Sakamoto, Hiroaki; Traeger, Christian

Working Paper

Self-enforcing Stable Sets

CESifo Working Paper, No. 12360

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Sakamoto, Hiroaki; Traeger, Christian (2025) : Self-enforcing Stable Sets, CESifo Working Paper, No. 12360, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/336060

CES ifo Working Papers

12360
2025

December 2025

Self-enforcing Stable Sets

Hiroaki Sakamoto, Christian Traeger

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Self-enforcing stable sets*

Hiroaki Sakamoto Christian Traeger

Version of December 2025

Abstract

We study coalition formation with externalities under voluntary, non-binding participation. Motivated by climate agreements, where standard modeling predicts small, inefficient coalitions, we propose a new solution concept—the self-enforcing stable set. It synthesizes the self-enforcing logic of non-cooperative approaches with the consistency requirement of cooperative forward-looking stability. By endogenizing players' beliefs about the eventual outcomes of negotiations, we show that rational foresight disciplines strategic free-riding and selects constrained Pareto efficient outcomes. In canonical climate-agreement models, this yields sharp predictions: stable coalitions must be large and only mildly fragmented, aligning closely with observed participation patterns.

Keywords: coalition formation; self-enforcing agreements; international agreements; public goods; climate change.

JEL codes: C71, F53, H41, Q54

*Sakamoto: Graduate School of Economics, Kobe University; contact: sakamoto@econ.kobe-u.ac.jp. Traeger: Department of Economics, University of Oslo, Norway & ifo Institute, Munich, Germany; contact: traeger@uio.no. Sakamoto acknowledges funding from Sumitomo Foundation and MEXT under grant 24K00258 and 24K21416. Traeger gratefully acknowledges funding from the Norwegian Research Council under grant 315878 (NIMICAR). We are grateful to Ken-Ichi Akao, Geir Asheim, Bard Harstad, Takashi Kamihigashi, Eiichi Miyagawa, Noritsugu Nakanishi, and Takashi Shimizu for comments and suggestions.

1 Introduction

The Coase theorem suggests that, when transaction costs are sufficiently small, bargaining among affected parties should resolve externality problems: rational agents ought to negotiate toward Pareto improvements whenever such gains exist.[1] Yet, as Dixit and Olson (2000) emphasize, the classic argument presumes that no party abandons negotiations to free ride. In environments where agreements must be voluntary and non-binding, this presumption is at odds with the strategic realities of cooperation. Without external enforcement, participants can always reconsider their commitments, and potential beneficiaries may have incentives to stand aside. As a result, Coasian bargaining is often regarded as ineffective in self-enforcing environments, and the prevailing view in the literature is that such settings inevitably fail to correct externalities, leading to small coalitions and inefficient agreements.

This paper reexamines that pessimistic conclusion. We study a coalition-formation game with externalities in which players can voluntarily join or leave an agreement, and where transfers or side payments are not enforceable. The framework is deliberately general and can be applied to a wide range of free-rider problems, but our main application is climate change. As Nordhaus (1977) put it nearly half a century ago, the problem of climate change is the "most extreme imaginable form of external diseconomy" and can only be solved through agreements among sovereign countries. In the absence of a supranational enforcement agency, such agreements must be self-enforcing: no country can be forced to participate, and every country reserves the right to leave at will. A general conclusion in the literature is that any stable agreement in such a self-enforcing environment involves substantial inefficiency, often with only a few countries remaining in the agreement. The predictions of our analysis are less pessimistic and more consistent with empirical evidence that some self-enforcing agreements attract many participants and manage to mitigate transboundary pollution (Breitmeier et al., 2006; Young, 2011; UNFCCC, 2016).

An important component of our analysis is farsightedness. We suggest that negotiating parties form rational expectations about the outcomes that will *eventually* emerge from the negotiation. As Aumann and Myerson (1988) emphasize, when a player or a group of players deviates, the actions of others should not be taken as given. Even a single player hinting at a change in participation can trigger reactions from insiders and outsiders alike, potentially setting off a cascade of further adjustments. Such responsiveness is characteristic of open-ended negotiation environments like climate agreements, where participation is fluid rather than a one-shot, irreversible decision. Hence, when contemplating a deviation, players must evaluate not only the immediate payoff change but also the equilibrium outcomes that

[1] Coase (1960) himself emphasizes the likely failure to reach efficient agreements due to transaction costs. In our primary application, the regular high-level meetings on climate change and the establishment of a permanent UN entity dedicated to finding solutions suggest that transaction costs are small compared to the stakes at the intergovernmental level. Our general framework accommodates transaction costs without qualitative changes to the results.

could ultimately prevail. We show that in self-enforcing environments, such rational foresight disciplines free riding and thereby places an effective bound on equilibrium inefficiency.

Our theoretical framework integrates two strands of literature on coalition formation. The first strand, focused on *self-enforcing agreements*, employs a non-cooperative participation game originally developed by d'Aspremont et al. (1983) and Palfrey and Rosenthal (1984) in industrial organization (see Finus (2001) and Barrett (2005) for surveys of the early literature, and Carattini et al. (2019) and Harstad (2024) for reviews of recent developments). In this framework, players decide independently whether to join a coalition, anticipating that subsequent actions and payoffs are determined by the resulting coalition structure (i.e., coalition size, composition, and membership status).[2] The framework was quickly adopted in the analysis of international environmental agreements by Hoel (1992), Carraro and Siniscalco (1993), and Barrett (1994), among others, and has evolved to be the literature's workhorse model. While this non-cooperative approach captures important aspects of reality, its predictions are primarily driven by myopic expectations. Potential defectors optimistically expect that they can reap the benefits of free riding without triggering reactions from others, whereas potential participants pessimistically expect that, if they join the coalition, nobody will follow suit.[3]

The second strand of literature is the cooperative game theory of coalition formation (see Ray (2007) and Ray and Vohra (2015a) for general overviews). At the core of this literature lies a *consistency condition*, which requires that any alternative to an existing agreement must itself be stable; if an alternative agreement is not stable, it does not constitute a valid objection. Such a consistency condition originates with von Neumann and Morgenstern's (1944) concept of stable sets (vNM stable sets), which partitions outcomes into two classes: stable outcomes, which are not "dominated" by other stable outcomes, and unstable outcomes, which must be "dominated" by some stable outcome. Harsanyi (1974) and Ray and Vohra (2015b), among others, extend the general concept of stable sets and refine the definition of "domination" to better capture farsightedness. Although conceptually appealing, the idea of stable sets has rarely been adopted in applied studies.[4] A primary reason for this limited adoption is that stable sets are hard to characterize and their existence is not guaranteed, especially when players' payoffs depend on the entire coalition structure due to

[2]Ray and Vohra (1997) adopt the same two-stage procedure, although their paper is not typically included in this literature.

[3]The opposite extreme is the assumption that any deviation from a coalition triggers the collapse of the coalition. Under such an assumption, the grand coalition becomes easy to stabilize (Chander and Tulkens, 1995, 1997; Germain et al., 2003). A notable application of this reasoning to climate change is Gerber and Wichardt (2009).

[4]Notable exceptions are Diamantoudi and Sartzetakis (2015, 2018), who apply (farsighted) stable sets of Harsanyi (1974) to the analysis of international environmental agreements based on two variants of effectivity set. Diamantoudi and Sartzetakis (2015) allow potential defectors to coordinate their actions whereas Diamantoudi and Sartzetakis (2018) only allow for unilateral deviations. Both papers rely on a particular model structure with symmetric players to ensure the existence and tractability of solutions. Their relatively optimistic findings follow from a threat of complete collapse of an agreement (Diamantoudi and Sartzetakis, 2015) or a domino effect of consecutive defections (Diamantoudi and Sartzetakis, 2018).

cross-coalition externalities. For example, a typical participation game of international environmental agreements can have no vNM stable sets. Ensuring the existence and tractability of stable sets often requires restrictive assumptions on the model structure.

These two strands of literature, though active for decades, have remained largely separate. They even use crucial terminology—internal and external stability—with different meanings. The present paper provides a bridge between the two by developing a new solution concept that synthesizes key features from each tradition. On the one hand, drawing on the non-cooperative literature, our solution concept incorporates the self-enforcing nature of agreements; stable coalitions must be robust against unilateral deviations by both insiders and outsiders. On the other hand, we follow the cooperative literature in imposing a consistency requirement; a coalition's stability can only be undermined by an alternative coalition that can eventually prevail as a stable outcome. We synthesize these ideas in a general framework and our core insights do not depend on a particular model structure. Unlike many existing approaches in the literature, our solution concept leads to equilibrium outcomes that are relatively easy to characterize, as we demonstrate using canonical models of international climate cooperation.

