following, we will assume that the identity of the coalition members in period t (for some given coalition size k_t) is randomly determined. Intuitively, as countries are ex-ante symmetric, there is no reason why any specific country should be more likely to join the coalition than another country. Hence, if countries coordinate on a coalition size of k_t, the ex-ante probability of any country to become coalition member is k_t/N. To fix ideas, we assume that countries coordinate with the help of a *randomization device* (nature) that selects an *assignment* (or an ordering) of countries (see Figure 1). Of course, the actual participation decision of a country can differ from the recommendation, as the participation decision of each country remains voluntary and non-cooperative. However, in equilibrium each country is willing to follow the recommendation. Conceptually, this corresponds to a correlated equilibrium, where, for a given coalition size k_t, the public randomization device selects randomly one of the $\binom{N}{k_t} = \frac{N!}{k_t!\,(N-k_t)!}$ equilibria with coalition size k_t.[18]

We say that the negotiations in period t are *successful* if the coalition signs an agreement. Otherwise, we say that the negotiations *have failed*. We assume throughout the paper that the coalition members in period t use an *unanimity rule* when deciding whether or not to sign a climate contract. Hence, every country that has joined the coalition has a veto right, and the negotiations in period t fail as soon as at least one coalition member uses its veto right.[19] For most results in this paper, the choice of the decision rule is inconsequential. Nevertheless, we would like to point out that giving each coalition member a veto right at the signature stage gives potentially rise to another coordination problem with other equilibria at the signature stage, where no agreement is signed (leaving each coalition member indifferent between signing and not). Similarly as we did regarding the participation stage, we assume that countries can overcome this coordination problem. To this end, we assume that coalition members select a Pareto dominant equilibrium (if such exists) at the signature stage. This is a plausible selection criterion since the countries are engaged in negotiations (see the discussion in Farrell and Maskin

Assumption 4 in our full model, threshold equilibria can only exist in the dynamic case, not in the static one.

[18]The alternative *deterministic membership approach* is analyzed in Section 6. Most of our analysis presented here, remains valid also in this case, as shown formally in Appendix A.2.

[19]This modeling choice is also inspired by the UN climate negotiations that led to the Paris Agreement, where each member country had a veto right. See also Finus and Rundshagen (2003) for an analysis of the role of an unanimity rule.

1989).[20, 21]

Whether the participating countries sign an agreement in period t, depends on their outside option. This is determined by the continuation value in case no agreement is signed. Assuming that $k_{t+1} = k$ countries sign an agreement in period $t+1$, a country achieves in the next period an expected welfare of

$$V(k) = p(k)\Pi_s(k) + (1 - p(k))\Pi_n(k), \tag{5}$$

where $p(k) = k/N$ denotes the probability of being assigned as a coalition member in period $t+1$.[22] Clearly, $\Pi_s(k) < V(k) < \Pi_n(k)$ for all $k \in (k_0, N)$. Moreover, due to Assumption 2, the function $V(k)$ is strictly increasing for $k > k_0$.

A coalition member (and hence, the whole coalition) is willing to sign the agreement in period t rather than not signing and thus delaying negotiations until period $t+1$, if and only if

$$\Pi_s(k_t) \geq \pi_0 + \delta V(k_{t+1}). \tag{6}$$

The left-hand side represents the welfare from signing an agreement among k_t countries in the current period. The right-hand side represents the welfare from a delay when there will be an agreement signed among k_{t+1} countries in the next period. This yields welfare of π_0 in the current period and expected welfare of $V(k_{t+1})$ in the following period.

Let us for any coalition size $k \in [k_0, N]$, define $\xi(k)$ as the critical coalition size such that $\xi(k) \in [k_0, N]$ and

$$\Pi_s(\xi(k)) = \pi_0 + \delta V(k), \qquad \text{or equivalently,} \qquad \xi(k) = \Pi_s^{-1}\left(\pi_0 + \delta V(k)\right). \tag{7}$$

We show in Appendix A.1 (Lemma 2) that the function ξ is well defined and increasing, and that $\xi(k) \in [k_0, N)$ for all $k \in [k_0, N]$. A coalition of size $k_t = \xi(k)$ thus leaves its members indifferent between signing an agreement in the current period, and delaying the negotiations until the next period where a coalition of size $k_{t+1} = k$ would be formed. Of course, $\xi(k)$ may not be an integer in general. Hence, let $\hat{\xi}(k) = \lceil \xi(k) \rceil$ be the smallest integer at least as large as $\xi(k)$. Note, that since k_t is an integer, the inequality $k_t \geq \xi(k_{t+1})$ is *equivalent* to $k_t \geq \hat{\xi}(k_{t+1})$. These considerations together with inequality (6) yield the following lemma (all proofs are relegated to the Appendix).

Lemma 1. *If countries anticipate that a coalition of size $k_{t+1} \geq k_0$ will sign an agreement in the following period (period $t+1$), provided no agreement is signed in period t, then*

[20] Note that if signing an agreement is an improvement for each coalition member, it is also an improvement for non-members who even obtain a higher payoff.

[21] As an alternative approach, one could assume that coalition members decide *collectively* whether to sign an agreement or not (e.g., by delegating their individual decisions to a social planner who decides on behalf of all signatories).

[22] Under the deterministic membership approach, $p(k) = 1$. See Section 6 for further details.

the coalition of size $k_t \geq k_0$ signs an agreement in period t, if and only if $k_t \geq \hat{\xi}(k_{t+1})$.

Note that the finding that function ξ is increasing reflects our intuition that, whenever countries anticipate the formation of a larger coalition in the next period (in case of a delay), the threshold $\hat{\xi}(k_{t+1})$ for signing a long-term agreement in the current period becomes (weakly) larger. In other words, countries are more demanding today when they anticipate a more favorable outcome in the future.

In the main part of this paper, we focus on *Markov perfect equilibria* in pure strategies.[23] This means that countries' strategies may depend only on payoff-relevant events. Given the stationary structure of the model (in particular the time-invariant payoff functions), the only payoff-relevant event is when an agreement is signed by a coalition in some period t. However, this ends the game, so under the Markov restriction, countries' equilibrium strategies are stationary.

First note, that there is a *trivial equilibrium* where countries never sign an agreement, if and only if $\Pi_s(1) \leq \pi_0/(1-\delta)$. Under this condition, it is not profitable for any country to join the coalition (and sign a single-country agreement), provided that no other country joins.[24] From now on we focus on equilibria where the countries indeed sign an agreement. Due to the stationary nature, in equilibrium, countries thus clearly sign the agreement in the first period.[25] Let us denote k^* the equilibrium coalition size. 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 of size $k_{t+1} = k^*$ is expected to form in the next period with probability 1.

In what follows we provide conditions for k^* to be an equilibrium coalition size. As a first observation, consider when the k^* countries are indeed willing to sign an agreement *on the path*. According to condition (6), this is the case when

$$\Pi_s(k^*) \geq \pi_0 + \delta V(k^*), \qquad \text{or equivalently,} \qquad k^* \geq \hat{\xi}(k^*), \tag{8}$$

where the equivalence follows from Lemma 1. Thus, condition (8) is necessary for k^* to be an equilibrium coalition size.[26] This captures a threshold effect regarding the coalition size that arises *endogenously* in our model. In particular, if a coalition forms in a period t that is perceived as "too small" by its members (i.e., $k_t < \hat{\xi}(k^*)$), then this coalition dissolves.[27] Below we show that this logic gives rise to a novel type of equilibrium, where

[23] In Supplementary Appendix B.2 we analyze a variant of the model with finitely many periods of negotiations, and in Supplementary Appendix B.3 we relax the Markov assumption and analyze equilibria with delay. The Markov restriction in the infinite-horizon model narrows the set of equilibrium coalition sizes and leads to sharper predictions.

[24] A single-country agreement can affect payoffs in games with complex interactions, for example, when the signatory becomes a Stackelberg leader (see the Example 4 in Section 4).

[25] Equilibria with delays are characterized in Supplementary Appendix B.3, where we relax the Markov restriction.

[26] Note that $k^* < \xi(k^*)$ would imply that in each period a coalition forms that remains inactive.

[27] Note that since $V(k) \geq \Pi_s(k)$, inequality (8) implies that $\Pi_s(k^*) \geq \pi_0/(1-\delta)$.

coalitional stability is driven by the endogenous threshold effect regarding the minimum size of an active coalition (i.e., a coalition that is sufficiently large for its members to be willing to sign a long-term agreement). Furthermore, there can also be an equilibrium that parallels the one in the *static model*. In that case the endogenous threshold effect does not play any role.

Next, we argue that (assuming that k^* satisfies (8)) the external stability condition is the same as in the static model. To see this, consider a potential deviation by an outsider who joins the coalition. In such a case, we obtain a coalition of size k^*+1 in the current period. This coalition clearly signs an agreement, because (by Lemma 1) even a smaller coalition of only k^* countries would sign an agreement. Comparing the payoff of the deviating country, $\Pi_s(k^*+1)$, to the equilibrium payoff $\Pi_n(k^*)$, we indeed obtain the same external stability condition (ES) as in the static case. In addition, notice that for all $k \geq k^{st}$, external stability is satisfied.

Finally, let us characterize internal stability in the dynamic game. To this end, consider a deviation by an insider (i.e., a country assigned to be a coalition member by the public randomization device in period t) who deviates by not joining the coalition. The payoff from such a deviation depends on whether the remaining k^*-1 countries sign an agreement or not. We distinguish two cases. First, let (8) hold with a strict inequality, i.e., $k^* > \hat{\xi}(k^*)$. Then, since both k^* and $\hat{\xi}(k^*)$ are integers, we have $k^*-1 \geq \hat{\xi}(k^*)$. Due to Lemma 1, the remaining k^*-1 countries sign an agreement, yielding a payoff of $\Pi_n(k^*-1)$ to the deviating country. Comparing it to the equilibrium payoff $\Pi_s(k^*)$ from no deviation, we obtain the same internal stability condition (IS) as in the static model. Now, since k^{st} is the only integer satisfying both external and internal stability, the only equilibrium coalition size that satisfies $k^* > \hat{\xi}(k^*)$ is the static one, i.e., $k^* = k^{st}$.

Second, let (8) hold with an equality, i.e., $k^* = \hat{\xi}(k^*)$. In that case, after the deviation of an insider, the coalition of size k^* is just large enough to sign an agreement. Thus, the coalition of the remaining k^*-1 does not sign an agreement in the current period, causing a delay with payoff of $\pi_0 + \delta V(k^*)$ to the deviating country. Such a deviation is not profitable, since by assumption $\Pi_s(k^*) = \pi_0 + \delta V(k^*)$ in this case. Thus, there is no profitable deviation for the insiders when $k^* = \hat{\xi}(k^*)$.

Let us sum up the above findings. For k^* to be an equilibrium coalition size in the dynamic model, then it must hold that either

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

or

$$\Pi_n(k^*) \geq \Pi_s(k^*+1), \quad \Pi_s(k^*) \geq \pi_0 + \delta V(k^*), \quad \text{and} \quad k^* = \hat{\xi}(k^*). \tag{10}$$

In both cases, the first two conditions reflect the individual incentives of a country whether or not to *join* the coalition (external and internal stability), whereas the third condition

reflects the incentives of the coalition members whether or not to *sign* an agreement. Conditions (9) and (10) specify two equilibrium types, which differ in what happens after one of the countries deviates and does not join the coalition: In the second equilibrium type that is based on the threshold effect regarding the coalition size, a deviation by a single country is sufficient to induce a delay in the negotiations. In the first equilibrium type, also a coalition of size $k^* - 1$ signs an agreement (following a deviation).

The case when $k^* = \hat{\xi}(k^*)$ (second equilibrium type) represents a crucial difference to the static model, where a coalition still signs an agreement after one of its member has deviated. Now a deviation by an insider causes a delay. As we have shown above, the second condition in (10) is implied by the third condition. The static internal stability condition is then effectively replaced by the new condition $k^* = \hat{\xi}(k^*)$, which can be rewritten as[28]

$$\xi(k^*) \leq k^* < \xi(k^*) + 1, \tag{11}$$

which is equivalent to $\Pi_s(k^* - 1) < \pi_0 + \delta V(k^*) \leq \Pi_s(k^*)$. As we will see later, the two inequalities in (11) yield an upper and a lower bound on the equilibrium coalition size.

Summing up the above arguments, we can formulate the following two propositions.

Proposition 1. *There is a Markov perfect equilibrium of the dynamic game with coalition size k^{st}, if and only if $k^{st} \geq \hat{\xi}(k^{st})$, or equivalently $k^{st} \geq \xi(k^{st})$.*

Proposition 2. *In any non-trivial Markov perfect equilibrium of the dynamic game, the coalition is at least as large as in the equilibrium of the static model ($k^* \geq k^{st}$). An integer $k^* > k^{st}$ (where $k^* \leq N$) is an equilibrium coalition size, if and only if it is a fixed point of function $\hat{\xi}$, i.e.,*

$$k^* = \hat{\xi}(k^*). \tag{DS}$$

As argued above, the new dynamic stability condition (DS) replaces the internal stability condition (IS) from the static model. We will be particularly interested in equilibria characterized by this condition, hence, equilibria with a coalition size k^* that is larger than in the static model. Let us thus analyze fixed points of the function $\hat{\xi}$. In Appendix A.1, we show that the function $\hat{\xi}$ has indeed a fixed point in the interval $(k_0, N]$ (Lemma 3).[29]

As follows from Proposition 2, such a fixed point represents an equilibrium coalition size, if it is larger than the static coalition size k^{st}. In general, the equilibrium coalition size in the overall game does not need to be unique. In particular, $\hat{\xi}$ may have several

[28]The former follows directly from the definition of $\hat{\xi}$, since the equality $\lceil \xi(k) \rceil = k$ can also be rewritten as $\xi(k) \leq k < \xi(k) + 1$ (see footnote 16).

