A simple dynamic climate cooperation model *

EUGEN KOVÁČ† ROBERT C. SCHMIDT‡

August 1, 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 Katuscák, 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. Financial support from the German Research Foundation (GRK 1569) 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-duis.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.\footnote{See Hoel (1992), Carraro and Siniscalco (1993), Barrett (1994), Dixit and Olson (2000), Karp and Simon (2013).} In such a case, countries always sign an agreement, but generally with few members.

In the main part of the paper we focus on Markov perfect equilibria. Our results indicate 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.\footnote{This latter approach has also been adopted by Battaglini and Harstad (2016).} 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 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).

The remainder of this paper is organized as follows. Section 2 contains a literature review. Section 3 introduces the basic model, thereby adopting a reduced-form approach without imposing specific functional forms, and analyzes participation under the random membership approach. Section 4 considers specific examples to show how the payoff functions follow from an underlying emissions game, and illustrates our main results for these examples. Section 5 extends the basic model to allow for short-term agreements. Section 6 shows that our main results are qualitatively preserved under the deterministic membership approach. Section 7 discusses the role of commitment, by focusing on a modified setup where countries can commit only for a single period. Section 8 concludes. Some of the formal proofs are relegated to Appendix A. The Supplementary Appendix B (for online publication only) considers additional extensions. Appendix B.1 analyzes a variant of our model where countries can negotiate only in finitely many periods, to demonstrate that our main results do not hinge on an infinite time horizon. Appendix B.2 relaxes the Markov assumption and demonstrates that there can be also equilibria that involve delay on the equilibrium path, that, however, do not entail higher participation than without delay.
2 Related literature

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) as well as other papers to analyze the formation of International Environmental Agreements (IEAs) using game-theoretic tools.\(^3\) We refer to this as the “static” approach, because there is a single participation stage in which each country decides whether or not to participate in the (single) coalition, followed by a second stage where countries choose their abatements (a cooperative choice by the signatories, and non-cooperative by the non-signatories).\(^4\) 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.\(^5\)

The closest papers to ours are Karp and Sakamoto (2018), henceforth KS, and Battaglini and Harstad (2016), henceforth BH. Similar to us, these authors also analyze dynamic climate coalition formation games.\(^6\) The dynamic framework of KS 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.\(^7\) BH, by contrast, apply a deterministic membership approach, similar to the one we adopt for our model in Section 6. However, while we focus on a pure abatement game, BH consider technology investments in addition to countries’ abatement choices. They demonstrate how an endogenous length of the commitment period of a climate contract, in conjunction with incomplete contracts that

---

\(^3\)For an overview, see Barrett (2005) and Finus (2008).

\(^4\)The abatement stage is sometimes split into two separate stages. The signatories are then typically assumed to move first, acknowledging that the coalition behaves more sluggishly which translates into a Stackelberg leadership position.

\(^5\)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).

\(^6\)There is a separate strand of 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).

\(^7\)KS nevertheless focus on pure strategies. Mixed strategies in countries’ participation decisions are considered by Dixit and Olson (2000) and Hong and Karp (2012, 2014) in static (one-shot) games. Similar to KS, we focus on pure strategies.
exclude countries’ technology investments (leading to a hold-up problem) can explain larger coalition sizes than predicted by static models.

Our approach shares with both of these models the important insight that large coalitions that achieve substantial welfare gains can be stable in equilibrium, in spite of countries’ free-rider incentives. To highlight our contribution, it is important to understand how our approach differs from the ones of KS and BH.\(^8\) It turns out that the mechanisms at work are fundamentally different in each of these models. This also explains why the conditions under which large coalitions can be stable in equilibrium, are rather different in each model, which can also lead to different policy implications. Surprisingly, in spite of our simple modeling framework (see Section 3), large welfare gains from cooperation can be achieved under fairly general conditions.

A crucial difference between our approach and the one by KS is that in our model, the game ends when coalition members sign a long-term agreement, since coalition members commit to a long-term agreement. In KS, by contrast, countries cannot commit for more than one period, and 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, each lasting only for a single period. Our model is simpler in that aspect, 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. As we explain in Section 7, our main results are preserved even if today’s coalition members could decide to dissolve the coalition at the beginning of the next period (as in KS). This holds as long dissolving the coalition causes delay. We believe that such delay after dissolving a coalition, in order to organize a new round of climate negotiations, is a realistic feature of our model. By contrast, in KS, the incentive to maintain an existing coalition stems from the risk of reaching a worse agreement if the randomization device is again used to select a new stable coalition.\(^9\)

BH emphasize that in order to obtain large coalitions that achieve substantial welfare gains in their model, it is necessary to assume an endogenous length of the commitment period of a climate contract, in conjunction with incomplete contracts that lead to a hold-up problem regarding countries’ technology investments. In this sense, the conditions under which large coalitions are stable, are much more specific than in our model where neither of these two conditions is required. In our model, it is the sheer possibility...
bility to negotiate again in the future that makes large agreements stable.\textsuperscript{10} The key to understand the differences between our approach and the one of BH are countries’ incentives to free-ride on short-term agreements. While our basic model (Section 3) abstracts from short-term agreements, the possibility to sign agreements that last only for a single period are a key feature of the model by BH.\textsuperscript{11} To facilitate a comparison of our approaches, we extend our model in Section 5 to allow also for short-term agreements (see also Section 6.1). However, we depart in one “detail” from the way in which BH model short-term agreements. In their model, a short-term agreement is implemented, if the members of today’s coalition decide to sign an agreement which lasts for only one period. This is the case if the coalition size is perceived as “too small”, e.g., due to the deviation of a country that was supposed to join the coalition (along the equilibrium path), but decided to stay out. Then the remaining coalition members prefer to sign only a short-term agreement, anticipating that the “missing” country will join in the next period. However, the remaining coalition members are then effectively locked-in, and may end up signing a short-term agreement with more members than is consistent with the condition of “internal stability” in a corresponding static setup. Hence, following the deviation of a single country, also other coalition members might prefer to reverse their participation decision in order to free-ride on the short-term agreement. This lock-in seems to be in conflict with the rules of international law, according to which countries are “free to participate in treaties or not as they please” (Barrett 2005). Therefore, when modeling negotiations about short-term agreements, we depart from BH by allowing countries that have joined a coalition (hoping to reach a long-term agreement) to withdraw from the subsequent negotiations about a short-term agreement that start if the coalition decides not to sign a long-term agreement.