A crucial step in synthesizing these key features is endogenizing players' expectations under strategic uncertainty. In complex, unstructured negotiations where negotiation protocols (who moves when, in what order, and with what information) are not clearly defined, the precise sequence of moves following a deviation is inherently unpredictable. Rational players anticipate that unsettling the status quo will ultimately lead to *some* stable outcome, but they cannot know which one. The existing literature typically resolves this uncertainty by imposing behavioral assumptions based on deterministic reaction chains. One common approach assumes optimism on the part of the defector: players deviate whenever there exists one reaction path that makes them better off, even if other plausible paths do not (Harsanyi, 1974). By contrast, Chwe (1994)'s largest consistent set assumes pessimistic players who deviate only if every plausible path makes them better off. As Ray and Vohra (2015a) show with simple examples, neither approach is satisfactory, especially with externalities.[5] We propose an alternative approach that directly addresses this strategic uncertainty by introducing probabilistic beliefs—a probability distribution over the set of potential stable outcomes. These beliefs allow players to evaluate the expected payoff from a deviation, which in turn determines the set of stable outcomes. Crucially, we endogenize these beliefs rather than imposing them exogenously. Equilibrium beliefs are defined as a fixed point: the set of outcomes players believe to be stable must be the set of outcomes that are stable given those beliefs. This consistency requirement disciplines players' expectations, moving beyond simple optimism or pessimism.

Our main characterization shows that a coalition structure is stable if and only if it

[5]See Dutta and Vohra (2017) and Ray and Vohra (2019) for recent refinements of plausible deviation paths.

achieves constrained Pareto efficiency, i.e., the efficiency attainable without transfers. In self-enforcing settings, negotiations among farsighted players only settle when everyone believes that the current arrangement is the unique stable outcome; if multiple outcomes were believed possible, someone would always deviate from their least-preferred option. Hence, in equilibrium, beliefs must converge on a single negotiation endpoint. Once such disciplined beliefs emerge, myopic free riding becomes unprofitable: a transient gain from defection is understood as futile because there is no rational basis for expecting negotiations to settle in the defector's favor. Consistency, the understanding that only stable equilibria constitute credible threats of deviating, then drives the outcome toward efficiency. If a Pareto-improving arrangement still exists, rational players recognize that the inefficient status quo is untenable as the unique endpoint of negotiations. Thus, the process continues until all constrained Pareto improvements are exhausted, with any residual inefficiency arising solely from institutional constraints that preclude enforceable transfers.[6]

Taken together, our analysis contributes to the literature on self-enforcing agreements along several dimensions. First, we offer a tractable benchmark that brings farsightedness into the standard participation-game approach: the self-enforcing stable set combines stability against unilateral deviation with a cooperative consistency requirement on long-run negotiation outcomes, while remaining simple to apply in general environments with externalities.[7] Second, we provide a general characterization and existence result: stable outcomes coincide exactly with constrained Pareto efficient coalitions (and, in our extension, coalition structures), yielding a transparent description of what self-enforcing negotiations can achieve. Third, this benchmark delivers sharp and empirically plausible implications in canonical models of climate cooperation—predicting high participation and limited fragmentation—and our second-order refinement further sharpens equilibrium predictions without sacrificing tractability.[8]

In Section 2, we describe the model and briefly review the two solution concepts com-

[6]Section 3.2 relates this characterization to prior results.

[7]We view the simplicity and abstraction from negotiation details as a key strength of our approach. That said, in Appendix A, we also provide an alternative theoretical underpinning by presenting a sequential bargaining game in which the stationary subgame-perfect equilibria coincide with our proposed solution. The model is not intended to replace full-scale dynamic approaches but to offer a complementary perspective on coalition formation. A recent contribution by Vosooghi et al. (2024) uses a sequential bargaining framework within a calibrated integrated assessment model to study climate coalitions under farsighted behavior, providing an empirically grounded complement to our theoretical analysis.

[8]There is a growing literature that revisits international climate cooperation using modern tools from political economy, contract theory, dynamic games, and trade policy. This body of work has yielded important insights by analyzing climate cooperation through a range of complementary lenses, including dynamic technology investment (Battaglini and Harstad, 2016; Harstad, 2016; Harstad et al., 2019), domestic political economy (Battaglini and Harstad, 2020; Harstad and Kessler, 2025), international trade (Nordhaus, 2015; Böhringer et al., 2016; Kortum and Weisbach, 2024; Farrokhi and Lashkaripour, 2025; Iverson, 2025), bargaining protocols (Harstad, 2023a,b), and transfer schemes (Okada, 2023; Dutta and Radner, 2025; Kerr et al., 2025). Our analysis complements this literature by abstracting from particular institutional dimensions in order to establish a general theoretical benchmark for self-enforcing climate cooperation under farsighted behavior.

monly used in the literature. We then present our new solution concept, the self-enforcing stable set, and explain how it relates to the existing approaches. Section 3 characterizes our solution in a general setting, relates it to the literature, and discusses several applications. To further sharpen predictions, Section 4 introduces a refinement that we refer to as second-order stability. Section 5 extends the model to a setting with multiple coexisting coalitions.

2 The model

This section explains the setup, reviews existing approaches, and formalizes our solution concept.

2.1 Setting

Let $N = \{1, 2, \ldots, n\}$ be the set of $n \geq 3$ players; players can be asymmetric. The set of all nonempty subsets of N, denoted by $\mathcal{N}$, represents the set of all possible coalitions. For now, we assume that only a single coalition can form; we relax this assumption in Section 5. Once a coalition forms, it uniquely determines the payoffs for all players, both insiders and outsiders.[9] These payoffs, which may incorporate transaction costs, are represented by reduced-form functions $u_i : \mathcal{N} \to \mathbb{R}$ for each $i \in N$. Notice that, by mapping each coalition to a unique payoff vector, we exclude the possibility of transfers or reallocation of surplus among players. Our examples below show how these reduced-form payoffs can be interpreted as the outcomes of an underlying game. Table 1 summarizes the key notation used throughout the paper.

We illustrate core concepts using two common examples from the international environmental agreement literature. In these examples, n countries emit a global pollutant and each country's emission level g_i depends on the prevailing coalition structure. Coalition members coordinate to maximize their joint payoff, whereas non-members choose emissions to maximize their individual payoffs.[10] The following examples introduce the reduced-form payoff function u_i for such settings.

Example 1. Player i's underlying utility consists of private benefits from his or her own

[9] Throughout, we use the term payoff in the game-theoretic sense, i.e., the objective each player maximizes. In our examples, payoffs correspond to the utility derived from emissions and climate.

[10] As Battaglini and Harstad (2016) show, the assumption of joint payoff maximization can be interpreted as an equilibrium outcome of a within-coalition bargaining. Gersbach and Winkler (2011) design a refunding scheme that sets within-coalition incentives for joint welfare maximization under a polluting externality (conditional on participation). Alternative behavioral assumptions are also possible as long as players' payoffs are uniquely determined for each coalition. One could assume, for example, that the coalitional surplus is divided among its members based on the Shapley value. See Myerson (1977), Hart and Kurz (1983) and Aumann and Myerson (1988) for such specifications.

Table 1: Key notation used throughout the paper

Symbol	Description
$N := \{1, 2, \ldots, n\}$	Set of $n \geq 3$ players
$\mathcal{N}$	Set of all nonempty subsets of N (possible coalitions)
$u_i : \mathcal{N} \to \mathbb{R}$	Reduced-form payoff function of player i
$M \sim M'$	Coalitions M and M' are payoff-equivalent
$\lvert M \rvert$	Cardinality of coalition M
$\mathcal{M} \subset \mathcal{N}$	Generic subset of possible coalitions
$\tilde{M} := \{M' \in \mathcal{N} \mid M' \sim M\}$	Indifference class of coalitions payoff-equivalent to M
g_i	Emission level of country i
$\gamma \geq 2$	Convexity parameter of abatement cost
$\xi > 0$	Marginal cost of externality
Coalition Structures (Section 5)	
$\boldsymbol{\mathcal{N}}$	Set of all partitions of N (coalition structures)
$\boldsymbol{M} = \{M_1, M_2, \ldots, M_L\}$	A coalition structure with L coexisting coalitions
$\lvert \boldsymbol{M} \rvert = L$	Number of coexisting coalitions in structure $\boldsymbol{M}$
$M_l \subset N$	Members of the l-th coalition
$\boldsymbol{N} := \{N\}$	Grand coalition structure

emissions and a negative externality from aggregate emissions:

$$\underbrace{-\frac{1}{\gamma}(\bar{g}_i - g_i)^\gamma}_{\text{private benefit}} \underbrace{- \xi \sum_{j \in N} g_j}_{\text{externality}},$$

with $\gamma \geq 2$ and $\xi > 0$. The private benefits are maximal under the "business-as-usual" (BAU) emission level, $\bar{g}_i$, and decline as the player engages in mitigation by choosing $g_i < \bar{g}_i$.[11] We constrain emissions such that $0 \leq g_i \leq \bar{g}_i$ and assume $\bar{g}_i \geq n^{\frac{1}{\gamma-1}} \xi^{\frac{1}{\gamma-1}}$ so that equilibrium emission levels remain non-negative. The parameter γ characterizes the convexity of abatement costs, while ξ represents the marginal damage from pollution (e.g., climate damage per ton of carbon).[12] Given that coalition members maximize their joint payoffs and non-members maximize their own payoffs, this game yields the following