[29]Note that in the special case when $\Pi_n(k_0) = \pi_0/(1-\delta)$, i.e., when the left condition in Assumption 4 holds with equality, a trivial fixed point of ξ is $k = k_0$. To see this, recall that due to Assumption 1, we have $V(k_0) = \Pi_s(k_0) = \Pi_n(k_0) = \pi_0/(1-\delta)$, and thus $\Pi_s(k_0) = \pi_0 + \delta V(k_0)$, or equivalently $k_0 = \xi(k_0)$. If, in addition, k_0 is an integer, it is also a fixed point of $\hat{\xi}$.

fixed points if ξ has several fixed points or when ξ has a slope close to 1. The latter case is illustrated in Figure 2, that shows the functions ξ and $\hat{\xi}$, based on a specification of payoff functions Π_s and Π_n from an example introduced in the following section. (We omit further details at this point.) As can be seen from the figure, ξ has (besides k_0) a single fixed point equal to approximately 5.4. However, $\hat{\xi}$ has three fixed points: $k = 6$, $k = 7$, and $k = 8$.

As suggested by (11), in order to characterize equilibrium coalition sizes, we need to identify the fixed points of functions ξ and $\xi + 1$. The following assumption provides sufficient conditions, assuring that each of these functions indeed has at most one fixed point. This allows for a particularly simple and parsimonious characterization of equilibrium outcomes. Below we also provide a rationale for the assumption.

Assumption 5 (Single-crossing)**.** (a) There exists $\underline{k} \in [k_0, N]$ such that $k \lesseqgtr \xi(k)$ if $k \lesseqgtr \underline{k}$.

(b) There exists $\bar{k} \in [k_0 + 1, N]$ such that $k - 1 \lesseqgtr \xi(k)$ if $k \lesseqgtr \bar{k}$.

Assumption 5 postulates that each of the functions ξ and $\xi + 1$ has at most one fixed point, and if it does, it is the point $\underline{k}$ and $\bar{k}$, respectively. Moreover, at the fixed point, the corresponding function (ξ or $\xi + 1$) crosses the 45°-line *from above*. The assumption also allows for cases, where $\underline{k} = k_0$, when ξ lies below 45°-line on the interval $[k_0, N]$, and where $\bar{k} = N$, when $\xi + 1$ lies above 45°-line on the interval $[k_0 + 1, N]$. In order to avoid a tedious discussion of non-generic cases, we assume that N is *not* a fixed point of $\xi + 1$ (i.e., that $\xi(N) \neq N - 1$). Moreover, it clearly follows from Assumption 5 that $\underline{k} < \bar{k}$.

Intuitively, the function ξ captures the willingness of coalition members to sign a long-term agreement (or to use their veto right instead). There are *two basic motives* that determine the willingness of countries to sign a long-term agreement today: (i) the requirement of a sufficiently strong agreement in order to compensate its members for the forgone opportunity to become free-riders in the future (with probability $1-p(k_{t+1})$); and (ii) the willingness to sign something weaker today than what would be expected in the future in order to avoid inefficient delay. A fixed point of the function ξ is where these two motives are balanced for a marginal country, i.e., when the integer constraint on k is neglected. The fixed points of the function $\hat{\xi}$ are the coalition sizes where the two motives are (almost) balanced for a non-marginal country, i.e., when the integer constraint on k is taken into account.[30]

The first of the two basic motives mentioned above becomes *weaker* when a larger coalition is expected to form tomorrow in case of a delay, because the probability to

[30]They are not exactly balanced, unless $\hat{\xi}$ coincides with ξ at a fixed point of ξ which requires that the latter is an integer. Otherwise, at a fixed point of $\hat{\xi}$ the coalition is just large enough to sign an agreement. Hence, the motive to sign may be slightly larger than the motive to delay.

become non-signatory in the next period ($1 - p(k_{t+1})$) is then smaller. The second effect results from impatience (discounting of future payoffs) and does not directly depend on the coalition size. Therefore, if the first effect is sufficiently strong for small k, it dominates (hence, $\Pi_s(k) < \pi_0 + \delta V(k)$ and thus $\xi(k) > k$), whereas it always vanishes for coalition sizes close to N (as $1 - p(k) = (N-k)/N$ so that $\Pi_s(k) > \pi_0 + \delta V(k)$ and $\xi(k) < k$ for k sufficiently large. Assumption 5 assures that there is a smooth transition from the region where the first effect dominates (for small k) to the region where the second effect dominates (for large k).

Note that Assumption 5 is trivially fulfilled if ξ is concave. Lemma 4 in Appendix A.1 provides an alternative sufficient (albeit not necessary) condition for Assumption 5(a) to hold, based on monotonicity of the functions Π_s and Π_n. It can be verified that the assumption in Lemma 4 is indeed satisfied in all our examples in Section 4 (in Examples 1 and 3, the ratio in Lemma 4 is simply a constant).

Now we are ready to provide a characterization of the set of equilibrium coalition sizes in the dynamic game. Recall from Proposition 1 that the sufficient and necessary condition for the stable coalition size in the static model, k^{st}, to be an equilibrium coalition size also in the dynamic game is $k^{st} \geq \xi(k^{st})$. Under Assumption 5, this is equivalent to $k^{st} \geq \underline{k}$. Moreover, it follows from Proposition 2 that any other equilibrium coalition size k^* is a fixed point of $\hat{\xi}$ such that $k^* > k^{st}$. It follows from (11) that, under Assumption 5, the set of all fixed points of $\hat{\xi}$ is in the interval $[\underline{k}, \bar{k})$. Depending on the size of k^{st} relative to $\underline{k}$ and $\bar{k}$, we thus obtain the following characterization of equilibrium coalition sizes:

Proposition 3. *The set of all equilibrium coalition sizes (in any non-trivial Markov perfect equilibrium in pure strategies) in the dynamic game is*

(a) all integers from the interval $[\underline{k}, \bar{k})$*, if* $k^{st} < \underline{k}$*;*

(b) all integers from the interval $[k^{st}, \bar{k})$*, if* $\underline{k} \leq k^{st} < \bar{k}$*;*

(c) $\{k^{st}\}$*, if* $k^{st} \geq \bar{k}$*.*

The proposition implies that any equilibrium coalition sizes in the dynamic game are bounded from below by $\max\{k^{st}, \underline{k}\}$ and from above by $\max\{k^{st}, \bar{k}\}$. In the case where $k^{st} < \underline{k}$ (case (a)), we obtain that the static equilibrium coalition size is *not* an equilibrium coalition size in the dynamic model. To understand the intuition why this is the case, suppose to the contrary that in the first period of the dynamic game, a coalition of size k^{st} forms and its members are willing to sign a long-term agreement (hence, k^{st} is an equilibrium coalition size). Then it must hold that in case of delay (i.e., if the coalition members do not sign an agreement in the first period), another coalition of size k^{st} forms in the next period. But since k^{st} is typically small (see Barrett 1994), this implies that for each member of the coalition in period 1, the probability to become a free-rider (non-signatory) in the next period is high. This makes it unprofitable to sign an agreement

in the first period. By contrast, equilibria that fulfill (DS) entail (in case (a)) a higher coalition size than k^{st} (see Section 4 for examples). The chances of becoming a free-rider in the next period in case of a delay are, then, smaller, so that the members of a larger coalition are willing to sign an agreement (in the first period). Figure 2 illustrates these equilibria for a specific example (Example 2 in Section 4) where case (a) from the above proposition applies for the underlying parameter values. Observe that all fixed points of $\hat{\xi}$ (namely $6, 7, 8$) are integers from the interval $[\underline{k}, \bar{k})$, where $\underline{k} \approx 5.4$ is a fixed point of ξ and $\bar{k} \approx 8.2$ is a fixed point of $\xi + 1$ (note that $k_0 = 1$ and $k^{st} = 3$ in this example).

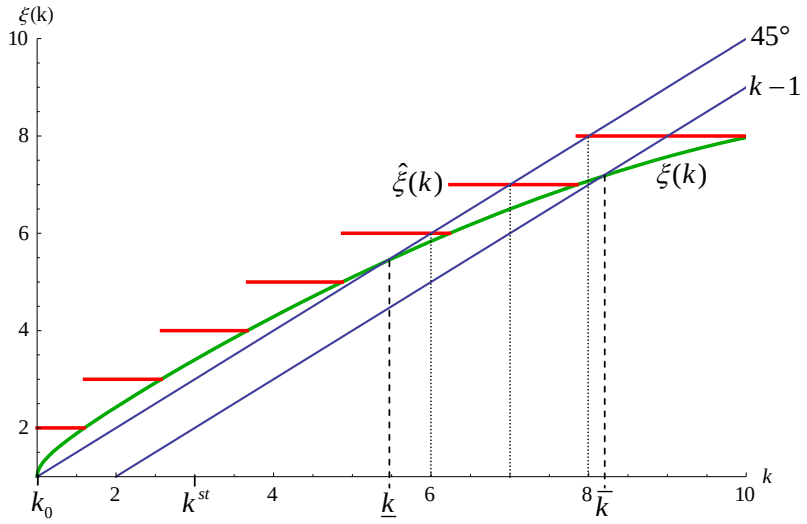

Figure 2: Illustration of $\xi(k)$ and $\hat{\xi}(k)$ for Example 2 (Section 4) with $\delta = 0.6$, $N = 10$

The multiplicity of equilibria in our model can be explained intuitively: If coalition members in period t are optimistic and anticipate 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 $\hat{\xi}(k^*)$ 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 delay, then also the critical coalition size $\hat{\xi}(k^*)$ is smaller and an agreement is signed immediately by fewer countries.

We now provide comparative statics with respect to the discount factor δ and the per-period payoff when no agreement is signed, π_0. Due to multiplicity of equilibria, we consider comparative statics on $\underline{k}$. Recall that Π_s and Π_n denote present values of payoffs. Until now we did not impose any intertemporal structure on the interaction of countries after an agreement is signed, which gives rise to these present values. In order to provide comparative statics results, more structure is needed, and we simply assume that these

present values are derived from time-independent interaction, so that

$$\pi_s(k) = (1-\delta)\Pi_s(k) \qquad \text{and} \qquad \pi_n(k) = (1-\delta)\Pi_n(k) \tag{12}$$

reflect the (constant) per-period payoffs once an agreement among k countries has been signed. Clearly, the internal and external stability conditions (first two conditions in (10) resp. (9)) can now be formulated for the per-period payoffs $\pi_s(k)$ and $\pi_n(k)$, that do not depend on δ nor on π_0. Thus, also the value of k^{st} does not depend on δ nor on π_0. In conjunction with Proposition 2, the next result thus indicates that in our dynamic climate cooperation model, an increase in δ or in π_0 leads to (weakly) larger equilibrium coalition sizes.

Proposition 4. *Assume that the per-period payoffs after signing an agreement are constant over time. If $\underline{k} > k_0$, then the value of $\underline{k}$ is increasing in δ and in π_0.*

To see the intuition, consider an equilibrium coalition size k^* (such that $k^* > k^{st}$). An increase in δ or in π_0 makes countries *less* eager to sign an agreement in period t and leads to a larger equilibrium coalition size. This is because in both cases, the outside option (i.e., when coalition members in period t do not sign an agreement) becomes more attractive, which leads to a larger endogenous threshold for the minimum size of an active coalition. Clearly, when π_0 increases, then not signing a long-term agreement in some period becomes less costly. This makes countries less eager to reach a long-term agreement in any period. At the same time, it implies that the critical coalition size that must be reached for countries to be willing to sign a long-term agreement increases. Consequently, the value of $\underline{k}$ increases in π_0. Similarly, when the discount factor increases, a delay in climate negotiations is relatively less costly because most of the benefits from cooperation are incurred in the future. Hence, again the countries are less eager to sign a long-term agreement in any period, and the value of $\underline{k}$ increases in δ.

The best way to sharpen our intuition for this model is to look at specific examples.

4 Examples

In this section we consider examples where we model the interaction that gives rise to the payoff functions Π_s and Π_n. Similarly as in Proposition 4, we consider the case where the present values Π_s and Π_n are outcomes of time-independent (and myopic) interaction (see (12)). In addition, we assume that in periods without an agreement, the countries obtain the non-cooperative equilibrium payoff $\pi_0 = \pi_n(0) = (1-\delta)\Pi_n(0)$.[31]

All examples that we consider in this section share some basic properties (described in the following). We refer to emission games that have these properties as *simple emission*

[31] In Supplementary Appendix B.1 (Lemma 5) we show that for the class of games considered in this section, we obtain $\pi_n(0) = \pi_n(k_0)$. This means that π_0 attains its lower bound from Assumption 4.

games. In such games, a country's payoff can be expressed in a simple benefit-cost form

$$B(X) - C(x_i),$$

where X denotes the aggregate abatement and x_i denotes the country i's abatement level. This captures the idea that all countries benefit equally from the overall abatement but each country bears the costs of its own abatement efforts. Furthermore, let $B' > 0$, $B'' \leq 0$, $C' > 0$, and $C'' > 0$, i.e., the costs and benefits are increasing but benefits are (weakly) concave in the aggregate abatement level, while the costs are increasing and convex in the country's abatement level. We also assume that the coalition acts as a Stackelberg leader vis-à-vis the non-signatories. This is a plausible assumption, as signatories commit themselves to long-term abatement targets while non-signatories choose their efforts on a short-term basis in all periods. We consider symmetric equilibria, where all signatories choose the same abatement level, denoted x_s, and also all non-signatories choose an identical abatement level, denoted x_n.