This seemingly minor change in the way in which negotiations about short-term agreements are modeled, explains why BH 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, while the possibility to sign short-term agreements tends to increase participation in a long-term agreement in our model. The sensitivity of our models to such modeling details, 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

\textsuperscript{10}Our finding that adding dynamics to a simple climate cooperation model can change the results (drastically) in the positive direction is not obvious. In other cases, the opposite holds true (we are grateful to an anonymous referee for pointing this out). E.g., countries may underinvest in green technology if access to a better technology reduces their bargaining power in future negotiations (see Buchholz and Konrad 1994, and Beccherle and Tirole 2011). Similarly, under convex damage costs from emissions, the free-riding incentive may be strengthened, see Fershtman and Nitzan (1991).

\textsuperscript{11}While these authors allow coalition members to freely determine the length of the commitment period, effectively, in their model coalition members will either choose to commit for infinitely many periods (long-term contract), or for a single period only (short-term contract).
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.\textsuperscript{12}

In terms of policy implications, BH conclude that it is important that countries can determine the length of the commitment period endogenously. Our results highlight that it can be helpful to leave the door open to future negotiations, in case current negotiations fail. This is because the sheer possibility to negotiate again in the future changes countries’ trade-offs whether or not to sign a long-term agreement today fundamentally, compared to a one-shot setting where players can negotiate only once. This conclusion may hold more generally, also for applications other than climate negotiations (e.g., cooperation on the provision of some public good).

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) and Nordhaus (2015) consider the possibility of trade sanctions to foster participation in a climate agreement resp. a climate club.\textsuperscript{13}

\section{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. The length of the period can be interpreted as the amount of time required for countries to organize a new round of climate negotiations.

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 individ-

\textsuperscript{12}Inspired by the Paris climate-change agreement, pledge-and-review bargaining is analyzed formally by Harstad (2019).

ually 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 $\Pi_s$ and $\Pi_n$ 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. In Section 4 we analyze a class of simple emission games where the welfare functions are derived from a simple cost-benefit analysis with common benefits but individual costs. We also provide several specific examples of emission games and derive corresponding welfare functions.

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.\footnote{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.} 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 $t$](image)

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, the kt 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, 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 per period.

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. 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:

\[
\begin{align*}
\Pi_n(k) & \geq \Pi_s(k + 1), \\
\Pi_s(k) & \geq \Pi_n(k - 1). 
\end{align*}
\]

External stability (ES) requires that outsiders (non-signatories) have no incentives to join the coalition. 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 \( \Pi_n(k) \leq \Pi_s(k) \) for \( k \leq k_0 \). 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). \( \Pi_n(0) \leq \frac{\pi_0}{1 - \delta} \leq \Pi_s(k_{st}) \).

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

---

15In 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 accommodate also this case.
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 reversed. While it is typically the case that \( k_0 \leq 1 \), or equivalently \( \Pi_n(k) \geq \Pi_s(k) \) for \( k \geq 1 \), it is also possible that \( k_0 > 1 \) depending on the shape of the underlying benefits from abatement (see Example 4 from Section 4). The value \( k_0 \) thus 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 this normally means that in both cases there are \( N \) countries that behave in the same way (see Lemma 4 and Section 4).

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 \( \pi_0 \) 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 \( \Pi_s \) and \( \Pi_n \).

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.\(^\text{16}\) 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.

Now consider some coalition of size \( k_t \) that would form in period \( t \), provided that period

\(^{16}\)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 \( x \).
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 $k_t$.\(^{17}\)

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.\(^{18}\) 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 1989).\(^{19,20}\)

\(^{17}\)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.

\(^{18}\)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.

\(^{19}\)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.

\(^{20}\)As an alternative approach, one could assume that coalition members decide collectively whether to
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

\[
V(k) = p(k)\Pi_s(k) + (1 - p(k))\Pi_n(k),
\]

where \( p(k) = k/N \) denotes the probability of being assigned as a coalition member in period \( t + 1 \). 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

\[
\Pi_s(k_t) \geq \pi_0 + \delta V(k_{t+1}).
\]

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 \( \pi_0 \) 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

\[
\Pi_s(\xi(k)) = \pi_0 + \delta V(k), \quad \text{or equivalently}, \quad \xi(k) = \Pi_s^{-1}(\pi_0 + \delta V(k)).
\]

**Lemma 1.** Function \( \xi \) is well defined and strictly increasing. Moreover, \( \xi(k) \in [k_0, N] \) for all \( k \in [k_0, N] \).

See Appendix A for a proof of the lemma. It follows that a coalition of size \( k_t = \xi(k) \) 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 (2) yield the following lemma.

**Lemma 2.** 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 the coalition of size \( k_t \geq k_0 \) signs an agreement in period \( t \), if and only if \( k_t \geq \hat{\xi}(k_{t+1}) \).

\[\text{sign an agreement or not (e.g., by delegating their individual decisions to a social planner who decides on behalf of all signatories).}\]

\[\text{Under the deterministic membership approach, } p(k) = 1. \text{ See Section 6 for further details.}\]
Note that the finding that function \( \xi \) 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.\(^22\) 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.\(^23\) 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.\(^24\) Let us denote \( k^\ast \) 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 \( k_{t+1} = k^\ast \) is expected to form in the next period with probability 1.

In what follows we provide conditions for \( k^\ast \) to be an equilibrium coalition size. As a first observation, consider when the \( k^\ast \) countries are indeed willing to sign an agreement on the path. According to condition (2), this is the case when

\[
\Pi_s(k^\ast) \geq \pi_0 + \delta V(k^\ast), \quad \text{or equivalently,} \quad k^\ast \geq \hat{\xi}(k^\ast),
\]

where the equivalence follows from Lemma 2. Thus, condition (4) is necessary for \( k^\ast \) to be an equilibrium coalition size.\(^25\) 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., \( k_t < \hat{\xi}(k^\ast) \)), then this coalition dissolves.\(^26\) Below we show that this logic gives rise to a novel type of equilibrium, where coalitional stability is driven by the endogenous threshold effect regarding the minimum coalition size.

\(^{22}\)In Supplementary Appendix B.1 we analyze a variant of the model with finitely many periods of negotiations, and in Supplementary Appendix B.2 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.

\(^{23}\)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).

\(^{24}\)Equilibria with delays are characterized in Supplementary Appendix B.2, where we relax the Markov restriction.