[11] Strictly speaking, $\bar{g}_i$ denotes the BAU emission level in a continuum-player setting, or when players fail to account for the harm caused by their own emissions. In our discrete-player model, even free riders internalize this self-inflicted damage and optimally choose $g_i = \bar{g}_i - \xi^{1/(\gamma-1)}$. We note that private benefits of emissions are equivalent to private abatement costs.

[12] More generally, ξ captures marginal damages relative to abatement costs as we have normalized the multiplicative constant on the benefit term to one.

reduced-form payoff function (derived in Appendix C.1):

$$u_i(M) = \begin{cases} \xi^{\frac{\gamma}{\gamma-1}} \left(|M|^{\frac{\gamma}{\gamma-1}} - |M| + n - \frac{1}{\gamma}|M|^{\frac{\gamma}{\gamma-1}} \right) - \xi \sum_{j \in N} \bar{g}_j & \forall i \in M \\ \xi^{\frac{\gamma}{\gamma-1}} \left(|M|^{\frac{\gamma}{\gamma-1}} - |M| + n - \frac{1}{\gamma} \right) - \xi \sum_{j \in N} \bar{g}_j & \forall i \notin M, \end{cases} \tag{1}$$

for each $M \in \mathcal{N}$, where $|M|$ is the coalition's cardinality (number of coalition members). The most widely used specification in the literature corresponds to $\gamma = 2$, delivering a linear-quadratic problem.

Example 2. Another canonical model, common in climate economics (Golosov et al., 2014; Hassler et al., 2016), specifies each player's utility as

$$\underbrace{\ln(g_i)}_{\text{private benefit}} - \xi \underbrace{\sum_{j \in N} g_j}_{\text{externality}}$$

for some $\xi > 0$. In this specification, benefits increase logarithmically with emissions, whereas damages remain linear in aggregate emissions. Applying the same behavioral assumptions as before yields the following reduced-form payoff (see Appendix C.1):

$$u_i(M) = \begin{cases} -\ln(\xi) - n - 1 + |M| - \ln(|M|) & \forall i \in M \\ -\ln(\xi) - n - 1 + |M| & \forall i \notin M. \end{cases} \tag{2}$$

2.2 Solution concepts

We now turn to the solution concept of the game. We begin by reviewing the standard equilibrium concept for self-enforcing agreements and then turn to von Neumann-Morgenstern stability. Our approach combines key elements of both concepts.

The standard approach in the self-enforcing agreement literature treats the formation of agreements as a one-shot participation game (d'Aspremont et al., 1983). The corresponding solution concept can be formalized as follows:

Definition 2.1 (One-shot stability)**.** A nonempty subset $\mathcal{M} \subset \mathcal{N}$ is a one-shot stable set if

$$M \in \mathcal{M} \iff \begin{cases} u_i(M) \geq u_i(M \setminus \{i\}) & \forall i \in M \quad \text{and} \\ u_i(M) > u_i(M \cup \{i\}) & \forall i \notin M. \end{cases} \tag{3}$$

The first inequality in condition (3) is what this strand of literature refers to as the *internal stability* condition; no member of a stable coalition has an incentive to leave the coalition. The second inequality is accordingly referred to as the *external stability* condition; no outsider has an incentive to join.[13] These internal and external stability conditions jointly

[13]Here, following other studies in the literature, we assume that players choose to be a member of a

capture the central feature of self-enforcing agreements: coalitions are not stable unless everyone agrees to settle. Technically, the one-shot stable set corresponds to the set of all Nash equilibria in the participation game, where no single player has an incentive to change their participation decision *given the participation decisions of the others*. While appropriate for simultaneous-move games with irreversible participation, this notion of stability fails to reflect the open-ended nature of real-world settings such as climate negotiations. The internal stability condition, for instance, compares payoffs under coalition M to those under $M \setminus \{i\}$, implicitly assuming that the other players cannot react to player i's defection. An analogous criticism applies to the external stability condition.

The solution concept of von Neumann and Morgenstern (1944) addresses this issue. We introduce it using the effectivity set, which formalizes the set of players whose membership status changes when the coalition shifts from M to M'.

Definition 2.2 (Effectivity set)**.** For each pair $(M, M') \in \mathcal{N} \times \mathcal{N}$, we define the effectivity set as $E(M, M') := (M \setminus M') \cup (M' \setminus M)$.

The sets $M \setminus M'$ and $M' \setminus M$ respectively represent the players leaving and joining the coalition in the transition from M to M'. Their union can thus be interpreted as the set of players who can collectively bring about the transition. We define a coalition M as dominated by another coalition M' if all players in the effectivity set $E(M, M')$ prefer M' over M.[14]

Definition 2.3 ($\mathcal{M}$-domination)**.** A coalition M is M'-dominated if

$$u_i(M') \geq u_i(M) \quad \forall i \in E(M, M'), \tag{4}$$

where at least one of the inequalities is strict. We say that M is $\mathcal{M}$-dominated if there exists a coalition $M' \in \mathcal{M}$ such that M is M'-dominated.

Equipped with the concept of $\mathcal{M}$-domination, we can define von Neumann-Morgenstern stability.

Definition 2.4 (vNM stable set)**.** A nonempty subset $\mathcal{M} \subset \mathcal{N}$ is a vNM stable set if

$$M \in \mathcal{M} \iff M \text{ is not } \mathcal{M}\text{-dominated.} \tag{5}$$

The literature on cooperative game theory refers to the right arrow in condition (5) as *internal stability* and to the left arrow as *external stability*. These are different from the internal and external stability concepts used in the self-enforcing agreement literature discussed above. From here on, we adopt the terminology in accordance with the cooperative

coalition whenever they are indifferent.

[14] This definition presumes that insiders and outsiders of a coalition can communicate and coordinate their actions upon deviations.

game theory literature. Internal stability (the right arrow) requires that a stable coalition cannot be dominated by another stable coalition. External stability (the left arrow) requires that every unstable coalition must be dominated by some stable coalition. Unlike one-shot stable sets, players compare a candidate coalition only to *stable* coalitions, which is the key conceptual advantage of vNM stable sets. However, characterizing vNM stable sets is generally difficult and there is no guarantee that a vNM stable set even exists.

Remark 1. No vNM stable set exists in Example 2 with $n = 4$ (see Appendix D.1).[15]

The difficulty in characterizing and even establishing the existence of vNM stable sets stems from the fact that the concept relies on finding a fixed point; to determine whether a coalition is stable, one must already know the entire set of stable coalitions. Proving the (non-)existence of such a fixed point typically requires detailed specifications of the underlying model and, in our example, it depends on the number of players involved.

Before turning to our contribution, it is helpful to compare vNM stability with one-shot stability by recasting the latter in the language of cooperative game theory. The essential step in one-shot stability is to test whether a candidate coalition is dominated by any "neighboring" coalition that emerges when a player unilaterally changes his or her participation status.

Definition 2.5 (Neighbor-domination)**.** Given coalition $M \in \mathcal{N}$, we define the map $M^o : N \to \mathcal{N}$ that characterizes for each player $i \in N$ the coalition emerging when that player changes participation status:[16]

$$M^o(i) := \begin{cases} M \setminus \{i\} & \text{if } i \in M \\ M \cup \{i\} & \text{if } i \notin M. \end{cases} \tag{6}$$

A coalition M is neighbor-dominated if

$$\exists i \in N \text{ such that } u_i(M^o(i)) \geq u_i(M), \tag{7}$$

where we require strict inequality if $i \in M$.

Given this definition, we have the following alternative definition of one-shot stable sets.