The coalition (with size k) chooses the abatement level x_s in order to maximize the coalition's joint welfare $k[B(X) - C(x_s)]$, where $X = kx_s + x_{n,1} + \cdots + x_{n,N-k}$. As to the behavior of non-signatories, we consider two possibilities. In Example 1, we assume that they are non-strategic and do not reduce their emissions at all, i.e., $x_{n,i} = 0$. In Examples 2 and 4, we assume that they are strategic and choose their abatement levels in order to maximize the welfare $B(X) - C(x_{n,i}) = B(x_{n,i} + X_{-i}) - C(x_{n,i})$, where $X_{-i} = X - x_{n,i}$ denotes the aggregate abatement level of all other countries. It follows from the above assumptions that this welfare is strictly concave in $x_{n,i}$. The first-order condition then becomes $B'(x_{n,i} + X_{-i}) = C'(x_{n,i})$, which under symmetry yields

$$B'(kx_s + (N-k)x_n) = C'(x_n). \tag{13}$$

Note that, even though the above considerations apply only for integer values of k, we can extend the equilibrium welfare functions also to non-integer values of k (by using condition (13) and the corresponding solution to the coalition's maximization problem). We provide further general results for simple emission games in Supplementary Appendix B.1 (Lemma 5).

Let us now turn to the specific examples. We first focus on a simple emissions game with linear benefits and quadratic costs of abatement, that has often been considered in the literature. In Example 4, we allow for concave benefits of abatement.

Example 1: Linear-quadratic example (basic case)

Suppose, in each period there is a constant marginal benefit of abatement $b > 0$. In this case, the benefit function is linear and has the form $B(X) = bX$. Moreover, assume that

the costs are quadratic and have the form $C(x_i) = \frac{1}{2}cx_i^2$ (where $c > 0$).

In the basic example we assume for simplicity that non-signatories do not regulate their emissions, i.e., $x_n = 0$ and $\pi_0 = 0$. Relaxing this assumption will lead us to our next Example 2 (below). The coalition then chooses the abatement level x_s to maximize the aggregate coalition payoff $k[B(X) - C(x_s)] = k(bkx_s - \frac{1}{2}cx_s^2)$. This yields the coalition's optimal abatement per signatory $x_s^*(k) = bk/c$ (in all periods once an agreement is signed) and the following welfare functions:

$$\pi_s(k) = \frac{b^2k^2}{2c} \qquad \text{and} \qquad \pi_n(k) = \frac{b^2k^2}{c}. \tag{14}$$

The discounted payoffs Π_s and Π_n are then given by (12). Note, that in this example, each non-signatory obtains a payoff that is twice that of a signatory: $\Pi_n(k)/\Pi_s(k) = \pi_n(k)/\pi_s(k) = 2$ for any $k > 0$. This nicely illustrates the free-rider incentives in this model.[32] For the above payoff functions it is easy to verify that $\tilde{k} = 1 + \sqrt{2}$ (see Assumption 3). Thus, in the static case we obtain the pessimistic result that only a coalition with $k^{st} = 3$ countries is stable.

In the dynamic game, it follows from the expressions in (14) and from $\pi_0 = 0$ that $\xi(k) = k\sqrt{\delta\,(2 - k/N)}$.[33] Now recall that $\underline{k}$ is a fixed point of the function ξ, i.e., $\underline{k} = \xi(\underline{k})$, if such exists. This is the case when $\delta \geq \frac{1}{2}$, which then yields

$$\frac{\underline{k}}{N} = 2 - \frac{1}{\delta}. \tag{15}$$

On the other hand, if $\delta < \frac{1}{2}$, then $\xi(k) < k$ for all $k > k_0 = 0$, and thus $\underline{k} = 0$.

The simple condition (15) nicely captures the central result of this paper: For sufficiently large values of δ (and N), the equilibrium coalition size is strictly larger than in the static model, and for δ close to 1, the grand coalition (i.e., $k^* = N$) is stable. Unlike in the static model, this holds even when the gains from cooperation are large. Only if the discount factor is close to or below $\frac{1}{2}$, the equilibrium coalition size is small. In this case, the coalition size is determined by conditions (9), rather than (10) and the dynamic model leads to identical results as the static one. Notice that in this example, the lower bound ($\underline{k}$) for the size of a stable coalition (for $\delta \geq \frac{1}{2}$), (15), is *independent* of the benefit and cost parameters b and c.

Example 2: Linear-quadratic example (standard case)

We now extend the previous linear-quadratic example by relaxing the (non-standard) assumption that non-signatories do not regulate their emissions. This assumption was made for simplicity, and simplified the algebra considerably. The standard case consid-

[32] For $k = 0$ we have $\Pi_s(0) = \Pi_n(0) = \pi_s(0) = \pi_n(0) = 0$. Thus, also $k_0 = 0$.

[33] Recall that $\pi_s(\xi(k)) = (1 - \delta)\pi_0 + \delta v(k)$, where $v(k) \equiv k/N \cdot \pi_s(k) + (1 - k/N)\pi_n(k)$.

ered in the literature is where all countries choose some positive abatement efforts. In equilibrium, each non-signatory chooses an abatement level that satisfies the first-order condition, which now becomes $b = cx_n$. Thus, $x_n^* = b/c$.[34] The coalition then chooses the abatement level x_s that maximizes $k[B(X) - C(x_s)]$, where $X = kx_s + (N-k)x_n$. It follows from the first-order condition that $x_s^* = kb/c$.

This yields the following per period payoffs (see also Barrett 2005):

$$\pi_s(k) = \frac{b^2}{2c}[k^2 + 2(N-k)] \quad \text{and} \quad \pi_n(k) = \frac{b^2}{2c}[2k^2 + 2(N-k) - 1]. \tag{16}$$

The payoff per country in a period without an agreement equals the non-cooperative payoff $\pi_0 = \pi_n(0) = b^2/(2c) \cdot (2N-1)$.[35] Then we obtain $\tilde{k} = 2$, an integer. Thus, in the static model we obtain equilibrium coalitions with 2 or 3 countries.[36]

Using (16) and (7), we obtain $\xi(k) = 1 + \sqrt{\delta k(k-1)[2-(k+1)/N]}$, and can determine $\underline{k}$ to obtain for the dynamic model:

$$\frac{\underline{k}}{N} = 1 - \frac{1}{2\delta} - \frac{1}{2N} + \sqrt{\frac{1}{\delta N} + \left(1 - \frac{1}{2\delta} - \frac{1}{2N}\right)^2}. \tag{17}$$

It is easy to show that for large values of δ, the values of $\underline{k}$ defined by (15) and (17) are close (see Figure 3 for an illustration). Thus, the simple formula (15) for the lower bound on the size of the stable coalition in the basic Example 1 where only signatories abate delivers a good *approximation* for the respective value in the standard case where all countries abate. As indicated in the intuition below Proposition 4, the equilibrium coalition size is rather large when the countries are patient (δ is large). As can be seen from (16), for large k, the payoff ratio $\Pi_n(k)/\Pi_s(k) = \pi_n(k)/\pi_s(k)$ is then close to 2, while this ratio is exactly 2 in the basic Example 1. This observation leads us to a straightforward generalization of the linear-quadratic example presented in the following.

Example 3: Generalized example with linear benefits of abatement

Now we generalize the basic Example 1 in another direction. Recall that in the basic example it is the case that the payoff ratio $\Pi_n(k)/\Pi_s(k)$ is constant and equal to 2 (when $k > 0$). Let us now consider the case where this ratio is constant but equal to some value $\alpha > 1$, i.e.,

$$\frac{\Pi_n(k)}{\Pi_s(k)} = \frac{\pi_n(k)}{\pi_s(k)} = \alpha \tag{18}$$

[34] Observe that due to the assumption of linear benefits, this abatement level is a dominant strategy for each non-signatory. Thus, we obtain identical equilibrium abatement levels also in a model where all countries choose their abatement levels simultaneously.

[35] Note, that now we have $\Pi_s(1) = \Pi_n(1) = \Pi_n(0)$. This also implies that $k_0 = 1$.

[36] Recall that for convenience we have assumed in the general analysis that $\tilde{k}$ is not an integer. This assumption is not crucial, it only facilitates the formulation of our results.

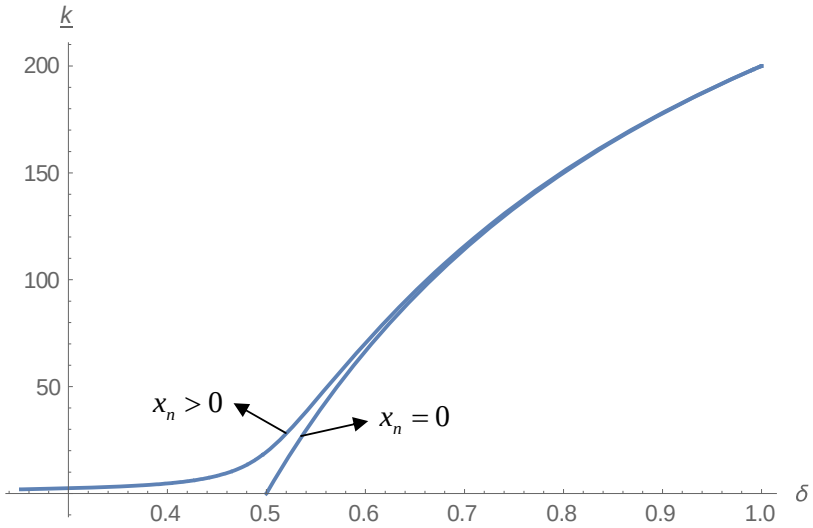

Figure 3: $\underline{k}$ as function of δ (for $N = 200$), in Example 1 ($x_n = 0$) and Example 2 ($x_n > 0$)

for all $k > 0$. Furthermore, let $\pi_0 = \pi_n(0) = 0$ as in the basic example.[37] The parameter α captures 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.[38] A fixed ratio α as in (18), can be obtained from the benefit-cost emission game with linear benefit function $B(X) = bX$ and cost function of the form $C(x_i) = cx_i^\gamma/\gamma$ (where $b, c > 0$ and $\gamma > 1$). In this case, $\alpha = \gamma/(\gamma - 1)$.[39] Hence, the free-rider incentives are more intense the closer the parameter γ is to 1. The case of quadratic costs from Example 1 is obtained for $\gamma = 2$.

In this example we cannot compute the function ξ explicitly. However, we can evaluate $\underline{k}$ as solution of the equation $\pi_s(k) = (1-\delta)\pi_0 + \delta v(k)$. After dividing by $\pi_s(k)$ and using the assumption $\pi_n(k)/\pi_s(k) = \alpha$, this equation becomes $1 = \delta[k/N + (1-k/N)\alpha]$. Solving the equation yields

$$\frac{\underline{k}}{N} = \frac{\alpha - 1/\delta}{\alpha - 1}, \tag{19}$$

when $\alpha\delta \geq 1$, while $\underline{k} = 0$ when $\alpha\delta < 1$. This generalizes the result (15).

Intuitively, if the ratio $\alpha = \pi_n(k)/\pi_s(k)$ is large, then it is very profitable to be a non-signatory when an agreement is signed. Hence, the first of the two basic motives mentioned earlier (which requires a large coalition in order for its members to sign a long-term agreement) is strong, so that only large coalitions are stable. Recall that with k^* large, the probability to become a non-signatory in the next period in case of a delay,

[37] 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.

[38] It is easy to show that in the static case, only a small coalition is stable when α is large.

[39] A straightforward computation yields the per period payoffs $\pi_s(k) = (\gamma - 1)/\gamma \cdot [(bk)^\gamma/c]^{1/(\gamma-1)}$ and $\pi_n(k) = [(bk)^\gamma/c]^{1/(\gamma-1)}$.

$1 - p(k^*) = (N - k^*)/N$, is small. This undermines the free-rider incentive and explains why a large coalition can indeed form in equilibrium.

Example 4: Concave benefits of abatement

Now we briefly consider a more complex example, where the benefits of abatement are concave in the overall abatement X. More specifically, we consider quadratic benefit function $B(X) = bX - \frac{1}{2}dX^2$ and quadratic cost function $C(x_i) = \frac{1}{2}cx_i^2$ (where $b, c > 0$ and $d \geq 0$).

Using the first-order condition for non-signatories and maximizing the coalition's welfare as a Stackelberg leader, we obtain the following abatement levels (see Supplementary Appendix B.1 for details):

$$x_n = \frac{b - dkx_s}{c + (N-k)d} \quad \text{and} \quad x_s = \frac{bck}{c^2 + d^2(N-k)^2 + cd(k^2 + 2(N-k))}. \tag{20}$$

Inserting these results back into the payoff functions, we can compute the countries' welfare as a function of the coalition size. For simplicity, we do not provide the full formulas here. It is straightforward to show that the equality $\pi_s(k_0) = \pi_n(k_0)$ yields $x_s = x_n$ and thus

$$k_0 = \frac{c + dN}{c + d}.$$

Moreover, the non-cooperative abatement level, $x_n(0) = b/(c + Nd)$, is directly obtained from (20) by setting $k = 0$.

Observe that $k_0 = 1$ for $d = 0$, i.e., in the case with linear benefits (see Example 2), while $k_0 > 1$ if $d > 0$. Intuitively, a small coalition strategically reduces the abatement efforts of its members in order to induce non-signatories to raise their efforts, thereby exploiting the first-mover advantage. This effect was absent in the examples with linear benefits of abatement and explains why a larger coalition size (greater than 1) is required to induce coalition members to internalize environmental externalities between them by reducing their emissions (by more than the non-signatories).