\(^{25}\)Note that \( k^\ast < \xi(k^\ast) \) would imply that in each period a coalition forms that remains inactive.

\(^{26}\)Note that since \( V(k) \geq \Pi_s(k) \), inequality (4) implies that \( \Pi_s(k^\ast) \geq \pi_0/(1 - \delta) \).
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 (4)) 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 2) 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 (4) 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 2, 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 (4) 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

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

or

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

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 (5) and (6) 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 (6) 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

\[ \xi(k^*) \leq k^* < \xi(k^*) + 1, \quad (7) \]

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 (7) yield an upper and a lower bound on the equilibrium coalition size.

Summing up the above arguments, we obtain 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.,

\[ k^* = \hat{\xi}(k^*). \quad (DS) \]

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 $\hat{\xi}$.

The following lemma provides an existence result (see Appendix A for a proof).

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

As follows from Proposition 2, a fixed point of $\hat{\xi}$ 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

---

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

28 Note that in the special case when $\Pi_s(k_0) = \pi_0/(1 - \delta)$, i.e., when the left condition in Assumption 4 holds with equality, a trivial fixed point of $\xi$ is $k = k_0$. To see this, recall that due to Assumption 1, we have $V(k_0) = \Pi_s(k_0) = \Pi_s(k_0) = \pi_0/(1 - \delta)$, 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 $\xi$.
fixed points if $\xi$ has several fixed points or when $\xi$ has a slope close to 1. The latter case is illustrated in Figure 2, that shows the functions $\xi$ and $\hat{\xi}$, based on a specification of payoff functions $\Pi_s$ and $\Pi_n$ from an example introduced in the following section. (We omit further details at this point.) As can be seen from the figure, $\xi$ 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 (7), in order to characterize equilibrium coalition sizes, we need to identify the fixed points of functions $\xi$ 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 $\bar{k} \in [k_0, N]$ such that $k \leq \xi(k)$ if $k \leq \bar{k}$.

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

Assumption 5 postulates that each of the functions $\xi$ and $\xi + 1$ has at most one fixed point, and if it does, it is the point $\bar{k}$ and $\tilde{k}$, respectively. Moreover, at the fixed point, the corresponding function ($\xi$ or $\xi + 1$) crosses the 45°-line from above. The assumption also allows for cases, where $\bar{k} = k_0$, when $\xi$ lies below 45°-line on the interval $[k_0, N]$, and where $\tilde{k} = N$, 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 $\bar{k} < \tilde{k}$.

Intuitively, the function $\xi$ 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 $\xi$ 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.29

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

---

29They are not exactly balanced, unless $\hat{\xi}$ coincides with $\xi$ at a fixed point of $\xi$ which requires that the latter is an integer. Otherwise, at a fixed point of $\xi$ 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 \(\xi\) is concave. Lemma 5 in Appendix A.1 provides an alternative sufficient (albeit not necessary) condition for Assumption 5(a) to hold, based on monotonicity of the functions \(\Pi_s\) and \(\Pi_n\). It can be verified that the assumption in Lemma 5 is indeed satisfied in all our examples in Section 4 (in Examples 1 and 3, the ratio in Lemma 5 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 \(k^{st} \geq \bar{k}\). 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 (7) that, under Assumption 5, the set of all fixed points of \(\hat{\xi}\) is in the interval \([\bar{k}, \tilde{k}]\). Depending on the size of \(k^{st}\) relative to \(\bar{k}\) and \(\tilde{k}\), 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 \([k, \bar{k}]\), if \(k^{st} < \bar{k}\);

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

(c) \(\{k^{st}\}\), if \(k^{st} \geq \tilde{k}\).

The proposition implies that any equilibrium coalition sizes in the dynamic game are bounded from below by \(\max\{k^{st}, \bar{k}\}\) and from above by \(\max\{k^{st}, \tilde{k}\}\). In the case where \(k^{st} < \bar{k}\) (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 \([k, \bar{k})\), where \( k \approx 5.4 \) is a fixed point of \( \xi \) 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).

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 \( \delta \) and the per-period payoff when no agreement is signed, \( \pi_0 \). Due to multiplicity of equilibria, we consider comparative statics on \( \bar{k} \). Recall that \( \Pi_s \) and \( \Pi_n \) 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

![Figure 2: Illustration of \( \xi(k) \) and \( \hat{\xi}(k) \) for Example 2 (Section 4) with \( \delta = 0.6, N = 10 \)](image)
present values are derived from time-independent interaction, so that

\[ \pi_s(k) = (1 - \delta)\Pi_s(k) \quad \text{and} \quad \pi_n(k) = (1 - \delta)\Pi_n(k) \]  

(8)

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 (5) resp. (6)) can now be formulated for the per-period payoffs \( \pi_s(k) \) and \( \pi_n(k) \), that do not depend on \( \delta \) nor on \( \pi_0 \). Thus, also the value of \( k^{st} \) does not depend on \( \delta \) nor on \( \pi_0 \).

In conjunction with Proposition 2, the next result thus indicates that in our dynamic climate cooperation model, an increase in \( \delta \) or in \( \pi_0 \) leads to (weakly) larger equilibrium coalition sizes.

**Proposition 4.** Assume that the per-period payoffs after signing an agreement are constant over time. If \( k > k_0 \), then the value of \( k \) is increasing in \( \delta \) and in \( \pi_0 \).

See Appendix A for a proof. To see the intuition, consider an equilibrium coalition size \( k^* \) (such that \( k^* > k^{st} \)). An increase in \( \delta \) or in \( \pi_0 \) 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 \( \pi_0 \) 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 \( k \) increases in \( \pi_0 \). 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 \( k \) increases in \( \delta \).

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 \( \Pi_s \) and \( \Pi_n \). Similarly as in Proposition 4, we consider the case where the present values \( \Pi_s \) and \( \Pi_n \) are outcomes of time-independent (and myopic) interaction (see (8)). In addition, we assume that in periods without an agreement, the countries obtain the non-cooperative equilibrium payoff \( \pi_0 = \pi_n(0) = (1 - \delta)\Pi_n(0) \).

\[ \text{(30)} \]

Lemma 4 below shows that for the class of games considered in this section, we obtain \( \pi_n(0) = \pi_n(k_0) \). This means that \( \pi_0 \) attains its lower bound from Assumption 4.
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 \textit{simple emission 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

\[ B'((kx_s + (N-k)x_n) = C'(x_n). \]  

The coalition, being a Stackelberg leader, anticipates the equilibrium abatement of non-signatories and chooses its abatement level. Let us denote \( x_s^*(k) \) and \( x_n^*(k) \) the equilibrium abatement levels of a signatory and a non-signatory. 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 (9) and the corresponding solution to the coalition’s maximization problem).