Remark 2. One-shot stable sets can equivalently be defined as the set $\mathcal{M} \subset \mathcal{N}$ such that

$$M \in \mathcal{M} \iff M \text{ is not neighbor-dominated.}$$

[15] If direct dominance in the vNM solution is replaced by the indirect dominance of Harsanyi (1974), stable sets exist in this example. For instance, the grand coalition indirectly dominates other coalitions through the threat of agreement collapse. This mechanism, however, relies on the specific structure of this example. In general, neither the existence nor the tractability of (farsighted) stable sets is guaranteed, especially when players are asymmetric or when multiple coalitions can coexist.

[16] Some readers might prefer defining the map generally as $^o : \mathcal{N} \times N \to \mathcal{N}$, which also characterizes the coalition $M^o(i)$ by equation (6).

This alternative definition allows us to make a direct comparison between one-shot stability and vNM stability along two important dimensions. First, as evident from condition (7), the neighbor domination is triggered whenever any player benefits from unilateral deviation, capturing the self-enforcing nature of the solution concept. In essence, a single player can unsettle the status quo if the grass looks greener on the other side. By contrast, the domination criterion adopted in vNM stability (condition (4)) implies that a deviation is possible only when all players on the transition path agree. Second, to evaluate one-shot stability, we merely have to compare a candidate coalition to its given neighbor coalitions (one for each player) regardless of whether these neighbors are themselves stable. On the other hand, vNM stability requires comparing a candidate coalition to a set of endogenously determined stable coalitions.

2.3 Self-enforcing stable sets

Our solution concept of self-enforcing stable sets synthesizes components of one-shot stability and vNM stability along the two dimensions discussed above. A crucial component we adopt from the one-shot concept of self-enforcing agreements is that a single player can unsettle the situation if he or she thinks that the grass is greener on the other side. Yet, the other side is no longer the immediate neighbor. From von Neumann and Morgenstern's stability concept and the subsequent farsightedness literature, we adopt the requirement that players look ahead and anticipate the long-run stable outcome(s) triggered by a deviation.

A key innovation of our framework is that, in synthesizing these features, we explicitly model players' beliefs about the long-run outcomes as part of the equilibrium. Players recognize that the game has a set of alternative stable outcomes and that any deviation will trigger a sequence of reactions eventually leading to one of these alternatives. The consequence of defection is therefore not uniquely determined a priori. The existing literature typically resolves this uncertainty through behavioral assumptions: potential defectors are either optimistic, deviating whenever some reaction path leads to improvement (Harsanyi, 1974), or pessimistic, deviating only when all paths lead to improvement (Chwe, 1994). In complex, unstructured, real-world negotiations, however, assuming players can predict the exact sequence of moves and countermoves is unrealistic. This fundamental uncertainty motivates the methodological shift from tracing specific deviation paths to focusing on the endpoints of negotiation via probabilistic beliefs. Rather than imposing optimism or pessimism, we allow players to form beliefs over the eventual outcomes, and assume they deviate if and only if they are better off in expectation.

To formalize these ideas, we introduce an alternative domination criterion. In our approach, domination is no longer defined relative to a single neighboring coalition or a deterministic set of stable alternatives, but is instead defined with respect to a belief and the expectations it induces. Let $\mathcal{M}$ be the set of (believed-to-be) possible long-run out-

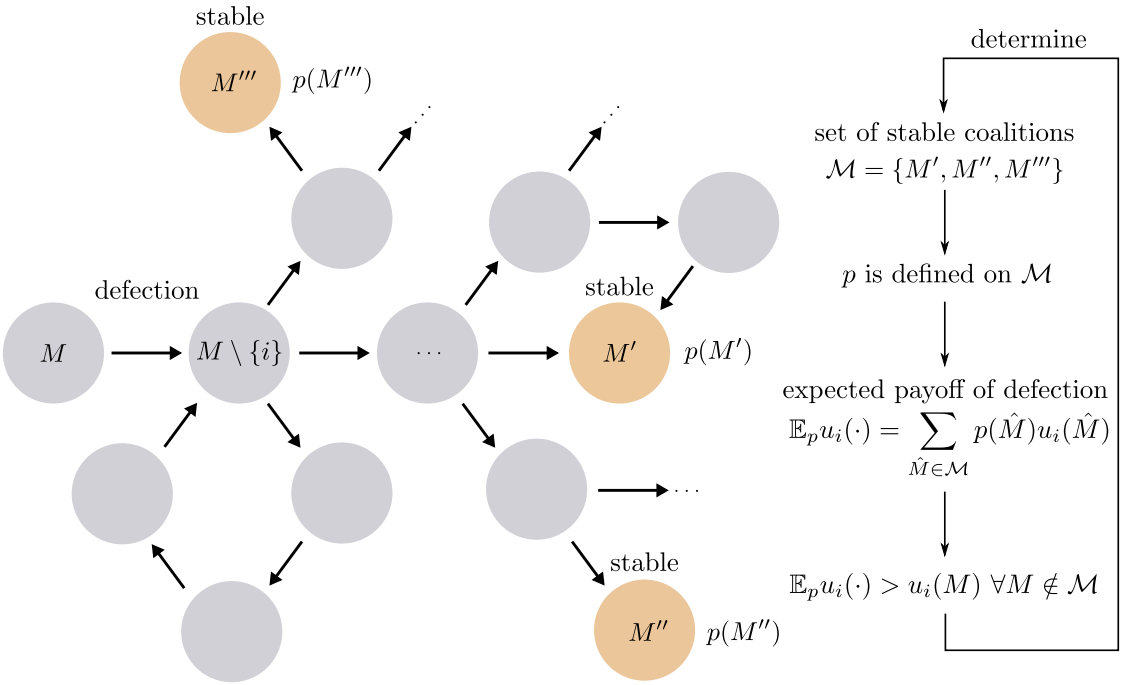

Figure 1: Schematic diagram illustrating the idea of self-enforcing stable sets.

comes of the negotiation process (see Figure 1 for an illustration). We define a corresponding belief as $p : \mathcal{M} \to (0, 1]$ satisfying $\sum_{M \in \mathcal{M}} p(M) = 1$. It assigns strictly positive weight to potentially stable coalitions $\mathcal{M}$. Player i's expected payoff under belief p is $\mathbb{E}_p u_i(\cdot) := \sum_{M \in \mathcal{M}} p(M) u_i(M)$, which characterizes the expected consequence of unsettling the negotiations. We say that a coalition M is dominated under belief p if some player is better off in expectation under belief p than he or she would be by sticking with M.

Definition 2.6 (p-domination)**.** A coalition M is p-dominated if

$$\exists i \in N \text{ such that } \mathbb{E}_p u_i(\cdot) > u_i(M).$$

At first glance, allowing *any* player to unsettle the status quo may seem like a strong requirement. Yet this individual ability is tightly disciplined by common rational expectations: a deviator compares payoffs only with coalitions that *all* players view as feasible long-run outcomes. Unlike other approaches in the literature, no single player can unilaterally hold optimistic or pessimistic expectations regarding the outcome of defection. In our framework, expectations are an endogenously determined component of equilibrium rather than a fixed behavioral assumption.

We view negotiations as a process of converging to a common belief p that governs the set of possible long-term stable coalitions. While our analysis assumes a common belief, we interpret it as an equilibrium outcome rather than a premise. Appendix B demonstrates that even if players begin with heterogeneous beliefs, the stability conditions force convergence to a common belief in equilibrium. Either way, we abstract from the details of any specific

Table 2: Comparison of different solution concepts.