We do not seek to provide a general characterization of equilibrium outcomes in this example. Instead, we merely want to check if qualitatively similar results are obtained as in our previous examples. In particular, we want to investigate if the function ξ fulfills the basic properties that we assumed in Section 3. Due to the algebraic complexity of the involved functions, in particular $\pi_s(k)$, $\pi_n(k)$, and $\xi(k)$, the latter of which involves higher-order polynomials (not presented here), we content ourselves with a simple numerical inspection of these functions.

Figure 4 illustrates the shape of the payoff functions $\pi_s(k)$ and $\pi_n(k)$ in this example for a specific set of parameter values. Varying these parameter values, the basic properties of the payoff functions remain (not shown), while scales of course differ. We observe that

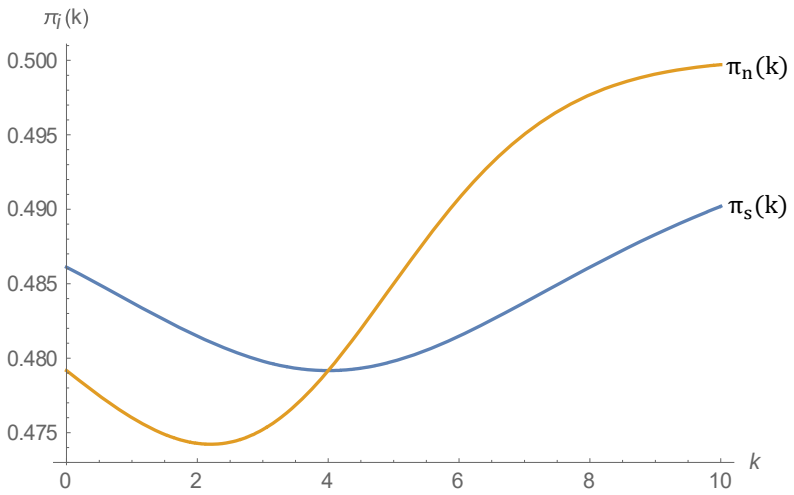

Figure 4: Payoffs as function of k for Example 4 with concave benefits of abatement (for $N = 10$, $b = 1$, $c = 2$, and $d = 1$)

in this example $\pi_s(k) > \pi_n(k)$ for all $k < k_0 = 4$, and it is easy to verify that these functions fulfill Assumption 3(b). Hence, there is no stable coalition size below k_0. Also our Assumption 1 is obviously fulfilled. Only Assumption 3(a) is not fulfilled for large values of k (the free-rider incentive $\pi_n(k) - \pi_s(k+1)$ is declining for large k), but this is inconsequential because $\pi_n(k)$ is significantly larger than $\pi_s(k)$ in this range so that the external stability condition remains fulfilled for all $k \geq k_0$. Also the function ξ fulfills our assumed properties. In particular, we find that it has a simple concave shape for $k \geq k_0$ (not shown), so that ξ and $\xi + 1$ obviously have at most one fixed point above k_0, and at a fixed point, ξ crosses the 45°-line from above. Now we conclude that our simple characterization of equilibrium outcomes from Section 3 can be applied also to this more complex example. In this example (for the parameter values underlying Figure 4), we obtain $\tilde{k} \approx 4.16$, and thus the stable coalition size in the static model is $k^{st} = 5$. In the dynamic model we obtain for $\delta = 0.9$: $\underline{k} \approx 9.04$ and $\bar{k} = 11$ (since $\xi(10) \approx 9.69$ and $\xi(10) + 1 > 10$), so that the only equilibrium coalition size for these parameter values is $k^* = 10$ (i.e., the grand coalition).

5 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 (yielding a payoff of π_0), 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 allow for this possibility. We maintain our earlier assumption that countries'

payoffs under a long-term agreement are derived from time-independent interaction with per-period payoffs as given by (12).

There are several ways how negotiations about short-term agreements could be modeled. These negotiations start if at least one coalition member (out of k_t members) uses its veto right to block the negotiations about a long-term agreement in period t. One possibility is to assume that the coalition dissolves, and new negotiations about a short-term agreement start within period t. Another approach is to assume that the coalition remains intact, and the remaining coalition members decide whether or not to sign a short-term agreement (instead of a long-term agreement).[40] We find both approaches somewhat extreme. The first approach may not be plausible because countries that are already in a coalition may be perceived as more likely candidates for a short-term agreement than outsiders.[41] The second approach is extreme in the sense that countries are *locked-in* inside the coalition, even if some of them would prefer to exit the negotiations about a short-term agreement. This is indeed the case whenever $k_t > k^{st}$, because the internal stability condition, (IS), is then violated. Furthermore, as Barrett (2005) points out, "*under the rules of international law, countries are free to participate in treaties or not as they please*". Hence, it seems implausible to assume that a country that has joined the coalition in period t can only sign a short-term agreement (once negotiations about a long-term agreement have failed in that period), or use its veto right to block any short-term agreement.

To address these problems, we assume that a country that has joined the coalition in period t can withdraw from the negotiations if it wishes to do so. Hence, to model short-term agreements, we add two additional stages to the negotiations within period t that parallel our modeling of negotiations about the long-term agreement (see Figure 1). First, if a long-term agreement is not signed by the (initial) k_t coalition members, each of them has the possibility to stay in the coalition that starts to negotiate about a short-term agreement, or to exit the coalition. Second, the remaining coalition members then decide whether to sign a short-term agreement that lasts only for one period (until the end of period t), or not to sign any agreement at all in this period.[42]

As a first observation, note that for any $k_t > k_0$, a short-term agreement is signed in period t (provided that this period is reached and that no long-term agreement is signed) since $\pi_s(k_t) > \pi_s(k_0) = \pi_n(0)$. By our earlier assumptions, countries' decisions whether or not to sign a short-term agreement have no direct impact upon their payoffs

[40] We comment on this approach in more detail in Section 6.1, where we also compare our approach with the one chosen by Battaglini and Harstad (2016).

[41] If new negotiations about a short-term agreement start in period t, all countries are equally likely to become members of the short-term coalition, irrespective of the composition of the initial coalition.

[42] A short-term agreement is signed if none of the remaining coalition members uses its veto right. Abandoning this stage would not change the results. Note, that exiting the (initial) coalition always delivers welfare that is at least as large as staying inside and blocking a short-term agreement (since outsiders benefit more from an agreement than insiders).

in future periods. Thus, under the Markov restriction there is no impact upon countries' continuation values either. Hence, if $k_t > k_0$, countries are better off by signing a short-term agreement in period t, rather than to sign no agreement at all. If $k_t \geq k^{st}$, the negotiations about a short-term agreement and the resulting payoffs within period t, thus, bring us back to the static game that (see Section 3). Hence, due to our assumption that countries can withdraw from the negotiations about a short-term agreement, the coalition size of any short-term agreement is equal to k^{st} whenever $k_t \geq k^{st}$. If $k_0 \leq k_t < k^{st}$, the coalition size in a short-term agreement is k_t, while no short-term agreement (or a short-term agreement with k_t members) is signed if $k_t < k_0$.[43]

Let us now proceed with the equilibrium analysis of the full dynamic game. We first ask whether equilibria that entail a long-term agreement signed by (at most) k^{st} countries can exist also when countries have the possibility to sign a short-term agreement. If $k_0 \leq k_t \leq k^{st}$, the coalition signs a long-term agreement in period t if[44]

$$\Pi_s(k_t) \geq \pi_s(k_t) + \delta V(k_{t+1}).$$

Since $\pi_s(k_t) = (1-\delta)\Pi_s(k_t)$, the condition simplifies to

$$\Pi_s(k_t) \geq V(k_{t+1}).$$

This replaces our earlier condition (6). However, for $k_{t+1} = k_t = k^*$ and $k_0 < k^* \leq k^{st}$, this condition is never satisfied, because $V(k^*)$ is a convex combination of $\Pi_s(k^*)$ and $\Pi_n(k^*)$, and $\Pi_n(k^*) > \Pi_s(k^*)$ due to Assumption 1. Intuitively, if a coalition of size k^{st} (or smaller) forms in period t, and countries expect the formation of another coalition of size k^{st} in the next period (provided that no long-term agreement is signed today), then the coalition members prefer to sign only a short-term agreement today. This way, they enjoy the benefits of free-riding in future periods with a positive probability, while the coalition size in the future is not smaller than it is today. Hence, an equilibrium where a coalition of size k^{st} signs a long-term agreement does not exist when countries have the possibility to sign short-term agreements.[45]

On the other hand, the above intuition suggests that there might be an equilibrium, where a coalition forming in each period fails to sign a long-term agreement, but signs only a short-term agreement instead. Such an equilibrium yields the same joint welfare

[43] If $k_t < k_0$ and $k_0 > 1$ (see Eample 4 in Section 4), then the k_t coalition members sign a short-term agreement.

[44] Note that since all $k_t \leq k^{st}$ satisfy the internal stability condition (IS), no coalition member has an incentive to withdraw from the negotiations when $k_t \leq k^{st}$.

[45] Note, that there is also no equilibrium where a coalition of size $k^* < k_0$ signs a long-term contract, as (due to Assumption 3(b)) additional countries would have an incentive to join. This holds also when the decision of an outsider to join may induce coalition members to sign only a short-term agreement, as follows from Assumption 1.

as a long-term contract signed by k^{st} countries and from an *ex ante* perspective yields the same expected welfare to each country.

Proposition 5. *Under the possibility to sign short-term agreements, there is no equilibrium where at most k^{st} countries sign a long-term agreement. There is an equilibrium where a coalition of size k^{st} forms in each period and signs only a short-term agreement, if and only if*

$$\Pi_s(k^{st}+1) \leq \pi_n(k^{st}) + \delta V(k^{st}). \tag{21}$$

There is no other equilibrium where a coalition signs a short-term agreement in each period.

Intuitively, if (21) is satisfied, then an equilibrium exists with k^{st} countries joining the coalition in each period to sign a short-term agreement, because there is no incentive for an additional country to join, even if the new coalition of $k^{st}+1$ countries signs a long-term agreement. However, the equilibrium fails to exist if (21) does not hold. In this case, there is an incentive for an additional country, that is assigned to become a non-member in period t, to join. Although this reduces the payoff of this country in the current period (from $\pi_n(k^{st})$ to $\pi_s(k^{st}+1)$), the continuation value from the next period onwards may be increased: The per-period payoff after the deviation (equal to $\pi_s(k^{st}+1)$) is higher in those periods where (without the deviation) the country would be assigned as a coalition member (yielding payoff $\pi_s(k^{st})$). Hence, there is an incentive to join in order to lock the other k^{st} coalition members into a long-term climate contract, inducing them to raise their abatement efforts. If this incentive is sufficiently strong, it outweighs the free-rider incentives of this country. An equilibrium with k^{st} countries signing a short-term agreement in each period then fails to exist.

Now consider equilibria with coalition sizes strictly larger than k^{st}. Recall that, without the possibility to sign short-term agreements, the stable coalition size in such an equilibrium must be a fixed point of the function $\hat{\xi}$. We show below that (with a small modification) the characterization of these equilibria remains the same. Formally, for $k_t > k^{st}$, the coalition in period t signs a long-term agreement if no country uses its veto power in order to free-ride on k^{st} other countries signing a short-term agreement:

$$\Pi_s(k_t) \geq \pi_n(k^{st}) + \delta V(k_{t+1}). \tag{22}$$

This replaces our earlier condition (6) for coalition sizes *greater* than k^{st}. Hence, we can define the function ξ by condition (7) as before, if we replace π_0 by $\pi_n(k^{st}) = (1-\delta)\Pi_n(k^{st})$. With this modification, the result of Proposition 2 remains valid, so any equilibrium coalition size larger than k^{st} is a fixed point of the function $\hat{\xi}$. Thus, we can characterize the set of all equilibrium coalition sizes using the points $\underline{k}$ and $\bar{k}$.

Proposition 6. *Under the possibility to sign short-term agreements, the set of all equilibrium coalition sizes that sign a long-term agreement are all integers from interval* $[\underline{k}, \bar{k})$. *Moreover,* $\underline{k} > k^{st} + 1$.

The proposition parallels Proposition 3 that provides a characterization of equilibrium coalition sizes. However, here we only obtain case (a) where $\underline{k} > k^{st}$. In addition, it holds that $\underline{k} > k^{st} + 1$, so the lowest possible equilibrium coalition size (for a long-term agreement) is $k^{st}+2$. Intuitively, there cannot be an equilibrium with coalition size $k^{st}+1$, because out of the $k^{st}+1$ countries, an individual coalition member would always prefer to drop out of the coalition. It does not lose by this: Its own participation choice only affects the duration of the agreement reached in period t, but not the remaining number of (other) countries that sign the agreement (in this period and from the next period onwards). The situation is fundamentally different with at least $k^{st}+2$ countries. If one country assigned as coalition member stays outside or blocks a long-term agreement, then at least one other country will also drop out of the coalition before the negotiations about a short-term agreement start. This gives the first country an additional incentive to stay in the coalition and to sign a long-term agreement. This explains why coalitions with larger participation levels that sign a long-term agreement can occur in equilibrium also in this version of the model.