The following lemma provides some basic properties of simple emission games (see Appendix A for a proof).

\textbf{Lemma 4.} \textit{In equilibrium of the simple emission game the following properties hold:}

\begin{itemize}
  \item[(i)] \( \pi_n^*(k) \leq \pi_s^*(k) \) if and only if \( x_n^*(k) \geq x_s^*(k) \).
\end{itemize}
(ii) $\pi_n(0) = \pi_n(k_0)$.

(iii) $\pi'_s(k_0) = 0$.

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 $C(x_i) = \frac{1}{2}cx^2_i$ (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 $k[B(X) - C(x_s)] = k(bkx_s - \frac{1}{2}cx^2_s)$. 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:

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

The discounted payoffs $\Pi_s$ and $\Pi_n$ are then given by (8). 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.\(^{31}\) For the above payoff functions it is easy to verify that $\bar{k} = 1 + \sqrt{2}$ (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 (10) and from $\pi_0 = 0$ that $\xi(k) = k\sqrt{\delta(2-k/N)}$.\(^{32}\) Now recall that $\bar{k}$ is a fixed point of the function $\xi$, i.e., $\bar{k} = \xi(\bar{k})$, if such exists. This is the case when $\delta \geq \frac{1}{2}$, which then yields

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

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

The simple condition (11) nicely captures the central result of this paper: For sufficiently large values of $\delta$ (and $N$), the equilibrium coalition size is strictly larger than in the static model, and for $\delta$ 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

\(^{31}\)For $k = 0$ we have $\Pi_s(0) = \Pi_n(0) = \pi_s(0) = \pi_n(0) = 0$. Thus, also $k_0 = 0$.

\(^{32}\)Recall that $\pi_s(\xi(k)) = (1 - \delta)\pi_0 + \delta v(k)$, where $v(k) \equiv k/N \cdot \pi_s(k) + (1 - k/N)\pi_n(k)$.
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 (5), rather than (6) and the dynamic model leads to identical results as the static one. Notice that in this example, the lower bound \( k \) for the size of a stable coalition (for \( \delta \geq \frac{1}{2} \)), (11), 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 considered 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, \( x_n^* = b/c \). 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):

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

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

Using (12) and (3), we obtain \( \xi(k) = 1 + \sqrt{\delta k(k - 1)[2 - (k + 1)/N]} \), and can determine \( k \) to obtain for the dynamic model:

\[
\frac{k}{N} = 1 - \frac{1}{2\delta} - \frac{1}{2N} + \sqrt{\frac{1}{\delta N} + \left(1 - \frac{1}{2\delta} - \frac{1}{2N}\right)^2}.
\] (13)

It is easy to show that for large values of \( \delta \), the values of \( k \) defined by (11) and (13) are close (see Figure 3 for an illustration). Thus, the simple formula (11) 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 (\( \delta \) is large). As can be seen from (12), for large \( k \), the payoff ratio \( \Pi_n(k)/\Pi_s(k) = \pi_n(k)/\pi_s(k) \) is then close to 2, while

\(^{33}\)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.

\(^{34}\)Note, that now we have \( \Pi_n(1) = \Pi_n(1) = \Pi_n(0) \). This also implies that \( k_0 = 1 \).

\(^{35}\)Recall that for convenience we have assumed in the general analysis that \( \bar{k} \) is not an integer. This assumption is not crucial, it only facilitates the formulation of our results.
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.

![Figure 3: $k$ as function of $\delta$ (for $N = 200$), in Example 1 ($x_n = 0$) and Example 2 ($x_n > 0$)](image)

**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.,

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

(14)

for all $k > 0$. Furthermore, let $\pi_0 = \pi_n(0) = 0$ as in the basic example.\(^{36}\) The parameter $\alpha$ captures the free-rider incentives. If $\alpha$ 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 $\alpha$ is large, then being a non-signatory is a lot more profitable than being a member of an active agreement.\(^{37}\) A fixed ratio $\alpha$ as in (14), 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, $\alpha = \gamma/(\gamma - 1)$.\(^{38}\) Hence, the free-rider incentives are more intense the closer the parameter $\gamma$ is to 1. The case of quadratic costs from Example 1 is obtained for $\gamma = 2$.

In this example we cannot compute the function $\xi$ explicitly. However, we can evaluate

\(^{36}\)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 $\pi_0$ is sufficiently small.

\(^{37}\)It is easy to show that in the static case, only a small coalition is stable when $\alpha$ is large.

\(^{38}\)A straightforward computation yields the per period payoffs $\pi_s(k) = (\gamma - 1)/\gamma \cdot [(bk)^\gamma/c]^1/(\gamma - 1)$ and $\pi_n(k) = [(bk)^\gamma/c]^1/(\gamma - 1)$. 

23
\( k \) 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
\[
\frac{k}{N} = \frac{\alpha - 1/\delta}{\alpha - 1},
\]
when \( \alpha \delta \geq 1 \), while \( k = 0 \) when \( \alpha \delta < 1 \). This generalizes the result (11).

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, \( 1 - p(k^*) = (N - k^*)/N \), 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:
\[
x_n = \frac{b - dkx_s}{c + (N - k)d} \quad \text{and} \quad x_s = \frac{bck}{c^2 + d^2(N - k)^2 + cd(k^2 + 2(N - k))}.
\]

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

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 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 $\xi$ fulfills our assumed properties. In particular, we find that it has a simple concave shape for $k \geq k_0$ (not shown), so that $\xi$ and $\xi + 1$ obviously have at most one fixed point above $k_0$, and at a fixed point, $\xi$ 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 $\hat{k} \approx 4.16$, and thus the stable coalition size in the static model is $k^{st} = 5$. In the dynamic model we obtain for $\delta = 0.9$: $\bar{k} \approx 9.04$ and $\tilde{k} = 11$ (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 $\pi_0$), 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 (8).

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).\(^{39}\) 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.\(^{40}\) 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

\(^{39}\)We comment on this approach in more detail in Section 6.1, where we also compare our approach with the one chosen by BH.