	Solution	**Comparing to**	**Relevant players for**	**destabilizing condition**
One-shot	$\mathcal{M} \subset \mathcal{N}$	neighbor	$\exists i \in N$	$u_i(M^o(i)) \geq u_i(M)$ inequality strict for insiders
SES	$\mathcal{M}, p : \mathcal{M} \to (0,1]$	stable alternatives $\mathcal{M}$ using belief p	$\exists i \in N$	$\mathbb{E}_p u_i(\cdot) > u_i(M)$
vNM	$\mathcal{M} \subset \mathcal{N}$	stable alternatives $\mathcal{M}$ using each element	$\forall i \in E(M, M')$	$u_i(M') \geq u_i(M)$ one inequality strict

Note: Dashed lines indicate conceptual similarity between neighboring cells. One-shot = standard concept for self-enforcing equilibria in the literature. SES = self-enforcing stable sets. vNM = von Neumann-Morgenstern stability.

negotiation process and merely formulate the requirements for a set of coalitions to qualify as eventual outcomes of the negotiation. A set of coalitions $\mathcal{M}$ is a self-enforcing stable set if there exists a corresponding belief p over $\mathcal{M}$ such that $\mathcal{M}$ consists of exactly those coalitions that are not p-dominated.[17]

Definition 2.7 (Self-enforcing stable set)**.** A nonempty subset $\mathcal{M} \subset \mathcal{N}$ is self-enforcing stable (SES) if there exists a belief $p : \mathcal{M} \to (0,1]$ such that:

$$M \in \mathcal{M} \iff M \text{ is not } p\text{-dominated.} \tag{8}$$

The SES solution, like von Neumann and Morgenstern's (1944) solution concept, is defined implicitly as a fixed point: $\mathcal{M}$ appears on both sides of condition (8). We demonstrate in the subsequent sections that, despite this implicit definition, the proposed solution generally exists and can be characterized even in highly abstract settings. The right arrow in condition (8) defines internal stability, which requires that any stable outcome must not be dominated (here, in expectation) by other stable outcomes. The left arrow characterizes our version of external stability; whenever all players (weakly) prefer a coalition to the expected outcome implied by the stable set, that coalition must be part of the stable set.[18]

Table 2 summarizes the three stability concepts we discussed, illustrating how our solution concept combines key features of self-enforcing one-shot stability and forward-looking vNM stability. The "Solution" column highlights that our concept employs a belief over

[17]We abstract from negotiation histories by treating continuation beliefs as stationary over long-run endpoints (Appendix A). History may still matter for equilibrium selection when multiple stable endpoints exist: stability implies convergence to a common belief in equilibrium (Appendix B), and once convergence occurs, Lemma 3.1 will imply that beliefs concentrate on a payoff-equivalent singleton.

[18]Note that the logic of external stability is analogous to subgame perfection: when evaluating a candidate equilibrium outcome, we cannot ignore off-equilibrium alternatives that would be stable once proposed. Appendix A formalizes this intuition, showing that self-enforcing stable sets can be supported as subgame-perfect equilibria of a sequential bargaining game in which proposers are selected at random.

the stable outcomes rather than only the outcomes themselves. The "Comparing to" column shows that the one-shot stability looks only at neighbor coalitions, which are typically transient short-term outcomes, whereas the SES and vNM solutions both compare the status quo with long-term stable alternatives. To define long-term stability, the SES solution adopts a belief over stable outcomes, whereas the vNM stability relies on pairwise comparison with individual stable outcomes. The "Relevant players" column indicates that self-enforcing solutions—both one-shot and SES—must withstand any unilateral deviation, whereas the vNM solution requires a potentially large set of players to be on board with a change. The final column details the specific payoff comparison, which is similar in spirit across all concepts: one or more players expect a higher payoff from deviating.

3 Characterization

This section characterizes self-enforcing stable sets, first in general, and then for the climate-related examples of Section 2.1.

3.1 General results

An attractive feature of our solution concept is that the equilibrium outcomes can be characterized even in general settings. To streamline the presentation, we begin by introducing notation to handle payoff-equivalent coalitions.

Definition 3.1 (Distinct/payoff-equivalent coalitions)**.**

(i) Coalitions $M, M' \in \mathcal{N}$ are *distinct* if there exists $i \in N$ such that $u_i(M) \neq u_i(M')$.

(ii) If M and M' are not distinct, we write $M \sim M'$.

(iii) We refer to a set of coalitions $\mathcal{M} \subset \mathcal{N}$ as *effectively singleton* if $M \sim M'$ for all $M, M' \in \mathcal{M}$.

(iv) We write $\tilde{M} := \{M' \in \mathcal{N} \mid M' \sim M\}$ for M's indifference class, i.e., the set of payoff-equivalent coalitions containing M.

The following lemma reveals how rational foresight disciplines equilibrium beliefs in self-enforcing settings (all proofs are provided in Appendix D).

Lemma 3.1. *A self-enforcing stable set cannot contain distinct coalitions.*

Lemma 3.1 implies that any self-enforcing stable set must be effectively singleton, containing only coalitions with identical payoff profiles. If a stable set contained coalitions with distinct payoffs, there would always be at least one player for whom one of these stable coalitions is worse than the others. Such a player would refuse to settle for that particular coalition and thus the coalition cannot be part of the stable set. This logic is shared by all

players. Every player understands that a coalition is not stable if any player can identify a more favorable alternative that others also recognize as stable. Consequently, negotiations can only settle when consensus emerges on the eventual outcome. This finding can be interpreted as a process of belief discipline through negotiations. While players may begin with arbitrary beliefs, as negotiations unfold, only a subset of those beliefs survives the consistency requirements of self-enforcing stable sets. Any equilibrium belief that ultimately prevails must assign positive probability only to an effectively singleton set.

The critical question, then, is which effectively singleton sets qualify as self-enforcing stable sets. We find that they are characterized by Pareto efficiency.

Definition 3.2 (Pareto efficient coalition)**.** A coalition is Pareto efficient if no other coalition leaves all players equally well off while making at least one player strictly better off.

While this definition is standard, its interpretation in our framework is crucial. Because we consider settings where transfers among players are restricted (as embedded in the reduced-form payoffs), this notion of Pareto efficiency must be understood as constrained efficiency, i.e., the efficiency attainable without transfers. This contrasts with full efficiency, which assumes unrestricted transfers.[19] Achieving full efficiency would require compensating players who benefit from a fragmented structure, which is ruled out in our self-enforcing framework.

The following theorem establishes the central result of our analysis.

Theorem 3.1. *A subset $\mathcal{M} \subset \mathcal{N}$ is a self-enforcing stable set if and only if there exists a Pareto efficient coalition M such that $\mathcal{M} = \tilde{M}$. There always exists a self-enforcing stable set.*

Every self-enforcing stable set is the indifference class of a Pareto efficient coalition, which immediately establishes its existence through the existence of Pareto efficient coalitions. The key driver of this result is the external stability condition (the left arrow in condition (8)). To see why, suppose that coalition M is in a self-enforcing stable set. By Lemma 3.1, all other coalitions in that self-enforcing stable set must be payoff-equivalent to M. Thus, if M were not Pareto efficient, all stable coalitions would be Pareto dominated by some *unstable* coalition M'. But once M' were proposed as an outcome of the game, every player would unanimously accept it. The external stability condition then implies that M' must be stable, yielding a contradiction. In equilibrium, all players agree on a particular Pareto efficient coalition as the (effectively) unique possible outcome of the game.

[19] If transfers were costlessly enforceable, it is well known that the fully efficient outcome can be stable (see, for example, Okada (2023)). Under such conditions, our framework would identify the grand coalition as the unique stable structure, as surplus redistribution allows it to Pareto-dominate any fragmented arrangement. We note that, once transfers are allowed, each coalition structure corresponds to a continuum of payoff outcomes (each determined by a particular transfer arrangement), thereby expanding the space of reduced-form payoffs. Consequently, the self-enforcing stable set under enforceable transfers would consist of all payoff vectors implementable within the grand coalition.

One might suspect that the Pareto efficiency in Theorem 3.1 is a direct consequence of the unanimity embedded in the internal stability condition (the right arrow in condition (8)). That is not the case. The primary role of internal stability is to ensure that any stable set is effectively singleton, which by itself does not imply efficiency. It is instead the external stability condition that requires stable outcomes to be Pareto efficient.[20] A grossly inefficient outcome would invite the expectation that a Pareto-dominating alternative will eventually be proposed, thereby destabilizing the inefficient state. This filtering process continues until all Pareto improvements have been exhausted. A key insight from Theorem 3.1 is therefore that, in self-enforcing settings, the consistency of farsighted beliefs is crucial for restoring the efficiency suggested by the Coase theorem.[21]

3.2 Relation to prior work

The literature has long sought to establish a connection between stability and efficiency. Typically, the logic for stabilizing efficient outcomes involves a form of coalition breakdown: defections are deterred by the threat of triggering unfavorable fragmentation, or efficient outcomes indirectly dominate inefficient alternatives through a temporary agreement collapse. While powerful, such mechanisms are inherently model-specific, making general insights difficult to obtain. Our approach provides a distinct logic for efficiency. Focusing on self-enforcing settings and endogenizing players' expectations as probabilistic beliefs, we show that efficiency follows directly from general consistency requirements imposed on equilibrium beliefs. Internal stability—no stable outcome is dominated in expectation by the set of stable outcomes—forces beliefs to converge on a single endpoint, whereas external stability—any unanimously preferred outcome must itself be considered stable—rules out the expectation of inefficient stable outcomes. Together, these forces deliver an exact equivalence between stability and constrained efficiency without relying on model-specific structures.

To put this result in context, we contrast it with two prominent approaches that also link stability with efficiency in games with externalities.