Let us finally compare the equilibrium coalition sizes in the dynamic model with and without the possibility to sign short-term agreements. As we have shown above, introducing short-term agreements corresponds to an increase in the parameter π_0 (from $\pi_n(k_0)$ to $\pi_n(k^{st})$), that captures the payoff in a period without a long-term agreement. Thus, it follows directly from Proposition 4 that the possibility to sign short-term agreements has a *stabilizing* effect upon long-term cooperation.[46]

6 Deterministic membership approach

So far we have assumed that the identity of the countries that become members of the coalition in period t (for some given coalition size k_t) is determined randomly. However, there is an alternative approach that is used in the literature, and there seems to be no consensus 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. Under the alternative approach, the countries have persistent identities. For any coalition size k_t, the identity of the coalition members is then pre-determined and commonly known (see Battaglini and Harstad 2016, among others). From a theoretical perspective, these

[46]For instance, in the Example 2, Section 4, given the possibility to sign short-term agreements, we obtain for the parameter values that are underlying Figure 2 (i.e., $\delta = 0.6$ and $N = 10$): $\underline{k} \approx 6.8$ and $\bar{k} \approx 8.9$, so that the equilibrium coalition sizes are $k^* = 7$ and $k^* = 8$, whereas without short-term agreements we had $\underline{k} \approx 5.4$ and $\bar{k} \approx 8.2$, so that $k^* = 6$ was also an equilibrium coalition size.

identities may be simply selected randomly at the beginning of the game.[47] From an applied perspective, they may reflect countries' (known) willingness to cooperate (or reputation) in climate-related issues. 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 country 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 will (in equilibrium) be part of the coalition and which countries will be the outsiders. More specifically, if we denote the countries as $1, 2, \ldots, N$, then for a coalition of size k_t: countries $1, 2, \ldots, k_t$ will become members, while countries $k_t + 1, \ldots, N$ will not. In this section we investigate how our previous results change under this alternative approach.

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 specification of the probability to be re-assigned as a coalition member in the next period in case of a delay, for a country that is assigned to become a coalition member today. Under the deterministic membership approach, this probability is $p(k^*) = 1$ (instead of $p(k^*) = k^*/N$) and it follows from (5) that $V(k) = \Pi_s(k)$. The condition (7) then simplifies to

$$\Pi_s(\xi(k)) = \pi_0 + \delta\Pi_s(k), \tag{23}$$

where π_0 is again treated as an independent parameter, bounded by Assumption 4 (as in Section 3). Then $\underline{k}$, which is a fixed point of the function ξ, satisfies

$$\Pi_s(\underline{k}) = \frac{\pi_0}{1-\delta}. \tag{24}$$

In contrast to the random membership case, now we don't need to impose the single-crossing property in Assumption 5(a) on function ξ; it follows from Assumption 4 and the monotonicity of Π_s. However, we still impose the single-crossing property on function $\xi+1$ from Assumption 5(b). Then $\bar{k}$ satisfies $\Pi_s(\bar{k}-1) - \delta\Pi_s(\bar{k}) = \pi_0$ if $\Pi_s(N-1) - \delta\Pi_s(N) \leq \pi_0$, while $\bar{k} = N$ otherwise.

Under these modifications, the analysis remains the same, and as we show in Appendix A.2, our general results from Section 3, in particular Lemma 1 and Propositions 1–4 remain valid. Moreover, due to Assumption 4, property (24) implies that $\underline{k} \leq k^{st}$. This implies that k^{st} is always an equilibrium coalition size and only the cases (b) and (c) in Proposition 3 apply.

[47] The difference to the random membership approach is, then, only that these identities remain unchanged later on.

Finally, note that equation (24) 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, recall that, for k^{st} small, $\underline{k}$ is only the lower bound on the stable coalition size. The upper bound, the fixed point of function $\xi + 1$, may still be large. To illustrate this point, let us review our Examples 1–3 from Section 4.

Example 1'. Consider first the basic linear-quadratic Example 1 from Section 4, where non-signatories do not regulate their emissions (and thus $\pi_0 = 0$). Inserting the payoff function $\Pi_s(k)$ as given by (14) into (23), we find that $\xi(k) = \sqrt{\delta}\, k$. The function ξ is linear and it follows that $\underline{k} = 0$ and $\bar{k} = \min\left\{1/(1-\sqrt{\delta}),\, N\right\}$. Thus, the set of stable coalition sizes (fixed points of $\hat{\xi}$) is large when δ is close to 1 (recall that $k^{st} = 3$ remains the same).

Example 2'. It is easy to verify that this result also translates to the extended Example 2, where all countries may regulate their emissions. For the payoffs given by (16), we obtain $\xi(k) = 1 - \sqrt{\delta} + \sqrt{\delta}\, k$. Again, the function $\xi(k)$ is linear, and we have $\underline{k} = 1$ and $\bar{k} = \min\left\{1 + 1/(1-\sqrt{\delta}),\, N\right\}$. We then again obtain that large coalitions are stable when δ is large. Figure 5 illustrates this property (for $N = 10$ and $\delta = 0.8$). For those parameter values, we find that *all* integer values from 2 to 10 represent equilibrium coalition sizes (recall that in the static model there are two equilibria: $k^{st} = 2$ and $k^{st} = 3$).

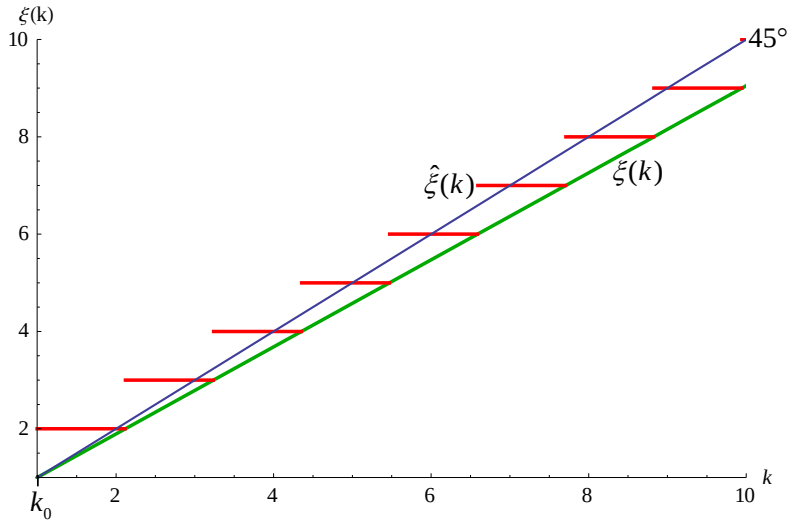

Figure 5: $\xi(k)$ and $\hat{\xi}(k)$, for $\delta = 0.8$ and $N = 10$, deterministic membership case (example with $x_n > 0$)

Example 3'. Finally, consider the generalized Example 3, with linear benefits and cost function $C(x_i) = cx_i^{\gamma}/\gamma$ (where $\gamma > 1$ and $\alpha = \gamma/(\gamma - 1)$). In that case, we obtain $\underline{k} = 0$ and $\bar{k} = \min\{1/(1-\delta^{1/\alpha}),\, N\}$. Hence, larger coalition sizes can be supported

in equilibrium when the free-rider incentive increases, which mirrors our results from Section 4 (Example 3).

Despite these similarities in the results for the two cases (random and deterministic membership), the underlying intuition is quite different. Recall that under random membership, large coalitions are stable (for δ sufficiently large) because with k^* large, the probability to become a non-signatory in the next period in case of a delay is small. This undermines the free-rider incentive. On the contrary, under deterministic membership, large coalitions are stable because of a self-fulfilling prophecy: When countries are optimistic and expect the formation of a large coalition in the future in case negotiations fail today, then the current coalition must be sufficiently large, too, in order to induce members to sign a long-term agreement already today. This feedback effect also explains the larger degree of multiplicity of equilibria under deterministic membership (see Figure 5), as compared to the random membership case (see Figure 2).

6.1 Short-term agreements under deterministic membership

As we have argued in Section 5, a short-term agreement is always signed in a period where the negotiations about a long-term agreement have failed. Given Assumption 4, the number of countries signing a short-term agreement is then equal to k^{st} if $k_t \geq k^{st}$ countries have joined the coalition in period t, is equal to k_t when $k_0 \leq k_t < k^{st}$, and no short-term agreement (or an agreement with k_t members) is signed when $k_t < k_0$. These findings remain unaffected under the deterministic membership approach. To see this, observe that for $k_t > k^{st}$, the coalition in period t signs a long-term agreement if

$$\Pi_s(k_t) \geq \pi_n(k^{st}) + \delta\Pi_s(k_{t+1}).$$

This condition is analogous to the condition (22) when $V(k_{t+1}) = \Pi_s(k_{t+1})$. Hence, again when replacing π_0 by $\pi_n(k^{st})$ as in Section 5, the definition of the function ξ, (23), and the corresponding equilibrium conditions remain valid.

By analogous arguments as in Proposition 5 for the random membership case, we again obtain that there is an equilibrium where a coalition of size k^{st} signs a short-term agreement in each period, if and only if

$$\Pi_s(k^{st}+1) \leq \pi_n(k^{st}) + \delta\Pi_s(k^{st}). \tag{25}$$

However, in contrast to Proposition 5, under the deterministic membership approach, there is also an equilibrium where k^{st} countries sign a *long-term* agreement when (25) holds. To see this, note that by the stationarity in the assignment of countries as signatories and non-signatories under deterministic membership, the identity of the k^{st} countries

that would sign a short-term agreement in each period would stay the same. Hence, these countries may as well sign a long-term agreement immediately.

Any equilibrium with $k^* > k^{st}$ is (in this version of the model), thus, characterized by the dynamic stability condition, (DS) with $\pi_0 = \pi_n(k^{st})$. All that remains to be done in order to characterize the effect of the possibility to sign short-term agreements on the set of equilibria (for simple emission games) is, thus, to study the effect of an increase in π_0 upon $\underline{k}$ and $\bar{k}$. It is straightforward to see from (24) that $\underline{k}$ increases in π_0. Now consider the effect on the upper bound on the stable coalition size, $\bar{k}$. Recall that $\bar{k}$ now satisfies $\Pi_s(\bar{k}-1) - \delta\Pi_s(\bar{k}) = \pi_n(k^{st})$. It follows from the single-crossing Assumption 5(b) that $\Pi_s(k-1) - \delta\Pi_s(k)$ is increasing at $k = \bar{k}$. Consequently, also $\bar{k}$ is increasing in π_0 and, thus, introducing the possibility to sign short-term agreements increases the values of $\underline{k}$ and $\bar{k}$.

Hence, as in the case with random membership, also under deterministic membership we find that the possibility to sign short-term agreements tends to have a stabilizing effect upon long-term agreements. This is especially true for smaller values of δ. For instance, in the Example 2 in Section 4, for $\delta = 0.5$ we obtain $\underline{k} \approx 4.7$ and $\bar{k} \approx 6.7$, yielding equilibrium coalition sizes $k^* = 5$ and $k^* = 6$, with the possibility to sign short-term agreements, while $\underline{k} = 1$ and $\bar{k} \approx 4.4$, yielding equilibrium coalition sizes $k^* = 3$ and $k^* = 4$, without it (recall that $k^{st} = 3$ in that example). For larger values of δ, the impact on the set of equilibrium coalition sizes (in particular on $\bar{k}$) is less pronounced (not shown here).

Finally, since the model analyzed in this subsection is close to the setup used by Battaglini and Harstad (2016), let us briefly discuss their relation. Battaglini and Harstad (2016) also allow for short-term agreements, and use a deterministic membership approach. Apart from their focus on technology investments (that we rule out by assumption), the crucial difference between our approach and their way of modeling short-term agreements is that we allow countries that have joined a coalition to withdraw from the negotiations about a short-term agreement when the coalition decides not to sign a long-term agreement. As we have argued above, we believe that our approach is more in line with the rules of international law. It is straight-forward to show that if countries cannot withdraw from the negotiations (as assumed by Battaglini and Harstad 2016), then also in our model, there is no equilibrium in which more than k^{st} countries sign a long-term agreement. As pointed out by Battaglini and Harstad (2016), this is due to countries' incentives to free-ride on short-term agreements.

7 Conclusion

Allowing for the possibility that parties who are negotiating about a binding long-term agreement (such as a climate treaty) can meet again in the future and re-start nego-

tiations in case no agreement is reached today, captures an important aspect of many real-world negotiations. The main insight from our analysis is that the sheer *possibility* of future negotiations can drastically change the outcome of the negotiations. As we have demonstrated in the context of climate agreements, under mild conditions, a large coalition that achieves substantial welfare gains forms immediately in equilibrium. By contrast, it is well-known from the literature that in static models where countries can negotiate only once, the stable coalition size is generally small precisely when the potential welfare gains from cooperation are large.

Our results are driven by a threshold effect regarding the coalition size: Having the possibility to re-start negotiations in the future, countries become more demanding in the current period and are only willing to sign an agreement if the agreement achieves a lot, i.e., if the coalition is sufficiently large (from their perspective). Otherwise, they prefer to delay negotiations by one period, anticipating that a better outcome will be reached in the next period. Outcomes based on this threshold effect involve different trade-offs regarding countries' participation decisions, as compared to a static framework where a country's incentive to join a coalition typically reflects the positive effect upon the other signatories' abatement decisions. In our model, a country joins (primarily) to prevent inefficient delay. We have demonstrated that our main results are robust to different extensions or modifications in the setup, such as random vs. non-random identities of countries joining the coalition, and the possibility to sign short-term agreements.

To facilitate our formal analysis, we have abstracted from a variety of issues that play an important role in real-world climate negotiations, and that may help to explain why these negotiations repeatedly failed in the past. This includes countries' heterogeneity, different perceptions of the issue of climate change, stock pollution of greenhouse gases, technological change, asymmetric information, to name just a few examples. Including some of these issues in a dynamic analysis such as ours, may be an interesting starting point for future research.

References

[1] d'Aspremont, C., A. Jacquemin, J.J. Gabszewicz, and J.A. Weymark (1983). On the Stability of Collusive Price Leadership. *Canadian Journal of Economics*, 16, 17–25.