\(^{40}\)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.
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.\(^{41}\)

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 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 \( k_t < k_0 \).\(^{42}\)

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\(^{43}\)

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

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 (2). 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 \( \Pi_s(k^*) > \Pi_s(k^{st}) \) 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

\(^{41}\)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).

\(^{42}\)If \( k_t < k_0 \) and \( k_0 > 1 \) (see Example 4 in Section 4), then the \( k_t \) coalition members sign a short-term agreement.

\(^{43}\)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 \( k_t \leq k^{st} \).
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.\footnote{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.}

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

\textbf{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

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

(17)

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

Intuitively, if (17) 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 (17) 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:

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

(18)

This replaces our earlier condition (2) for coalition sizes greater than \( k^{st} \). Hence, we can define the function \( \xi \) by condition (3) as before, if we replace \( \pi_0 \) 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 \( \bar{k} \) and \( \tilde{k} \).

**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 \([k, \bar{k})\). Moreover, \( \bar{k} > k^{st} + 1 \).

The proposition parallels Proposition 3 that provides a characterization of equilibrium coalition sizes. However, here we only obtain case (a) where \( \bar{k} > k^{st} \). In addition, it holds that \( \bar{k} > k^{st} + 1 \), 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 \( \pi_0 \) (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.\(^{45}\)

\(^{45}\)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 \( N = 10 \)): \( \bar{k} \approx 6.8 \) and \( \approx 8.9 \), so that the equilibrium coalition sizes are \( k^* = 7 \) and \( k^* = 8 \), whereas without short-term agreements we had \( \bar{k} \approx 5.4 \) and \( \approx 8.2 \), so that \( k^* = 6 \) was also an equilibrium coalition size.
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 BH, among others). From a theoretical perspective, these identities may be simply selected randomly at the beginning of the game.\(^{46}\) 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 pre-determined 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 (1) that \( V(k) = \Pi_s(k) \). The condition (3) then simplifies to

\[
\Pi_s(\xi(k)) = \pi_0 + \delta \Pi_s(k),
\]

where \( \pi_0 \) is again treated as an independent parameter, bounded by Assumption 4 (as in Section 3). Then \( k^* \), which is a fixed point of the function \( \xi \), satisfies

\[
\Pi_s(k^*) = \frac{\pi_0}{1 - \delta}.
\]

In contrast to the random membership case, now we don’t need to impose the single-
crossing property in Assumption 5(a) on function \( \xi \); it follows from Assumption 4 and the monotonicity of \( \Pi_s \). 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 2 and Propositions 1–4 remain valid. Moreover, due to Assumption 4, property (20) implies that \( k^{st} \) is always an equilibrium coalition size and only the cases (b) and (c) in Proposition 3 apply.

Finally, note that equation (20) 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, \( \bar{k} \) 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 (10) into (19), we find that \( \xi(k) = \sqrt{\delta} k \). The function \( \xi \) is linear and it follows that \( \bar{k} = 0 \) and \( \bar{k} = \min \{1/(1 - \sqrt{\delta}), N\} \). Thus, the set of stable coalition sizes (fixed points of \( \hat{\xi} \)) is large when \( \delta \) 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 (12), we obtain \( \xi(k) = 1 - \sqrt{\delta} + \sqrt{\delta} k \). Again, the function \( \xi(k) \) is linear, and we have \( \bar{k} = 1 \) and \( \bar{k} = \min \{1 + 1/(1 - \sqrt{\delta}), N\} \). We then again obtain that large coalitions are stable when \( \delta \) 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 \( k^{st} = 3 \)).

**Example 3’**. Finally, consider the generalized Example 3, with linear benefits and cost function \( C(x_i) = ax_i^\gamma/\gamma \) (where \( \gamma > 1 \) and \( \alpha = \gamma/(\gamma - 1) \)). In that case, we obtain \( \bar{k} = 0 \) 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 \( \delta \) 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 (18) when $V(k_{t+1}) = \Pi_s(k_{t+1})$. Hence, again when replacing $\pi_0$ by $\pi_n(k^{st})$ as in Section 5, the definition of the function $\xi$, (19), 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 $k^\text{st}$ signs a short-term agreement in each period, if and only if

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

(21)

However, in contrast to Proposition 5, under the deterministic membership approach, there is also an equilibrium where $k^\text{st}$ countries sign a long-term agreement when (21) 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 $k^\text{st}$ 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 $k^* > k^\text{st}$ is (in this version of the model), thus, characterized by the dynamic stability condition, $(DS)$ with $\pi_0 = \pi_n(k^\text{st})$. 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 $\pi_0$ upon $k$ and $\bar{k}$. It is straightforward to see from (20) that $\bar{k}$ increases in $\pi_0$. Now consider the effect on the upper bound on the stable coalition size, $\bar{k}$. Recall that $\bar{k}$ now satisfies $\Pi_s(\bar{k} - 1) - \delta \Pi_s(\bar{k}) = \pi_n(k^\text{st})$. 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 $\pi_0$ and, thus, introducing the possibility to sign short-term agreements increases the values of $k$ 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 $\delta$. For instance, in the Example 2 in Section 4, for $\delta = 0.5$ we obtain $\bar{k} \approx 4.7$ and $\bar{k} \approx 6.7$, yielding equilibrium coalition sizes $k^* = 5$ and $k^* = 6$, with the possibility to sign short-term agreements, while $\bar{k} = 1$ and $\bar{k} \approx 4.4$, yielding equilibrium coalition sizes $k^* = 3$ and $k^* = 4$, without it (recall that $k^\text{st} = 3$ in that example). For larger values of $\delta$, the impact on the set of equilibrium coalition sizes (in particular on $\bar{k}$) is less pronounced (not shown here).

Finally, we would like to point out that the model analyzed in this subsection is close to the setup used by BH. Let us briefly discuss their relation (see also our more general discussion in Section 2). BH 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 in Section 5, we believe that our approach is more in line with the rules of international law. It is straightforward to show that
if countries cannot withdraw from the negotiations (as assumed by BH), then also in our model, there is no equilibrium in which more than \( k \) countries sign a long-term agreement. As pointed out by BH, this is due to countries’ incentives to free-ride on short-term agreements.