Sequential bargaining with irreversible commitments

A key approach in the literature is the sequential bargaining framework of Ray and Vohra (1999, 2001). In their framework, coalition formation is modeled as an extensive-form game with irreversible commitments. Players negotiate sequentially, and once a player commits to a singleton coalition (becoming a free rider), that player permanently exits the negotiation. The remaining players then continue bargaining exclusively among themselves. This irre-

[20] To be more precise, both internal and external stability act together. The unanimity condition of internal stability plays an *indirect* role by ruling out the possibility that two or more distinct coalitions are simultaneously stable. Without it, external stability alone would not be able to eliminate inefficiency.

[21] We reiterate that our payoff functions can incorporate transaction costs. We are interested in the settings where the stakes are sufficiently high to overcome such frictions.

versibility allows the use of backward recursion to characterize a unique equilibrium coalition structure.

Their results show that equilibrium inefficiency is bounded, and full efficiency can sometimes be attained, depending on the total number of players. In their framework, efficiency is achieved through deterrence: a player considering defecting from the grand coalition anticipates that the remaining players may continue to fragment, potentially resulting in a highly inefficient coalition, reducing the defector's payoff below the grand coalition's payoff. This threat is credible when the process is irreversible: once the first player leaves, forming the grand coalition becomes impossible. Consequently, smaller coalitions are evaluated only against the restricted set of outcomes still achievable, allowing even Pareto inefficient outcomes to serve as credible threats.

Our framework operates through a fundamentally different mechanism. Since participation is non-binding, the threats of triggering the process to end up with highly inefficient outcomes are no longer credible. Instead, defection is deterred by its futility. As shown in Lemma 3.1, internal stability requires beliefs to concentrate on an effectively unique stable endpoint. Thus, *once an equilibrium belief emerges*, any deviation is perceived as pointless, as players have no rational basis to expect a better eventual outcome. While the sequential bargaining framework of Ray and Vohra (1999, 2001) is suited to structured environments with strong commitment technologies, our belief-based approach applies to unstructured negotiations where agreements must be self-enforcing.

Farsighted stability with optimistic deviations

Another point of comparison is the work of Diamantoudi and Xue (2007) and Diamantoudi and Sartzetakis (2015), who study environments closer to ours, characterized by open-ended negotiations. Extending the notion of equilibrium binding agreement of Ray and Vohra (1997), they employ vNM stable sets based on indirect dominance. In their framework, a coalition is unstable if there exists any sequence of deviations that ultimately benefits all players involved in the reaction chain. This approach resolves strategic uncertainty by assuming optimism: a deviation is considered viable if at least one favorable path exists.

Under this criterion, a constrained- or fully efficient agreement can be sustained if players can optimistically foresee a multi-step path leading back to the efficient outcome from any inefficient alternative. Such paths often involve a temporary, coordinated collapse of cooperation. For instance, members of an existing coalition might strategically defect, either simultaneously or sequentially, to unravel the inefficient status quo, thereby creating incentives for free riders to join a new, more efficient coalition. The credibility of such paths, however, is highly model-specific. In fact, as Diamantoudi and Xue (2007) show with counterexamples, this approach can also admit inefficient equilibria, even when transfers are permitted.

Our approach offers a different resolution to strategic uncertainty. Instead of assuming optimism and tracing specific deviation paths, we endogenize expectations as equilibrium beliefs over the set of stable endpoints. Efficiency arises not from the threat of a temporary collapse, but from general consistency requirements on these beliefs. Thus, while their studies characterize what can be stable under optimistic expectations, our self-enforcing stable sets identify what must be stable when expectations themselves are part of the equilibrium.

3.3 Applications

We now examine self-enforcing stable sets in the two examples introduced in Section 2.1. These examples represent common frameworks in the literature on international environmental agreements, especially those addressing climate change. For comparison, we also characterize the one-shot stable sets frequently analyzed in this literature.

Isoelastic payoff

For the isoelastic payoff specification in Example 1, the reduced-form payoff function is given by equation (1). The quadratic cost case ($\gamma = 2$) is widely applied in the literature. Karp and Sakamoto (2021) characterize the one-shot stable sets for general isoelastic cost convexities. In particular, they find the following.

Proposition (Karp and Sakamoto, 2021)**.** *The one-shot stable set in Example 1 is*

$$\mathcal{M} = \{M \in \mathcal{N} \mid |M| = m_\gamma\},$$

where m_γ is an integer satisfying $m_\gamma = 3$ for $\gamma = 2$ and $m_\gamma = 2$ for $\gamma > 2$.

The one-shot stability concept determines a unique coalition size, but not the identity of the members. We observe that (i) the size of stable coalitions, m_γ, is independent of the total number of players, n, and that (ii) these coalitions are very small and the resulting allocation is highly inefficient for large n. Applied to climate change agreements, this solution concept predicts that, of approximately 200 nation-states, only two or three will participate in the final agreement.

By contrast, the self-enforcing stable sets in this example are sensitive to the number of players and the associated predictions are much less pessimistic.

Proposition 3.1. *The set of self-enforcing stable sets in Example 1 is*

$$\mathscr{C} = \left\{\tilde{M} \subset \mathcal{N} \mid |M| > z_\gamma(n)\right\}, \tag{9}$$

where the minimal participation threshold $z_\gamma(n) \in (1, n)$ is the unique positive root of

$$n^{\frac{\gamma}{\gamma-1}} - n - \frac{1}{\gamma} n^{\frac{\gamma}{\gamma-1}} = z^{\frac{\gamma}{\gamma-1}} - z - \frac{1}{\gamma} \tag{10}$$

for a given number of players n and cost convexity parameter γ. The minimal participation threshold $z_\gamma(n)$ and the corresponding minimal membership share $r_\gamma(n) := z_\gamma(n)/n$ are both increasing in n. As the number of players grows, the membership share converges to $\lim_{n\to\infty} r_\gamma(n) = (1 - 1/\gamma)^{1-1/\gamma}$. Moreover, for $n \geq 12$, both $z_\gamma(n)$ and $r_\gamma(n)$ increase in the abatement cost convexity γ.

For the linear-quadratic model ($\gamma = 2$), the threshold has a closed-form expression $z_2(n) = \frac{1+\sqrt{1+2(n-1)^2}}{2}$, and the asymptotic minimal membership share is $1/\sqrt{2} \approx 0.707$.

The inequality in equation (9) reflects Theorem 3.1, specialized to this symmetric free-rider environment. In any self-enforcing stable set, a coalition must give every player at least as much as she expects to get by triggering renewed negotiations. Because our payoffs depend only on coalition size, this expected-deviation condition reduces to a simple comparison of payoffs in the grand coalition and free-riding on a smaller coalition. In this example, each self-enforcing stable set is a singleton.

A coalition is self-enforcing stable if and only if it has strictly more than $z_\gamma(n)$ members, formally $\lfloor z_\gamma(n) \rfloor + 1$, where $\lfloor z \rfloor = \max\{k \in \mathbb{N} \mid k \leq z\}$ is the floor function.[22] The threshold $z_\gamma(n)$ depends on the number of players, n, as illustrated in Figure 2. The left panel in the figure plots the minimal size of stable coalitions, $\lfloor z_\gamma(n) \rfloor + 1$, whereas the right panel plots the minimal membership share, $z_\gamma(n)/n$, both as functions of n. For the common linear-quadratic model, the minimal membership share converges to $1/\sqrt{2} \approx 0.707$; thus, as the number of players grows large, at least 70% of them must participate for a coalition to be stable. This finding contrasts sharply with the predictions of one-shot stability, where a stable coalition has at most three members, irrespective of the total number of players in the game.

The reasoning behind the contrasting predictions is revealing. In the one-shot framework, stability requires that the payoff from free riding be *small* enough to deter myopic members from defecting, a condition rarely met when externalities are substantial. Consequently, one-shot stability contains free riding only when it is not a major concern to begin with.

In our framework, by contrast, stability for any non-grand coalition requires that free riders earn sufficiently *large* payoffs. If a coalition fails to provide such benefits, all players would unanimously prefer a more efficient alternative, such as the grand coalition. The recognition that the grand coalition can be a stable alternative then destabilizes an inefficient status quo. As the number of countries and/or the magnitude of the externality increases, the payoffs to cooperation increase, and countries will also require higher payoffs as outsiders. This logic is formalized in equation (10), where the left side represents the payoff to a member of the grand coalition, and the right side the payoff to a free rider on a coalition of size z. As n increases on the left side, z must increase on the right side as well; larger gains from

[22] Similarly, we denote the ceiling function (the smallest integer greater than or equal to z) by $\lceil z \rceil$, which we use later.