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

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

[4] Barrett, S. (2001). International cooperation for sale. *European Economic Review*, 45, 1835–1850.

[5] Barrett, S. (2003). Environment and Statecraft. Oxford: Oxford University Press.

[6] Barrett, S. (2005). The Theory of International Environmental Agreements. In *Handbook of Environmental Economics*, vol. 3, edited by K.-G. Mäler and J. R. Vincent. Amsterdam: Elsevier.

[7] Barrett, S. (2006). Climate Treaties and "Breakthrough" Technologies. *AEA Papers and Proceedings*, 96, 22–25.

[8] Barrett, S. (2013). Climate treaties and approaching catastrophes. *Journal of Environmental Economics and Management*, 66, 235–250.

[9] Battaglini, M. and B. Harstad (2016). Participation and Duration of Environmental Agreements. *Journal of Political Economy* 124, 160–204.

[10] Bloch, F. (1996). Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division. *Games and Economic Behavior*, 14, 90–123.

[11] Breitmeier, H., O. Young, and M. Zürn (2006) Analyzing International Environmental Regimes: From Case Studies to Database. Cambridge, MA: MIT Press.

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

[13] Carraro, C., J. Eyckmans, and M. Finus (2006). Optimal transfers and participation decisions in international environmental agreements. *Review of International Organizations* 1, 379–396.

[14] Diamantoudi, E. and E.S. Sartzetakis (2015). International Environmental Agreements - The Role of Foresight. *Working paper*.

[15] Dixit, A. and M. Olson (2000). Does Voluntary Participation Undermine the Coase Theorem? *Journal of Public Economic Theory*, 76, 309–335.

[16] Farrell, J. and E. Maskin (1989). Renegotiation in repeated games. *Games and economic behavior*, 1, 327–360.

[17] 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, 29–67.

[18] Finus, M. and S. Maus (2008). Modesty May Pay! *Journal of Public Economic Theory*, 10, 801–826.

[19] Finus, M. and B. Rundshagen (2003). How the Rules of Coalition Formation Affect Stability of International Environmental Agreements. *Working paper.*

[20] Harstad, B. (2016). Pledge-and-Review Bargaining. *Working paper.*

[21] Harstad, B., F. Lancia, and A. Russo (2019). Compliance Technology and Self-Enforcing Agreements. forthcoming in *Journal of the European Economic Association*, 17, 1–20.

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

[23] Hoel, M. (1992). International Environmental Conventions: The Case of Uniform Reductions of Emissions. *Environmental and Resource Economics*, 2, 141–159.

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

[25] 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, 395–406.

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

[27] Hong, F. and L. Karp (2014). International Environmental Agreements with Endogenous or Exogenous Risk. *Journal of the Association of Environmental and Resource Economists*, 1, 365–394.

[28] Karp, L. (2010). Incentives to Join International Environmental Agreements with Permit Trading and Safety Valves. *Working paper.*

[29] Karp, L. and H. Sakamoto (2018). Sober optimism and the formation of international environmental agreements. *Working paper.*

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

[31] Kolstad, C.D. and M. Toman (2005). The Economics of Climate Policy. In *Handbook of Environmental Economics*, vol. 3, edited by K.-G. Mäler and J.R. Vincent, Ch. 30, 1561–1618. Amsterdam: Elsevier

[32] Martimort, D. and W. Sand-Zantman (2016). A Mechanism Design Approach to Climate-Change Agreements. *Journal of the European Economic Association*, 14, 669–718.

[33] Palfrey, T.R., and H. Rosenthal (1984). Participation and the Provision of Discrete Public Goods: A Strategic Analysis. *Journal of Public Economics*, 24, 171–93.

[34] Ray, D. and R. Vohra (1997). Equilibrium Binding Agreements. *Journal of Economic Theory*, 73, 30–78.

[35] Ray, D. and R. Vohra (2001). Coalitional Power and Public Goods. *Journal of Political Economy*, 109, 1355–1384.

[36] Schmidt, R.C. (2017). Dynamic cooperation with tipping points in the climate system. *Oxford Economic Papers*, 69, 388–409.

[37] Young, O. (2011). Effectiveness of International Environmental Regimes: Existing Knowledge, Cutting-Edge Themes, and Research Strategies. *Proceedings of the National Academie of Sciences*, 108, 19854–19860.

[38] de Zeeuw, A. (2008). Dynamic Effects on the Stability of International Environmental Agreements. *Journal of Environmental Economics and Management*, 55, 163–74.

A Appendix: Proofs

Proof of Lemma 1. The proof follows from the main text. □

Proof of Propositions 1–3. The proofs follow directly from the main text. □

Proof of Proposition 4. Recall that ξ is monotonically increasing (Lemma 2) and that $\underline{k}$ is now a fixed point of ξ (due to the assumption $\underline{k} > k_0$). Moreover, by Assumption 5(a), its slope at $\underline{k}$ is smaller than 1. Thus, it is sufficient to show that the value $\xi(k)$ increases in some neighborhood of $\underline{k}$, when δ or π_0 increases.

Let us now rewrite $\xi(k)$ (as given in (7)) using the functions $\pi_s(k) = (1-\delta)\Pi_s(k)$ and $v(k) \equiv (1-\delta)V(k)$ that correspond to per period payoffs and thus do not depend on δ:

$$\begin{aligned} \pi_s(\xi(k)) &= (1-\delta)\pi_0 + \delta v(k), \\ \text{or equivalently,} \quad \xi(k) &= \pi_s^{-1}\big((1-\delta)\pi_0 + \delta v(k)\big). \end{aligned} \tag{26}$$

If follows from Assumption 2 that the function π_s is increasing. Thus, its inverse π_s^{-1} is increasing as well. This immediately implies that $\xi(k)$ increases when π_0 increases (for all $k > k_0$).

Moreover, (26) implies that $\xi(k)$ increases when δ increases for all k such that $v(k) > \pi_0$. In order to complete the proof, it remains to show that $v(\underline{k}) > \pi_0$. Recall that for $k > k_0$ we have $V(k) > \Pi_s(k)$, which can be rewritten as $v(k) > \pi_s(k)$. This indeed implies that $v(\underline{k}) > \pi_s(\underline{k}) > \pi_0$. □

Proof of Proposition 5. The proof of the first claim follows from the arguments in the main text.

Now consider the case with a series of short term agreements among k^* countries. We first show that $k^* = k^{st}$. As argued in the main text in any short term agreement, the coalition size is a most k^{st}. Thus, $k^* \leq k^{st}$. Consider $k^* < k^{st}$. Since k^* violates external stability (ES), non-signatories have incentives to join the coalition anticipating that again only a short term agreement would be signed. Thus, k^{st} is the only coalition size for which the countries sign a series of short-term agreements in equilibrium.

Now let $k^* = k^{st}$. Since k^{st} satisfies internal stability (IS), there is no profitable deviation by an insider. The payoff from a deviation by an outsider depends on whether the new coalition of size $k^{st}+1$ signs a long-term agreement or only a short-term agreement among k^{st} countries. If (21) holds, a deviation by an outsider is not profitable, when the new coalition of $k^{st}+1$ signs a long-term agreement. Moreover, a deviation by an outsider is not profitable, if the new coalition only signs a short-term agreement among k^{st} countries. In such a case, the payoff of the deviating country (in the current period) becomes $\pi_s(k^{st})$ or $\pi_n(k^{st})$ depending on whether the country participates in the short-

term agreement or not. However, none of these payoffs exceeds the original payoff $\pi_n(k^{st})$ from simply staying out.

If, on the other hand, (21) holds with the opposite inequality, i.e., $\Pi_s(k^{st}+1) > \pi_n(k^{st}) + \delta V(k^{st})$, then after an outsider joins the coalition, the $k^{st}+1$ countries prefer to sign a long-term agreement (as opposed to a short-term one with k^{st} countries). To see this, observe that the right-hand side is equal to the payoff of a country that would not sign the short-term agreement. Since $\pi_n(k^{st}) > \pi_s(k^{st})$, the payoff of a country that would sign the short-term agreement is even smaller. At the same time, the inequality $\Pi_s(k^{st}+1) > \pi_n(k^{st}) + \delta V(k^{st})$ implies that a deviation by an outsider who joins the coalition is profitable. Thus, signing only short-term agreements among k^{st} countries is not an equilibrium. □

Proof of Proposition 6. Recall that for equilibrium coalition size $k^* > k^{st}$, the analysis from Section 3, and in particular, the characterization from Proposition 3, remain valid. We will argue that only case (a) applies, i.e., that $\underline{k} > k^{st}$, or equivalently $k^{st} < \xi(k^{st})$. Recall that we now replace π_0 by $\pi_n(k^{st}) = (1-\delta)\Pi_n(k^{st})$. The inequality $k^{st} < \xi(k^{st})$ then becomes $\Pi_s(k^{st}) < (1-\delta)\Pi_n(k^{st}) + \delta V(k^{st})$. Using (5) to substitute for $V(k^{st})$ and rearranging, this becomes $0 < [1-\delta p(k^{st})]\,[\Pi_n(k^{st}) - \Pi_s(k^{st})]$, which indeed holds.

Now it remains to show that $\underline{k} > k^{st}+1$. For $k_{t+1} = k_t = k^*$, condition (22) becomes $\Pi_s(k^*) \geq \pi_n(k^{st}) + \delta V(k^*)$. After substituting for $V(k^*)$, this (necessary) condition can be (after rearranging) rewritten as:

$$\pi_s(k^*) - \pi_n(k^{st}) \geq \delta[1-p(k^*)]\,[\Pi_n(k^*) - \Pi_s(k^*)]. \tag{27}$$

Condition (27) cannot be fulfilled for $k^* = k^{st}+1$, since the left-hand side is negative due to (ES), that holds for $k^* = k^{st}$, while the right-hand side is positive. Thus, $k^{st}+1 < \xi(k^{st}+1)$, which implies that indeed $\underline{k} > k^{st}+1$. □

A.1 Additional technical lemmas (full dynamic game)

Lemma 2. *Function ξ is well defined and strictly increasing. Moreover, $\xi(k) \in [k_0, N)$ for all $k \in [k_0, N]$.*

Proof of Lemma 2. First we show that the function ξ is well defined. Recall that it follows from Assumptions 1–4 that

$$V(k_0) = \Pi_s(k_0) \leq \frac{\pi_0}{1-\delta} < \Pi_s(N) = V(N). \tag{28}$$

For $k \in [k_0, N]$ we then obtain

$$\Pi_s(k_0) \leq \pi_0 + \delta V(k_0) \leq \pi_0 + \delta V(k) \leq \pi_0 + \delta V(N) < \Pi_s(N). \tag{29}$$

The first and the last inequalities follow from (28), and the second and the third inequalities follow from the monotonicity of V. Continuity and monotonicity of Π_s then imply that there is a unique $k' \in [k_0, N)$ such that $\Pi_s(k') = \pi_0 + \delta V(k)$. Then we set $\xi(k) = k'$.

Second, we show that $\xi(k)$ is strictly increasing. Recall that by the definition of $\xi(k)$ we have $\xi(k) = \Pi_s^{-1}(\pi_0 + \delta V(k))$. By assumption, Π_s and V are strictly increasing. Hence, also the inverse function Π_s^{-1} is increasing, which shows that ξ is increasing.

Finally, the property $\xi(k) \in [k_0, N)$ follows directly from (29) and the definition of the function ξ. □

Lemma 3. *Function $\hat{\xi}$ has a fixed point in the interval $(k_0, N]$.*

Proof of Lemma 3. Since the function ξ is increasing, it follows from the inequalities in Lemma 2 that ξ maps the interval $[k_0, N]$ into the interval $[k_0, N)$. Let $\eta = \lfloor k_0 \rfloor + 1$ be the smallest integer (strictly) larger than k_0. Below we show that $\hat{\xi}(\eta) \geq \eta$ and $\hat{\xi}(N) \leq N$. Since the function $\hat{\xi}$ is weakly increasing, it maps the interval $[\eta, N]$ into itself and we can apply *Tarski's Fixed Point Theorem.* This implies that $\hat{\xi}$ has a fixed point in the interval $[\eta, N]$.

Now it remains to show that $\hat{\xi}(\eta) \geq \eta$ and $\hat{\xi}(N) \leq N$. It follows from the definition of η that $\eta > k_0$. Thus, $\hat{\xi}(\eta) \geq \xi(\eta) > \xi(k_0) \geq k_0$, where the first inequality follows from the definition of $\hat{\xi}$, the second one from ξ being strictly increasing (Lemma 2), and the third one from Lemma 2. Since $\hat{\xi}(\eta)$ is an integer and it is larger than k_0, we obtain $\hat{\xi}(\eta) \geq \eta$. Moreover, $\xi(N) < N$ due to Lemma 2. Because N is an integer, $\hat{\xi}(N) \leq N$, which completes the proof. □

Lemma 4. *A sufficient condition for Assumption 5(a) to hold is that* $\dfrac{\Pi_n(k) - \pi_0/(1-\delta)}{\Pi_s(k) - \pi_0/(1-\delta)}$ *is weakly decreasing for values of k such that $\Pi_s(k) > \pi_0/(1-\delta)$.*

Proof of Lemma 4. Recall that $V(k) = p(k)\Pi_s(k) + [1 - p(k)]\Pi_n(k)$, where $p(k) = k/N$ which is increasing in k. We discuss two cases.