7 Limited commitment

In our dynamic model, we assume that once an agreement is signed, it will last forever. Such an assumption relies on countries’ possibility to commit. BH also assume that countries can commit for infinitely many periods, if they choose to do so. KS, by contrast, assume that countries cannot commit for more than one period. This begs the question, if the assumption of full commitment is essential for our results to hold, or not. In this section, we therefore investigate the role of commitment for our results. We consider a modified version of our dynamic coalition formation game (introduced in Section 3), where we allow the countries to dissolve an existing coalition (similarly as in KS). Hence, an agreement that is signed can last forever, if it is never dissolved by its members. The specification of the per period game differs depending on whether there is an active agreement or not. For this we define two regimes: (N) and (A).

Regime (N): There is no active agreement.\(^{47}\) In this case, the stages within period \( t \) are the same as before, with a participation stage followed by a signature stage (see Section 3, Figure 1). This applies both under the random membership approach, as well as under the deterministic membership approach (see Section 6). If an agreement is signed in period \( t \), the abatement levels are chosen cooperatively by the coalition members (with non-signatories behaving non-cooperatively), and the game enters the next period in regime (A). If no agreement is reached in period \( t \), then there is delay and the game enters the next period again in regime (N); in this case, all countries choose their abatement levels non-cooperatively in the current period (payoff \( \pi_0 \)).

Regime (A): There is an active agreement. In this case there is no participation stage, only the signature stage that is now interpreted as a review stage, where each coalition member (signatory of the active agreement) decides whether to stay or leave the coalition. If any of its members leaves the coalition, it is dissolved and countries behave non-cooperatively in the current period (payoff \( \pi_0 \)). There are no further negotiations in the current period and the game enters the next period in regime (N). If no signatory leaves the coalition, the payoffs are realized according to the abatements as specified

\(^{47}\)KS use a reduced-form model and assume that a climate coalition is always formed (or maintained) in each period. This implies that Regime (N) cannot occur after period 1. By contrast, we explicitly model countries’ participation decisions in a standard game-theoretic way, which means that we allow for deviations regarding countries’ participation decisions. On the equilibrium path, of course, deviations do not take place.
under the agreement (with non-signatories behaving non-cooperatively), and the game enters the next stage in regime (A).

Such a modification yields a richer strategy space than in our main setup as introduced in Section 3. We again consider Markov perfect equilibria and will argue that the equilibria characterized in Section 3 continue to exist (under the same conditions) when adjusted to the new setup. More specifically, under the Markov restriction, countries’ strategies depend now on the regime, and in regime (N) also on the size of an existing coalition. Fix an equilibrium from the original model (Section 3) where \( k^* \) countries sign an agreement. We will argue that the following establishes an equilibrium, where \( k^* \) countries sign an agreement that will last forever, in the modified game: In regime (N), using the public randomization device, \( k^* \) countries form a coalition and sign an agreement. In regime (A), each coalition member stays if the coalition has size at least \( k^* \) and leaves otherwise. Under this strategy profile, the countries anticipate that there will be an agreement signed among \( k^* \) countries in the next stage if entered in regime (N) and the coalition of \( k^* \) countries will not be dissolved in the next stage if entered in regime (A).

Consider first a period in regime (N). If a coalition forms in that period, and its members sign an agreement, then they do so anticipating, that this coalition will (on the path) be maintained also in all future periods. Thus, the countries face the same trade-offs as in the model specified in Section 3. In particular, the conditions (5)–(6) describe the case where no country has incentives to deviate from the equilibrium behavior. Second, consider a period in regime (A). The trade-offs involved in the decision of the coalition members of whether or not to maintain this agreement or dissolve it, are the same as the trade-offs that the members of a coalition (of the same size and with the same country identities) face in regime (N) at the signature stage: if they sign an agreement, their abatement levels are chosen cooperatively in this period, and the next period is entered in regime (A); if they don’t sign an agreement, abatements are chosen non-cooperatively, and the next period is entered in regime (N). We thus conclude, that the equilibria characterized in Section 3 for the model with long-term commitment, continue to exist in the modified model with single-period commitment, under the same conditions as specified in Section 3.48

Let us briefly consider also the case where countries can sign a short-term agreement, in a period in which negotiations about a long-term agreement have failed (see Section 5). We now argue that it is possible to extend the model with single-period commitment also in a way, that encompasses our results and insights from that version of our model. Let us again extend the game with single-period commitment by two more stages, if no

48Since the strategy space is richer in the modified game without long-term commitment, we cannot rule out the possibility that other equilibria may exist. A full characterization of equilibria in this modified setup is beyond the scope of this paper.
agreement is reached or an agreement is dissolved in the current stage, i.e., if the game is supposed to enter the next stage in regime (N). The two additional stages are the same as in Section 5: one stage where coalition members can drop out if negotiations about a long-term agreement fail, and a signature stage where the remaining coalition members can decide whether or not to sign a short-term agreement which lasts for a single period only. Note that the difference between a “long-term agreement” and a “short-term agreement”, however, is now rather subtle. The “long-term agreement” is dissolved only if its members decide to in the next (or some future stage), so that the next period starts in regime (A). The “short-term agreement” is dissolved automatically after one period, so that the next period starts in regime (N).

In this modified setup, it is clear that any agreement on a short-term climate contract in the current period (after negotiations about a long-term agreement have failed), does not affect what happens in any future period. Therefore, as discussed in Section 5, if a short-term agreement is signed, it coincides with the outcome in the static case. Apart from this, the trade-offs in the overall game are as discussed above. We conclude that also our results in Section 5 do not depend on the assumption of long-term commitment.

It is important to point out that, although the equilibria are preserved under the model with single-period commitment introduced above, this might not be the case in all alternative specifications. For instance, if new negotiations about a new agreement can then start immediately (as in KS), our original equilibrium might cease to exist and the outcome of the dynamic game collapses to the outcome of the static game. To see this, notice that in any period, countries will sign some agreement, as the agreement can be dissolved again “without cost” (since there us no delay) in the beginning of the next period. But if countries always sign some agreement, even an agreement with few members, then an individual country that is initially “supposed to be” in the coalition, can drop out, while the remaining coalition members still sign an agreement in that period. Hence, the “usual” free-rider incentive well-known from the static model reappears. In our full dynamic game (with long-term commitment), this is not the case, because if an individual country drops out, the remaining coalition members do not sign an agreement, in order to avoid locking in an inefficient agreement. Thus, it is rather the opportunity cost of signing (and preserving) a weak agreement today (thereby giving up a possibly better agreement tomorrow) that makes small agreements unstable in our framework.  