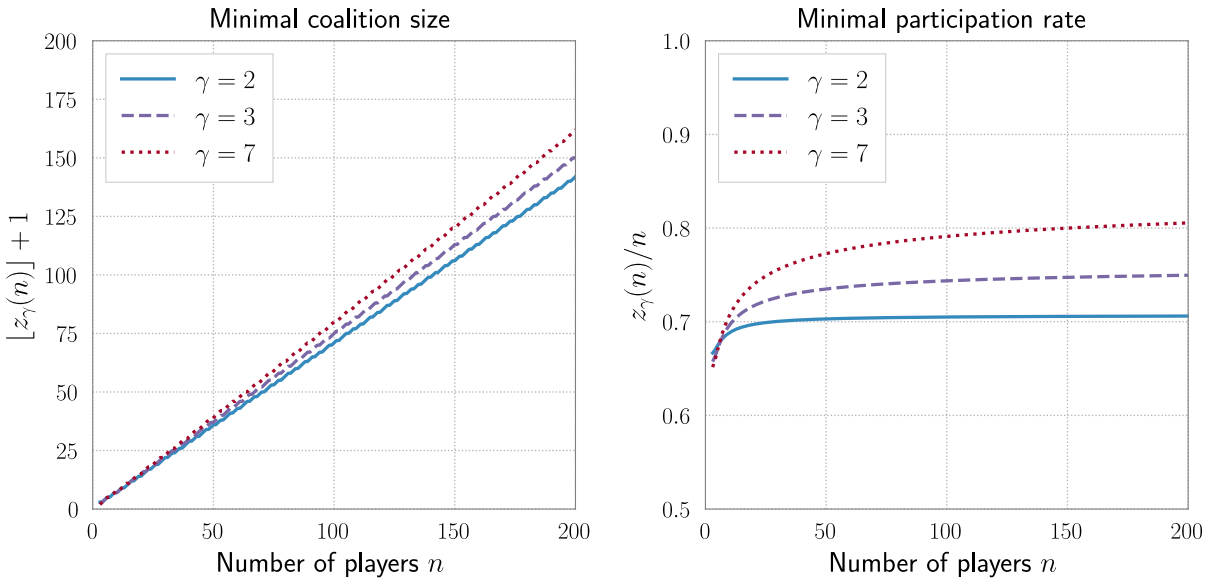

Figure 2: Minimal self-enforcing stable coalition size $\lfloor z_\gamma(n) \rfloor + 1$ (left) and participation rate $z_\gamma(n)/n$ (right) as functions of the number of players n for different values of the cost convexity γ in Example 1. The depicted minimal coalition size is also the unique second-order stable coalition size (see Section 4).

full cooperation require that the minimal stable coalition size rise to ensure high enough payoffs to free riders.[23]

Figure 2 also shows that higher cost convexity γ leads to greater participation. Proposition 3.1 establishes this observation as a general result for $n \geq 12$.[24] When marginal abatement costs rise more steeply, the socially optimal abatement level falls, reducing the required abatement effort, which lowers the membership cost of the grand coalition relative to free riding.[25] Consequently, to compete with the relatively higher net benefits of full cooperation, non-member payoffs must be larger to prevent Pareto domination, which in turn requires a higher threshold for the stable coalition size. Again, this result stands in stark contrast to the prediction from the one-shot stable set, where more convex costs imply reduced participation.

[23] Large free-riding benefits do not induce defections because farsighted members recognize that any free-rider status gained from deviating would be merely transient and that there is no rational basis for expecting a better outcome than the current arrangement.

[24] The requirement $n \geq 12$ is sufficient, but not necessary. Numerical computations indicate that the statement is true for all $n \geq 8$.

[25] The effect is visible in equation (10), where the terms $\frac{1}{\gamma} n^{\frac{\gamma}{\gamma-1}}$ and $\frac{1}{\gamma}$ represent the abatement costs for grand coalition members and free riders, respectively. Both cost terms decrease in γ, but the reduction is greater for grand coalition members, especially for large n, because of the additional abatement. The increasing γ in the exponent of $n^{\frac{\gamma}{\gamma-1}}$ reflects higher costs due to stronger convexity, whereas the decreasing (and dominating) $\frac{1}{\gamma-1}$ captures the reduction in the optimal abatement target induced by the cost increase.

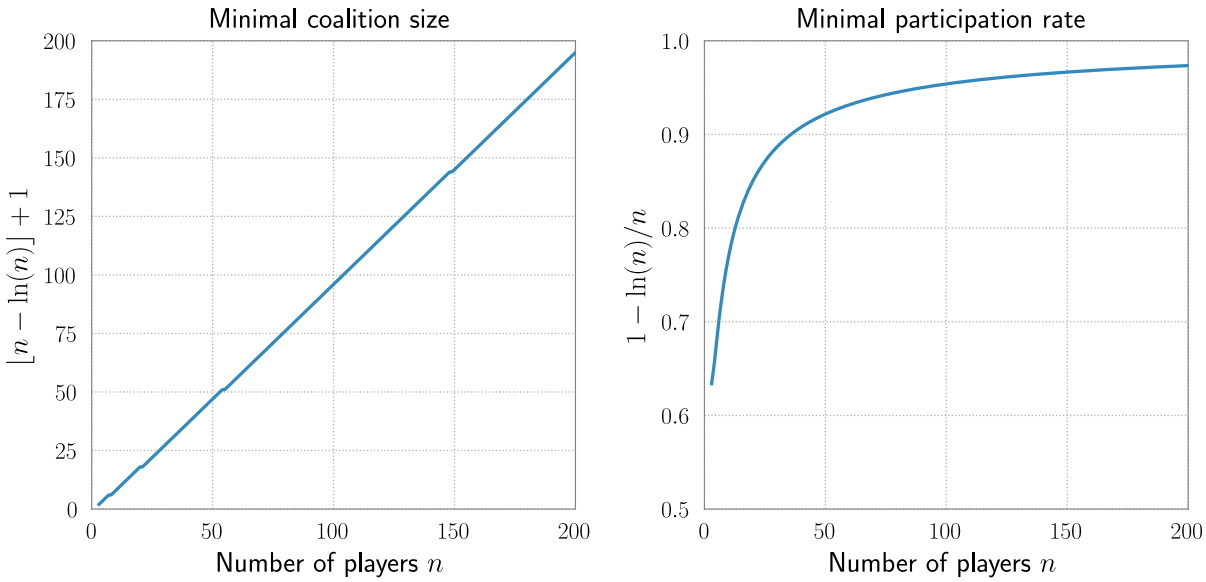

Figure 3: Minimal coalition size $\lfloor n - \ln(n) \rfloor + 1$ (left) and participation rate $(n - \ln(n))/n$ (right) as functions of the number of players n in Example 2. The depicted minimal coalition size is also the unique second-order stable coalition size (see Section 4).

Logarithmic payoff

In Example 2, the one-shot stable set gives an even more pessimistic outlook and the self-enforcing stable sets are even easier to characterize.

Proposition 3.2. *In Example 2 with reduced-form payoffs given by equation (2), a) the unique one-shot stable set is*

$$\mathcal{M} = \{M \in \mathcal{N} \mid |M| = 2\};$$

b) the set of self-enforcing stable sets is

$$\mathscr{C} = \left\{\tilde{M} \subset \mathcal{N} \mid |M| > n - \ln(n)\right\},$$

where the minimal size of stable coalitions is strictly increasing in n.

One-shot stability suggests that only two players can participate in a stable coalition, no matter how many players exist in the game. On the other hand, similarly to Example 1 in Proposition 3.1, self-enforcing stability gives rise to a much more optimistic prediction. A coalition is self-enforcing stable if and only if its size is greater than a threshold value, $n - \ln(n)$, eliminating excessively inefficient outcomes. Figure 3 depicts the minimal size of stable coalitions, $\lfloor n - \ln(n) \rfloor + 1$, and the minimal participation rate, $1 - \frac{\ln(n)}{n}$, both of which increase as the number of players grows.

A notable result from this particular specification is that *the minimal participation rate converges to 1 for $n \to \infty$*. This result follows from the model's structure: while the

damages from pollution are linear, the utility from emissions is logarithmic. To be more precise, coalition M is self-enforcing stable if and only if

$$\ln(1/\xi) + |M| > \ln(1/(\xi n)) + n,$$

which requires that the payoff to a free rider (left side) be greater than the payoff to a member of the grand coalition (right side). Relative to free riders, members of the grand coalition benefit more from avoided pollution ($n > |M|$), but at the cost of reduced consumption ($1/(\xi n) < 1/\xi$). As the number of players, n, increases, the benefits of forming the grand coalition rise linearly, whereas the utility gain from free riding grows only logarithmically. Consequently, the minimal share of coalition members must grow such that free-riding benefits can still compete with cooperation benefits. To illustrate the speed of this convergence, we consider the case of $n = 195$ countries, as in international climate negotiations. In this scenario, the minimal membership share is already very close to 1, requiring that at least 97% of potential participants become members. This figure aligns closely with the observed participation rate of the Paris Agreement on Climate Change,[26] demonstrating the empirical relevance of our solution concept.