First, let $\Pi_s(k) < \pi_0/(1-\delta)$. A straightforward computation yields that the inequality $\xi(k) > k$ or $\pi_0 + \delta V(k) > \Pi_s(k)$ can be rewritten as

$$\begin{aligned} \delta p(k)\left(\Pi_s(k) - \frac{\pi_0}{1-\delta}\right) + \delta[1-p(k)]\left(\Pi_n(k) - \frac{\pi_0}{1-\delta}\right) &> \Pi_s(k) - \frac{\pi_0}{1-\delta}, \\ p(k) + [1-p(k)]\underbrace{\frac{\Pi_n(k) - \pi_0/(1-\delta)}{\Pi_s(k) - \pi_0/(1-\delta)}}_{\varphi(k)} &< \frac{1}{\delta}. \end{aligned} \tag{30}$$

Since $\Pi_n(k) > \Pi_s(k)$ for $k > k_0$, we have $\varphi(k) < 1$. Therefore, $p(k) + [1 - p(k)]\varphi(k) < 1 < 1/\delta$ which implies that (30) is indeed satisfied. This shows that $\xi(k) > k$ when $\Pi_s(k) < \pi_0/(1-\delta)$.

Second, let $\Pi_s(k) > \pi_0/(1-\delta)$. By an analogous computation, the inequality $\xi(k) < k$ is now equivalent to (30). Since $\Pi_n(k) > \Pi_s(k)$ for $k > k_0$, we now have $\varphi(k) > 1$. Moreover, by assumption $\varphi'(k) \leq 0$, which implies that the left-hand side of (30) is decreasing, since its derivative is $p'(k)[1 - \varphi(k)] + [1 - p(k)]\varphi'(k) < 0$. Thus, it can attain the value $1/\delta$ at most once, and if it does, we denote it $\underline{k}$. Otherwise, we set $\underline{k} = k_0$. In both cases we obtain that $\xi(k) > k$ when $k_0 \leq k < \underline{k}$, while $\xi(k) < k$ when $\underline{k} < k \leq N$. □

A.2 Discussion of Propositions 1–4 for deterministic membership (Section 6)

Recall that now $V(k) = \Pi_s(k)$ and that $\xi(k)$ now satisfies $\Pi_s(\xi(k)) = \pi_0 + \delta\Pi_s(k)$. Clearly, Lemma 1 still holds under these definitions. Also note that the value k^{st} is derived from the static game, and is thus independent on whether the identities are random or deterministic.

Proposition 1 applies by the same arguments. As argued in the main text, k^{st} is an equilibrium coalition size, when k^{st} countries are indeed willing to sign an agreement, i.e., when $k^* = k^{st}$ satisfies (8). This condition is equivalent to the condition $k^{st} \geq \xi(k^{st})$ from the proposition. Observe that now this condition simplifies to $\Pi_s(k^{st}) \geq \pi_0/(1-\delta)$, which is satisfied by Assumption 4.

Consider now Proposition 2. Any equilibrium coalition size k^* needs to satisfy external stability and thus $k^* \geq k^{st}$. By the same argument as in Proposition 2, internal stability requires that the countries delay negotiations if one country leaves the coalition. Thus, condition (11) or equivalently, $k^* = \hat{\xi}(k^*)$, with redefined V and ξ applies.

Proposition 3 is a straightforward consequence of Proposition 1 and 2, and thus applies as well. Moreover, due to Assumption 4, we only obtain cases (b) and (c).

Finally, Proposition 4 follows from the property that $\Pi_s(\underline{k}) = \pi_0/(1-\delta)$. Since, $\Pi_s(k)$ is increasing for $k > k_0$, we indeed obtain that $\underline{k}$ is increasing in δ and in π_0 when $\underline{k} > k_0$.

B Supplementary Appendix (for online publication)

B.1 Additional properties of simple emission games

In this section we provide some additional results for simple benefit-cost emission games as introduced in Section 4. Recall that in the basic static climate cooperation game, the coalition acts as a Stackelberg leader who anticipates the equilibrium abatement of non-signatories, as given by (13). Let us denote $\hat{x}_n(x_s, k)$ the non-signatories' equilibrium abatement level in the subgame (of the emissions game) following the signatories' abatement level x_s (when k countries have signed the agreement), as given by (13). The coalition then chooses x_s in order to maximize $B(X) - C(x_s) = B(kx_s + (N - k)\hat{x}_n(x_s, k)) - C(x_s)$.[48] Furthermore, let us denote x_s^* and x_n^* the equilibrium abatement levels of a signatory and a non-signatory, respectively, and let $X^* = kx_s^* + (N - k)x_n^*$ be the equilibrium aggregate abatement level. We sometimes use the notation $x_s^*(k)$, $x_n^*(k)$, and $X^*(k)$ to highlight the dependence on the coalition size k.

Lemma 5. *In equilibrium of the simple emission game the following properties hold:*

(i) $\pi_n(k) \lesseqgtr \pi_s(k)$ *if and only if* $x_n^*(k) \gtreqless x_s^*(k)$.

(ii) $\pi_n(0) = \pi_n(k_0)$.

(iii) $\pi_s'(k_0) = 0$.

B.2 Extension: Finite negotiations

Here we study a modified version of our full dynamic game in which the negotiations can take place only for a finite number of periods. All other features of the model remain the same. In particular, the time horizon where the payoffs are realized is still infinite, and $\Pi_n(k)$ and $\Pi_s(k)$ represent the present values of payoffs over this infinite time horizon. However, we abstract from the possibility of signing short-term agreements, and we only consider the random membership case. The only difference to our model from Section 3 is that if no treaty is signed by the end of period T (where $T > 1$), then no treaty is signed whatsoever, and each country receives a stream of payoffs π_0 per period from period $T + 1$ onwards. This modification of the model allows us to investigate to what extent our previous results depend on the assumption of an infinite time horizon. An infinite time horizon is usually required to sustain tacit collusion in dynamic pricing games, where collusion breaks down completely if the time horizon is finite. By contrast, we show in the following that in our model, a high degree of cooperation typically emerges if the number of periods in which countries can negotiate is finite but sufficiently large.

[48] We omit the constant factor k.

In order to facilitate the comparison to the game with infinite negotiations, we consider symmetric subgame perfect equilibria with no delays.[49] We also impose Assumption 4. This now implies that if period T is reached (without signing any agreement before), then the countries are essentially in the same situation as in the static model, and in equilibrium k^{st} countries sign an agreement.

Let us now consider such an equilibrium and denote k_t^* (where $t = 1, 2, \ldots, T$) the number of countries that sign an agreement in period t (conditional on reaching that period). The above argument shows that $k_T^* = k^{st}$. Intuitively, one could expect a similar effect as for a finitely repeated prisoner's dilemma, where repeating the static equilibrium is the only subgame perfect outcome. However, this turns out not to be the case here, when the counties have the opportunity to delay the negotiations.

For illustration, consider Example 2 from Section 4 with the parameter values as illustrated in Figure 2 (i.e., $N = 10$ and $\delta = 0.6$). As we have argued there, the equilibrium coalition size in the static model is 2 or 3 countries, while in the dynamic model we have $\underline{k} = 5.4$ and $\bar{k} = 8.2$ with equilibrium coalition sizes 6, 7, and 8. Thus, in the last period T (if it has been reached), $k_T^* = 3$ countries sign an agreement (assuming that countries coordinate on the equilibrium with higher participation). Now, in period $T-1$ the countries anticipate the equilibrium in period T and expect $k_T^* = 3$ countries to sign an agreement. Recall from Lemma 1 (that applies also to this modified model) that in period $T-1$, the number of countries that sign an agreement, denoted k_{T-1}^*, is at least $\hat{\xi}(k_T^*) = \hat{\xi}(3) = 4$. Much like in Proposition 2, the equilibrium coalition size in period $T-1$ must be just large enough so that k_{T-1}^* countries are willing to sign an agreement in period $T-1$, but $k_{T-1}^* - 1$ countries are not. Thus, $k_{T-1}^* = \hat{\xi}(k_T^*) = 4$. Proceeding backwards, we obtain by the same argument that $k_{T-2}^* = \hat{\xi}(4) = 5$ countries sign an agreement in period $T-2$ and $k_{T-3}^* = \hat{\xi}(5) = 6$ countries sign an agreement in period $T-3$. Now since $\hat{\xi}(6) = 6$, the number of countries that would sign an agreement in earlier stages would be again 6. The following proposition provides general statements that are analogous to Proposition 3.

Proposition 7. *In the game with finite negotiations (with $T > 1$), the following statements hold:*

(i) There is an equilibrium where a coalition of size k^{st} signs an agreement in the first period, if and only if $k^{st} \geq \underline{k}$.

(ii) If $k^{st} < \underline{k}$ and T is sufficiently large, then there is an equilibrium where a coalition of size $\lceil \underline{k} \rceil$ signs an agreement in the first period.

Hence, the outcome under a finite number of negotiation stages (T) is characterized by a *ratcheting-up* in the coalition size from later towards earlier periods (see Figure 6 for

[49] Equilibria with delays are characterized in Section B.3.

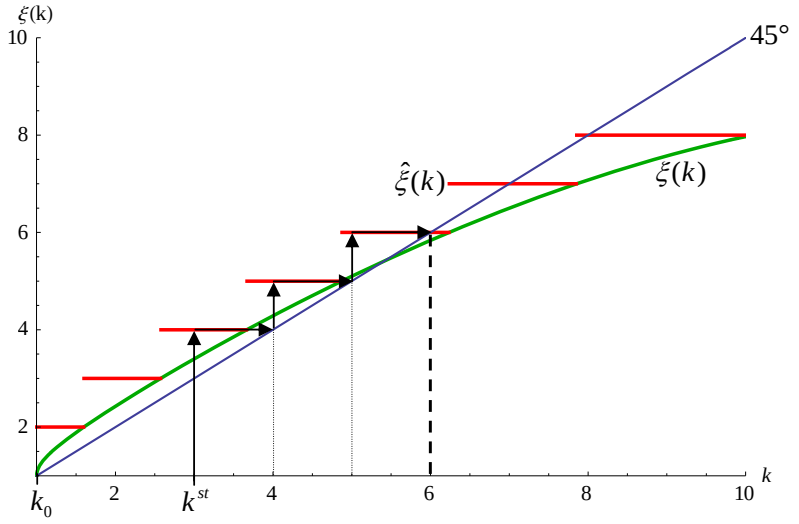

Figure 6: Illustration of *ratcheting-up* of the coalition size under finite T, for $\delta = 0.6$ and $N = 10$, random membership case (Example 2, Section 4)

a graphical illustration). This ratcheting-up stops when the maximum coalition size is reached, that coincides with the *smallest* stable coalition size ($\lceil \underline{k} \rceil$) under an infinite time horizon for the negotiations (in the case $\lceil \underline{k} \rceil > k^{st}$). Hence, the multiplicity of equilibria that we observed in the infinite horizon case (see Figure 2) vanishes.[50] Otherwise, the results are unchanged.[51]

B.3 Extension: Non-Markov equilibria and delay

In this section we explore what other kinds of equilibria (in pure strategies) can emerge in our dynamic coalition formation model when the Markov restriction, that was imposed in most sections of this paper (except Section B.2 where a finite time horizon T was assumed), is relaxed. We do not seek to provide a full characterization of all equilibria that exist. Instead, we focus on a subset of equilibria that deliver interesting new insights. Most importantly, we preserve the payoff structure from the previous sections. Thus, we rule out collusive strategies, where countries use their *emissions* to punish deviations from some collusive agreement. Such equilibria have been studied elsewhere (e.g., Barrett 1994; Harstad, Lancia, and Russo 2019) and are not the focus of this paper. Our focus is on binding long-term agreements, and the dynamics of reaching such an agreement given

[50] Here, we refer to the multiplicity of equilibrium coalition sizes that arises when the interval $[\underline{k}, \bar{k})$ contains several integers (see Proposition 3). Because Proposition 7 does not provide a full characterization of equilibrium outcomes for the model with finite negotiations, and the Markov restriction cannot be imposed here, some multiplicity may remain, especially with regards to non-Markov equilibria (see Section B.3 for further details).

[51] We have also analyzed finite negotiations under deterministic membership. Using similar arguments as above, we can show that there is always an equilibrium where a coalition of size k^{st} signs an agreement in the first period under deterministic membership.

the possibility to delay climate negotiations in one or several periods.

In particular, we maintain our earlier assumption that non-signatories choose their abatement efforts non-cooperatively and myopically in each period, while signatories of a long-term agreement choose their efforts so as to maximize their joint welfare. Furthermore, we do not allow for the possibility that countries can sign short-term agreements in periods where no long-term agreement has been reached yet. Hence, as in Section 3 countries' payoffs are fully captured by the functions Π_s and Π_n, by the size k of a coalition that signs a long-term agreement and the identity of the members of that coalition, as well as by the number of the period t in which the agreement is reached. We also maintain our assumption from Section 3 that the identity of coalition members (for a given coalition size k_t) is determined randomly in any period t (random membership case).

What is different when the Markov restriction is relaxed is that countries can condition their actions on the full history of the (participation) game up to that period. However, we preserve the assumption that at the signature stage the countries play a Pareto dominant equilibrium (if such exists).[52] In that respect, the countries may use only their non-participation, but not the use of veto power (during the signature stage) to punish deviations. It is well-known that strategies involving punishment (*grim trigger strategies*) can be used to sustain collusive agreements in infinitely repeated pricing games. We want to investigate if the *threat of delay* can be used in our setting to allow countries to reach a more cooperative outcome in the beginning of the game.