8 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-

---

49We are grateful to Hiroaki Sakamoto for pointing this out.
tations 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.

The main policy conclusion from our analysis is, that leaving the door open to future negotiations in case current negotiations fail, may help to achieve more effective agreements (with higher participation). Furthermore, in contrast to BH’s conclusion that the length of the commitment period should be determined during the negotiations and not ex ante fixed in order to achieve a better outcome, our results point in the opposite direction: a firm commitment (if such is feasible) to only negotiate about a long-term contract, thereby fixing the length of the commitment period ex ante, may increase the chances of reaching a successful agreement. Then short-term agreements that can potentially undermine long-term cooperation (as emphasized also by BH), are effectively ruled out. Although our results also show that the possibility to sign short-term agreements can actually foster participation in a long-term agreement, this conclusion is quite sensitive to the details of how short-term agreements are modeled (negotiated).

---

50 On page 163 of their paper, they write: “Second, it is important to let the final coalition negotiate the duration of the agreement rather than announcing a length before countries have fully committed on whether or not to join.”
References


Appendix: Proofs

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

$$V(k_0) = \Pi_s(k_0) \leq \frac{\pi_0}{1 - \delta} < \Pi_s(N) = V(N). \quad (22)$$

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

$$\Pi_s(k_0) \leq \pi_0 + \delta V(k_0) \leq \pi_0 + \delta V(k) \leq \pi_0 + \delta V(N) < \Pi_s(N). \quad (23)$$

The first and the last inequalities follow from (22), and the second and the third inequalities follow from the monotonicity of $V$. Continuity and monotonicity of $\Pi_s$ 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, $\Pi_s$ and $V$ are strictly increasing. Hence, also the inverse function $\Pi_s^{-1}$ is increasing, which shows that $\xi$ is increasing.

Finally, the property $\xi(k) \in [k_0, N)$ follows directly from (23) and the definition of the function $\xi$.

Proof of Lemma 3. Since the function $\xi$ is increasing, it follows from the inequalities in Lemma 1 that $\xi$ 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 $\eta$ 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 $\xi$ being strictly increasing (Lemma 1), and the third one from Lemma 1. 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 1. Because $N$ is an integer, $\hat{\xi}(N) \leq N$, which completes the proof.

Proof of Proposition 4. Recall that $\xi$ is monotonically increasing (Lemma 1) and that $k$ is now a fixed point of $\xi$ (due to the assumption $k > k_0$). Moreover, by Assumption 5(a), its slope at $k$ is smaller than 1. Thus, it is sufficient to show that the value $\xi(k)$ increases in some neighborhood of $k$, when $\delta$ or $\pi_0$ increases.

Let us now rewrite $\xi(k)$ (as given in (3)) using the functions $\tau_s(k) = (1 - \delta)\Pi_s(k)$ and
\( v(k) \equiv (1 - \delta)V(k) \) that correspond to per period payoffs and thus do not depend on \( \delta \):

\[
\pi_s(\xi(k)) = (1 - \delta)\pi_0 + \delta v(k),
\]

or equivalently, \( \xi(k) = \pi_s^{-1}((1 - \delta)\pi_0 + \delta v(k)). \) (24)

If follows from Assumption 2 that the function \( \pi_s \) is increasing. Thus, its inverse \( \pi_s^{-1} \) is increasing as well. This immediately implies that \( \xi(k) \) increases when \( \pi_0 \) increases (for all \( k > k_0 \)).

Moreover, (24) implies that \( \xi(k) \) increases when \( \delta \) increases for all \( k \) such that \( v(k) > \pi_0 \). In order to complete the proof, it remains to show that \( v(k) > \pi_0 \). 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 \( v(k) > \pi_s(k) > \pi_0 \).

Proof of Lemma 4. (i) Let us denote \( X^*(k) = kx^*_n(k) + (N - k)x^*_s(k) \). 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 (9), 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) Let \( \hat{x}_n(x_s, k) \) denote 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). Then \( \hat{x}_n(x_s, k) \) is implicitly defined by (9), 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 \( x_s \):

\[
B''(X) \left[ x_s - x_n + (N - k) \frac{\partial \hat{x}_n}{\partial k} (x_s, k) \right] = C''(x_n) \cdot \frac{\partial \hat{x}_n}{\partial k} (x_s, k). \tag{25}
\]

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

\[
\frac{\partial \hat{x}_n}{\partial k} (x^*_s(k_0), k_0) = 0. \tag{26}
\]

Now recall that the coalition chooses \( x_s \) in order to maximize \( B(X) - C(x_s) = B(kx_s +
\((N - k)\hat{x}_n(x_s, k)) - C(x_s)\). It follows from the Envelope theorem and from (25) that

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

Since, for \(k = k_0\) we have \(x_s^* = x_n^*\) (by (i)), it follows from (26) that indeed \(\pi_s'(k_0) = 0\).

**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, \(k^* \leq k_{st}\). 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 (17) 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, (17) 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 \(k > k_{st}\), or equivalently \(k_{st} < \xi(k_{st})\). Recall that we now replace \(\pi_0\) 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 (1) 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 $k > k^{st} + 1$. For $k_{t+1} = k_t = k^*$, condition (18) 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:

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

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 $k > k^{st} + 1$. \hfill \square

### A.1 Additional results

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

**Proof of Lemma 5.** 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

$$\delta p(k) \left(\Pi_s(k) - \frac{\pi_0}{1 - \delta}\right) + \delta \left[1 - p(k)\right] \left(\Pi_n(k) - \frac{\pi_0}{1 - \delta}\right) > \Pi_s(k) - \frac{\pi_0}{1 - \delta},$$

$$p(k) + \left[1 - p(k)\right] \frac{\Pi_n(k) - \pi_0/(1 - \delta)}{\Pi_s(k) - \pi_0/(1 - \delta)} < \frac{1}{\delta}.$$ (28)

Since $\Pi_n(k) > \Pi_s(k)$ for $k > k_0$, we have $\varphi(k) < 1$. Therefore, $p(k) + \left[1 - p(k)\right] \varphi(k) < 1 < 1/\delta$ which implies that (28) 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 (28). 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 (28) 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 $\overline{k}$. Otherwise, we set $k = k_0$. In both cases we obtain that $\xi(k) > k$ when $k_0 \leq k < \overline{k}$, while $\xi(k) < k$ when $\overline{k} < k \leq N$. \hfill \square
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 2 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 (4). 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 (7) or equivalently, $k^* = \xi(k^*)$, with redefined $V$ and $\xi$ 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 $\Pi_s(k) = \pi_0/(1 - \delta)$. Since, $\Pi_s(k)$ is increasing for $k > k_0$, we indeed obtain that $k$ is increasing in $\delta$ and in $\pi_0$ when $k > k_0$. 