Before we give an answer to this question, let us first demonstrate that delay can actually occur along the equilibrium path in our setup. This is an interesting insight, given that delay has occurred many times in actual climate negotiations. To highlight this point, let us first consider the case where $k_0 = 0$ and $\Pi_s(0) = \pi_0/(1-\delta)$. Recall that by Assumption 2, the payoff functions Π_s and Π_n are increasing above k_0, which implies $\Pi_s(1) > \pi_0/(1-\delta)$. Therefore, there cannot exist a trivial equilibrium where no long-term agreement is signed in any period, so the existence of equilibria with delay is clearly not based on this. Furthermore, there cannot exist an equilibrium where fewer than k^{st} countries sign an agreement in the first period of the game, even if countries play non-Markov strategies that may involve delay in future periods (conditional upon reaching those periods).[53] Nevertheless, even under this simplifying assumption, subgame perfect Nash equilibria (SPNE) can exist that exhibit delay along the equilibrium path.[54]

To see this, suppose the payoff functions Π_s and Π_n are such that the static model exhibits a non-trivial amount of cooperation in equilibrium, that is: $k^{st} \geq 2$. Then by

[52] This rules out equilibria where the countries use the signature behavior for punishment, for instance by joining the coalition, but not signing unless all other countries have joined. Technically, it implies that Lemma 1 still applies.

[53] To see this, recall that for $k < k^{st}$ the external stability condition is violated, so that it would always be profitable for another outsider to join the coalition in the first period.

[54] We focus on SPNE whenever the Markov restriction is relaxed.

Proposition 2, there is always an equilibrium (in Markov strategies) where a coalition of size $k^* \geq k^{st}$ signs a long-term agreement. Suppose, if period $\tau \geq 2$ is reached, countries indeed play Markov strategies and coordinate on the stable coalition size $k^*_\tau = k^*$. Then if the discount factor is not too small, there clearly exists an equilibrium in the full dynamic model (without the Markov restriction) where no agreement is reached in the first $\tau - 1$ periods, if all countries anticipate that an agreement will be reached by k^*_τ countries in period τ, yielding a payoff of

$$\pi_0 + \delta\pi_0 + \cdots + \delta^{\tau-2}\pi_0 + \delta^{\tau-1}V(k^*_\tau).$$

For this to be an equilibrium outcome, countries must adopt strategies that lead to a sufficiently small coalition size (e.g., zero) in all periods $t < \tau$, so that even if an additional country would join the coalition in any of these periods, the coalition members still prefer not to sign a long-term agreement (anticipating that a more favorable outcome will be reached in period τ). Then obviously for an individual country that is assigned not to join the coalition in any of these periods, it is not profitable to deviate.

For instance, in period $\tau - 1$, the critical coalition size is $\hat{\xi}(k^*_\tau)$, so that for any $k_t < \hat{\xi}(k^*_\tau) - 1$, an individual country has no incentive to deviate and join as this does not lead to the signature of a long-term agreement in that period. This logic, of course, extends readily towards earlier periods, so that if k^*_τ and δ are sufficiently large, no agreement is signed in any period $t < \tau$ even when τ is a large number, assuming that countries adopt such *delay strategies*.[55] Relaxing the assumption $k_0 = 0$ only strengthens this point, so for the rest of this section, we drop this simplifying assumption.

We are now ready to state the main result of this section. It reveals that relaxing the Markov assumption does not support larger coalition sizes.[56]

Proposition 8. *In any SPNE (in pure strategies), the equilibrium coalition size satisfies* $k^* \leq \max\{k^{st}, \lceil \bar{k} \rceil - 1\}$.

Intuitively, why does a strategy that involves the threat to revert to a period (or a larger number of periods) of delay not help to sustain a more cooperative outcome in the first period of the game? The answer is, that if a large number of countries ($k^*_1 > \max\{k^{st}, \lceil \bar{k} \rceil - 1\}$) joins the coalition in the first period on the equilibrium path to avoid the punishment phase, then each of them realizes that after a deviation, the remaining $k^*_1 - 1$ countries would sign an agreement in period 1 as well. This renders the deviation profitable, as internal stability is violated. Extending the length of the punishment phase cannot help to avoid this problem, because this only reduces the continuation value so

[55] Note, however, that any country not assigned as coalition member in some period $t < \tau$, weakly prefers to join the coalition. Furthermore, given the possibility to block an agreement (unanimity rule), a country can never end up being *trapped* in an unfavorable agreement.

[56] Note, that if $\bar{k} > k^{st}$, then the inequality in Proposition 8 simplifies to $k^* < \bar{k}$.

that the critical coalition size in the first period needed to sign a long-term agreement is then even smaller. The largest stable coalition size in any period is obtained under the most optimistic (rational) expectations about the stable coalition size in the following period (in case the next period is reached). Therefore, any threat to punish by future delay only makes countries more eager to sign an agreement today, which reduces the stable coalition size. Such threats are, thus, ineffective in raising the stable coalition size in our model.

B.4 Proofs for Appendix B

Proof of Lemma 5. (i) Recall that $\pi_n(k) = B(X^*(k)) - C(x_n^*(k))$ and $\pi_s(k) = B(X^*(k)) - C(x_s^*(k))$. Then $\pi_n(k) - \pi_s(k) = C(x_s^*(k)) - C(x_n^*(k))$. The statement follows from the fact that the cost function C is increasing.

(ii) It follows from (i), that in both cases $k = 0$ and $k = k_0$ we have N countries that all choose an identical abatement level: $x_n = x_n^*(0)$, and $x_n = x_s = x_n^*(k_0)$, respectively. It remains to show that these two values are the same. In both cases, the abatement level satisfies the first-order condition (13), that now becomes $B'(Nx_n) - C'(x_n) = 0$. By assumption $B'' \leq 0$ and $C'' > 0$, which implies that $B'(Nx_n) - C'(x_n)$ is strictly decreasing in x_n. Thus, there can be at most one value of x_n that satisfies this condition. Consequently, $x_n^*(0) = x_n^*(k_0)$, which, due to (i), yields $\pi_n(0) = \pi_n(k_0)$.

(iii) Recall that $\hat{x}_n(x_s, k)$ is implicitly defined by (13), i.e., $B'(kx_s + (N-k)\hat{x}_n(x_s, k)) = C'(\hat{x}_n(x_s, k))$. Taking the derivative with respect to k, we obtain from the *Implicit function theorem* that (with some abuse of notation) for all x_s:

$$B''(X)\left[x_s - x_n + (N-k)\frac{\partial \hat{x}_n}{\partial k}(x_s, k)\right] = C''(x_n) \cdot \frac{\partial \hat{x}_n}{\partial k}(x_s, k). \tag{31}$$

For $k = k_0$ and $x_s = x_s^*(k_0)$ we have $x_n = \hat{x}_n(x_s^*(k_0), k_0) = x_n^*(k_0) = x_s^*(k_0)$, due to (i), and thus

$$[(N-k_0)B''(X^*) - C''(x_n^*)]\frac{\partial \hat{x}_n}{\partial k}(x_s^*, k_0) = 0.$$

Since $B'' \leq 0$ and $C'' > 0$, the term in the square bracket is negative, and we obtain

$$\frac{\partial \hat{x}_n}{\partial k}(x_s^*(k_0), k_0) = 0. \tag{32}$$

Now recall that $\pi_s(k) = \max_{x_s} B(kx_s + (N-k)\hat{x}_n(x_s, k)) - C(x_s)$. It follows from the

Envelope theorem and from (31) that

$$\begin{aligned}\pi_s'(k) &= \frac{\partial}{\partial k}[B(kx_s + (N-k)\hat{x}_n(x_s,k)) - C(x_s)]\Big|_{x_s=x_s^*} \\ &= B'(X^*)\left[x_s^* - x_n^* + (N-k)\frac{\partial \hat{x}_n}{\partial k}(x_s^*,k)\right].\end{aligned}$$

Since, for $k = k_0$ we have $x_s^* = x_n^*$ (by (i)), it follows from (32) that indeed $\pi_s'(k_0) = 0$. □

Proof of Proposition 7. (i) Consider the case $k^{st} \geq \xi(k^{st})$. As argued above the proposition, upon arriving in the final period with negotiations, $t = T$, the stable coalition size is $k_T^* = k^{st}$. Much like in Proposition 1, in period $T-1$ (if this period is reached), k^{st} countries are willing to sign an agreement. Thus, $k_{T-1}^* = k^{st}$ and the same arguments can readily be applied also to all other periods $t < T-1$. This completes the proof of (i).

(ii) Before proceeding with the proof we state the following lemma. Its proof follows below the proof of Proposition 7.

Lemma 6. *Assume that $k^{st} < \underline{k}$. Consider the following sequence defined recursively:*[57]

$$l_0 = k^{st} = \lceil \tilde{k} \rceil, \qquad l_\beta = \hat{\xi}(l_{\beta-1}) \quad \textit{for } \beta = 1, 2, \ldots. \tag{33}$$

Then there is some $\tau \geq 0$ such that $l_0 < l_1 < \cdots < l_{\tau-1} < l_\tau = l_{\tau+1} = \cdots = \lceil \underline{k} \rceil$.

Now we show that $k_t^* = l_{T-t}$ for $t = 1, 2, \ldots, T$. The proof proceeds in the same way as the argument preceding the proposition. As argued there, $k_T^* = \lceil \tilde{k} \rceil = l_0$. For any $t \leq T-1$, if the countries in period t anticipate that $k_{t+1}^* = l_{T-t-1}$ countries sign an agreement in the next period, then according to Lemma 1, $k_t^* \geq \hat{\xi}(k_{t+1}) = \hat{\xi}(l_{T-t-1}) = l_{T-t}$. Thus, the countries prefer to sign the agreement in period t (when this period is reached). In addition, since $k_t^* \geq k^{st}$ for all t, external stability is satisfied in all periods.

Now consider internal stability. Similarly as in the arguments preceding Proposition 1 and 2, we distinguish two cases: Either $k_t^* = \hat{\xi}(k_{t+1}^*)$ or $k_t^* > \hat{\xi}(k_{t+1}^*)$. We show that the former case applies. Otherwise, if $k_t^* > \hat{\xi}(k_{t+1}^*)$, the coalition size k_t^* needs to satisfy both the external and internal stability conditions (ES) and (IS), and would thus be an equilibrium coalition size of the static game (i.e., $k_t^* = l_0$). This is a contradiction, since $k_t^* > \hat{\xi}(k_{t+1}^*) = \hat{\xi}(l_{T-t-1}) = l_{T-t} \geq l_0 = \lceil \tilde{k} \rceil = k^{st}$. Thus, indeed the former case applies, which yields $k_t^* = \hat{\xi}(k_{t+1}^*) = \hat{\xi}(l_{T-t-1}) = l_{T-t}$.

In order to complete the proof of (ii), it is sufficient to set $T > \tau + 1$, where τ is introduced in Lemma 6. Then $k_1^* = l_{T-1} = l_\tau = \lceil \underline{k} \rceil$. □

Proof of Lemma 6. Before proceeding with the actual proof, recall that due to Assumption 5, $k_0 < k < \underline{k}$ implies $k < \xi(k) < \underline{k}$.

[57] We use the subscript β for counting *backwards* in time (see below).

First, we show that $l_{\beta-1} < l_\beta \leq \lceil \underline{k} \rceil$ when $l_{\beta-1} < \lceil \underline{k} \rceil$. Since $l_{\beta-1}$ is an integer, the inequality $l_{\beta-1} < \lceil \underline{k} \rceil$ implies $l_{\beta-1} < \underline{k}$. Then it follows that $l_{\beta-1} < \xi(l_{\beta-1}) < \underline{k}$. Since $l_\beta = \hat{\xi}(l_{\beta-1}) = \lceil \xi(l_{\beta-1}) \rceil$, we obtain $l_{\beta-1} < l_\beta \leq \lceil \underline{k} \rceil$.

Next, we show that $l_\beta = l_{\beta-1}$ when $l_{\beta-1} = \lceil \underline{k} \rceil$. Since $l_{\beta-1}$ is a positive integer and $l_{\beta-1} \in [\underline{k}, \bar{k})$, it follows from the discussion preceding Proposition 3 that $l_{\beta-1}$ is a fixed point of $\hat{\xi}$. Thus, $l_\beta = \hat{\xi}(l_{\beta-1}) = l_{\beta-1}$.

Summing up, since $l_0 = k^{st} < \lceil \underline{k} \rceil$, the sequence $l_0, l_1, l_2, \ldots$ is bounded from above by $\lceil \underline{k} \rceil$ and is increasing before attaining this bound. Let us set τ such that $l_{\tau-1} < \lceil \underline{k} \rceil = l_\tau$. Then $l_\beta = l_\tau = \lceil \underline{k} \rceil$ for $\beta \geq \tau$, which completes the proof. □

Proof of Proposition 8. Proof by contradiction. Let k^{max} be the largest stable coalition size in the full set of SPNE (in pure strategies), and suppose (to the contrary of the statement in the proposition) that $k^{max} > \max\{k^{st}, \lceil \bar{k} \rceil - 1\}$, which is equivalent to $k^{max} \geq \max\{k^{st} + 1, \bar{k}\}$.

Now consider an equilibrium where a coalition of $k_t = k^{max}$ countries signs an agreement at some stage t. We show that there there is a profitable deviation not to join the coalition for some member. If the remaining $k_t - 1$ coalition members do not sign an agreement, then $V(k^{max})$ is the maximal payoff the deviating country can expect in the next round. Thus, the payoff of each country after such a deviation is at most $\pi_0 + \delta V(k^{max})$. However, since $k_t = k^{max} \geq \bar{k}$, we have $\Pi_s(k_t - 1) \geq \pi_0 + \delta V(k^{max})$ and thus, the remaining countries would sign an agreement in period t. However, anticipating that the remaining countries sign an agreement, not joining the coalition is indeed a profitable deviation, since $k_t = k^{max} > k^{st}$ violates internal stability. □