B Supplementary Appendix (for online publication)

B.1 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 \( \pi_0 \) 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.

In order to facilitate the comparison to the game with infinite negotiations, we consider symmetric subgame perfect equilibria with no delays.\(^{51}\) 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 \( \bar{k} = 5.4 \) and \( \tilde{k} = 8.2 \) 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 \( k^*_{T-1} = 3 \) countries to sign an agreement. Recall from Lemma 2 (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

\(^{51}\) Equilibria with delays are characterized in Section B.2.
\( \hat{\xi}(k_T^*) = \hat{\xi}(3) = 4 \). 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 (see Appendix B.3 for a proof).

**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 \( k^{st} \geq k \).

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

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)

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 a graphical illustration). This ratcheting-up stops when the maximum coalition size is reached, that coincides with the smallest stable coalition size (\( \lceil k \rceil \)) under an infinite time horizon for the negotiations (in the case \( \lceil k \rceil > k^{st} \)). Hence, the multiplicity of equilibria
that we observed in the infinite horizon case (see Figure 2) vanishes.\textsuperscript{52} Otherwise, the results are unchanged.\textsuperscript{53}

B.2 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.1 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 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 $\Pi_s$ and $\Pi_n$, 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 $k$) 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).\textsuperscript{54} In that respect, the countries may use only their non-participation, but not the use of veto power (during the signature stage).

\textsuperscript{52}Here, we refer to the multiplicity of equilibrium coalition sizes that arises when the interval $[k, \hat{k})$ 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.2 for further details).

\textsuperscript{53}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 $k$ signs an agreement in the first period under deterministic membership.

\textsuperscript{54}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 2 still applies.
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 \( \Pi_s \) and \( \Pi_n \) 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).\(^{55}\) Nevertheless, even under this simplifying assumption, subgame perfect Nash equilibria (SPNE) can exist that exhibit delay along the equilibrium path.\(^{56}\)

To see this, suppose the payoff functions \( \Pi_s \) and \( \Pi_n \) are such that the static model exhibits a non-trivial amount of cooperation in equilibrium, that is: \( k^{st} \geq 2 \). Then by 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 \( k^*_{\tau} \) countries in period \( \tau \), 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 \( \tau \)). 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,

\(^{55}\)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.

\(^{56}\)We focus on SPNE whenever the Markov restriction is relaxed.
extends readily towards earlier periods, so that if $k^*\tau$ and $\delta$ are sufficiently large, no agreement is signed in any period $t < \tau$ even when $\tau$ is a large number, assuming that countries adopt such delay strategies.\footnote{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.} 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. The following proposition reveals that relaxing the Markov assumption does not support larger coalition sizes (see Appendix B.3 for a proof).\footnote{Note, that if $k > k^{st}$, then the inequality in Proposition 8 simplifies to $k^* < \bar{k}$.}

**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 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.3 Proofs for Appendix B**

*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 \( k^{st} < k \). Consider the following sequence defined recursively:\(^{59}\)

\[
  l_0 = k^{st} = \lceil \hat{k} \rceil, \quad l_\beta = \hat{\xi}(l_{\beta-1}) \quad \text{for} \quad \beta = 1, 2, \ldots, (29)
\]

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

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, \( k^*_T = \lceil \hat{k} \rceil = l_0 \). For any \( t \leq T - 1 \), if the countries in period \( t \) anticipate that \( k^*_t = l_{T-t} \) countries sign an agreement in the next period, then according to Lemma 2, \( k^*_t \geq \hat{\xi}(k^*_{t+1}) = \hat{\xi}(l_{T-t-1}) = l_{T-t} \). 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 \( k^*_t = \xi(k^*_{t+1}) \) or \( k^*_t > \xi(k^*_{t+1}) \). We show that the former case applies. Otherwise, if \( k^*_t > \xi(k^*_{t+1}) \), 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 \( k^*_t > \xi(k^*_{t+1}) = \hat{\xi}(l_{T-t-1}) = l_{T-t} \geq l_0 = \lceil \hat{k} \rceil = k^{st} \). Thus, indeed the former case applies, which yields \( k^*_t = \xi(k^*_{t+1}) = \hat{\xi}(l_{T-t-1}) = l_{T-t} \).

In order to complete the proof of (ii), it is sufficient to set \( T > \tau + 1 \), where \( \tau \) is introduced in Lemma 6. Then \( k^*_1 = l_{T-1} = l_\tau = \lceil \hat{k} \rceil \). \( \square \)

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

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

Next, we show that \( l_\beta = l_{\beta-1} \) when \( l_{\beta-1} = \lceil \hat{k} \rceil \). Since \( l_{\beta-1} \) is a positive integer and \( l_{\beta-1} \in [\hat{k}, \lceil \hat{k} \rceil) \), 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 \( l_0 = k^{st} < \lceil \hat{k} \rceil \), the sequence \( l_0, l_1, l_2, \ldots \) is bounded from above by \( \lceil \hat{k} \rceil \) and is increasing before attaining this bound. Let us set \( \tau \) such that \( l_{\tau-1} < \lceil \hat{k} \rceil = l_\tau \). Then \( l_\beta = l_\tau = \lceil \hat{k} \rceil \) for \( \beta \geq \tau \), which completes the proof. \( \square \)

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 \( k^{max} > \max\{k^{st}, \lceil \hat{k} \rceil - 1\} \), which is equivalent to \( k^{max} \geq \max\{k^{st} + 1, \hat{k}\} \).

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

\(^{59}\)We use the subscript \( \beta \) for counting backwards in time (see below).
the coalition for some member. If the remaining \( k_t - 1 \) coalition members do not sign an agreement, then \( V(k_{\text{max}}) \) 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 \( \pi_0 + \delta V(k_{\text{max}}) \). However, since \( k_t = k_{\text{max}} \geq \bar{k} \), we have \( \Pi_s(k_t - 1) \geq \pi_0 + \delta V(k_{\text{max}}) \) 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 \( k_t = k_{\text{max}} > k^* \) violates internal stability.