4 General free-rider games and equilibrium refinement

The preceding applications show that our solution concept yields empirically plausible predictions. A potential limitation is that it may admit a large set of stable outcomes. This multiplicity can be interpreted as a second-order coordination problem: when several constrained-efficient endpoints are self-enforcing, parties may have incentives to engage in strategic positioning that affects which endpoint becomes focal. The present section addresses this issue by introducing a refinement mechanism that restricts the set of equilibria and delivers sharper theoretical predictions. We demonstrate the implications of this refinement by introducing a general class of symmetric free-rider games, for which we characterize the self-enforcing stable sets before and after applying the refinement.

4.1 Second-order stability

To capture additional considerations that can help narrow down the set of likely equilibria, we consider a meta-game. The meta-game is played on the set of self-enforcing stable coalitions, and we select its solutions by applying vNM stability. We denote by $\mathscr{C}$ the

[26] As of August 2025, 192 of the 195 countries (98.5%) have ratified the Paris Agreement. This figure includes Niue and the Cook Islands, which are self-governing states in free association with New Zealand. Looking only at UN member states (excluding Niue and the Cook Islands), 190 of 193 countries, or 98.4%, have ratified. If all parties to the United Nations Framework Convention on Climate Change (UNFCCC) are counted, including the European Union, the Holy See, and the State of Palestine, the total comes to 195 of 198, also 98.5%. The United States is expected to withdraw from the agreement in 2026.

collection of all self-enforcing stable sets. By Theorem 3.1, each element of $\mathscr{C}$ is of the form $\tilde{M}$ with $M \in \mathcal{N}$ Pareto efficient. Definition 4.1 extends von Neumann and Morgenstern's (1944) dominance relation from a subset of coalitions, $\mathcal{M} \subset \mathcal{N}$, to a subset of self-enforcing stable sets, $\mathscr{A} \subset \mathscr{C}$.

Definition 4.1 (Dominated *set* of coalitions.)**.** A self-enforcing stable set $\tilde{M} \in \mathscr{C}$ is dominated by a set $\mathscr{A} \subset \mathscr{C}$ of self-enforcing stable sets ($\mathscr{A}$-dominated) if there exists $\tilde{M}' \in \mathscr{A}$ such that M is M'-dominated.

Second-order stability is simply the application of the classic von Neumann and Morgenstern's (1944) solution concept to the meta-game played over the collection of self-enforcing stable sets, $\mathscr{C}$.

Definition 4.2 (Second-order stability)**.** A nonempty subset $\mathscr{A} \subset \mathscr{C}$ of self-enforcing stable sets is second-order stable if

$$\tilde{M} \in \mathscr{A} \iff \tilde{M} \text{ is not } \mathscr{A}\text{-dominated.} \tag{11}$$

The right arrow in condition (11), which we call the *second-order internal stability condition*, requires that no second-order stable solution of the meta-game should be dominated by another second-order stable solution. The *second-order external stability condition*, which is the left arrow in condition (11), requires that every second-order unstable outcome should be dominated by some second-order stable outcome.

4.2 Symmetric free-rider games

We noted in Section 2.2 that vNM stable sets, when applied to the entire outcome space, are often hard to characterize and may fail to exist. When vNM stability is applied to the meta-game over self-enforcing stable sets, neither of these issues arises in a large class of symmetric free-rider games. Moreover, our analysis of this class of games illustrates how second-order stability sharpens predictions by eliminating overly optimistic outcomes.

Our symmetric free-rider games capture the defining features of public good provision. On the one hand, enlarging a coalition benefits all players. On the other hand, each individual has an incentive to free ride on the efforts of others.

Assumption 1. Players' reduced-form payoff functions are symmetric and satisfy the following conditions. Let $M \subset M'$ with $|M| < |M'|$. Then,

$$\underbrace{u_i(M') > u_i(M)}_{\text{member payoff}}, \quad \underbrace{u_j(M') \geq u_j(M)}_{\text{non-member payoff}} \quad \text{for all } i \in M \text{ and } j \notin M' \tag{12}$$

and

$$\underbrace{u_i(M) \leq u_j(M)}_{\text{free-riding incentive}} \quad \text{for all } i \in M \text{ and } j \notin M. \tag{13}$$

The first inequalities in condition (12) assert that both coalition members and free riders are better off when the coalition is larger. Condition (13) requires that, for a given coalition, a free rider is weakly better off than a member. Both of our earlier examples satisfy Assumption 1.

Theorem 4.1. *Under Assumption 1, there exists an integer $m_\star$ such that a) the set of self-enforcing stable sets is*

$$\mathscr{C} = \left\{ \tilde{M} \subset \mathcal{N} \mid |M| \geq m_\star \right\};$$

b) a nonempty subset $\mathscr{A} \subset \mathscr{C}$ of self-enforcing stable sets is second-order stable if and only if

$$\mathscr{A} = \left\{ \tilde{M} \in \mathscr{C} \mid |M| = m_\star \right\}.$$

Part a) of Theorem 4.1 generalizes Propositions 3.1 and 3.2. A non-grand coalition in the symmetric free-rider game is Pareto efficient if and only if the free-riders' payoff is high enough to prevent Pareto domination by the grand coalition. Because the free-riding incentive increases with the size of a coalition, a coalition is Pareto efficient if and only if its size is greater than some threshold $m_\star$. In Examples 1 and 2, $m_\star$ is $\lfloor z_\gamma(n) \rfloor + 1$ and $\lfloor n - \ln(n) \rfloor + 1$, respectively. As in these examples, the general case also suggests that the class of self-enforcing stable coalitions can be large. In particular, the grand coalition is always self-enforcing stable, which might appear overly optimistic.

Part b) of the theorem characterizes the second-order stable outcomes of the games. Second-order stability pins down a unique coalition size, just as one-shot stability does. In contrast to one-shot stability, however, it predicts a Pareto efficient coalition that has just enough members—and thus provides just enough benefits to free riders—to avoid Pareto domination by the grand coalition. By selecting precisely the smallest coalition size that yields a (constrained) Pareto efficient outcome, second-order stability introduces a degree of pessimism to the otherwise optimistic predictions of the self-enforcing stability. In Examples 1 and 2, the sizes and the participation rates of the second-order stable coalitions are precisely those graphed in Figures 2 and 3.

5 Coalition structures

So far, we have assumed that players either free ride or join a single coalition. The present section extends the model to allow multiple coalitions to coexist.

5.1 Setting and general results

Let $\boldsymbol{\mathcal{N}}$ be the set of all partitions of N. Each partition $\boldsymbol{M} \in \boldsymbol{\mathcal{N}}$ represents a possible coalition structure, and $|\boldsymbol{M}|$ denotes the number of *coexisting coalitions* in $\boldsymbol{M}$. When $|\boldsymbol{M}| = L$, for example, $\boldsymbol{M}$ consists of L disjoint subsets that partition N, i.e., $\boldsymbol{M} = \{M_1, M_2, \ldots, M_L\}$ such that $\cup_{l=1}^{L} M_l = N$ and $M_l \cap M_k$ is empty for all $k \neq l$. The elements $M_1, M_2, \ldots, M_L$ are coexisting coalitions under $\boldsymbol{M}$, and every player belongs to exactly one of them. In this framework, what we previously termed a free rider corresponds to a member of a singleton coalition. As before, $u_i : \boldsymbol{\mathcal{N}} \to \mathbb{R}$ represents the payoff of player $i \in N$ under each possible coalition structure, and we write $\boldsymbol{M} \sim \boldsymbol{M'}$ if $\boldsymbol{M}$ and $\boldsymbol{M'}$ are payoff-equivalent for all players. We use $\boldsymbol{N} := \{N\}$ to denote the grand coalition structure.

Example 1*. Generalizing the isoelastic payoff model (Example 1) to the setting with multiple coalitions, we find that player i in any given coalition M_l chooses the emission level $g_i = \bar{g}_i - (|M_l|\xi)^{1/(\gamma-1)}$. The resulting reduced-form payoff to a member of M_l under coalition structure $\boldsymbol{M}$ (derived in Appendix C.2) is

$$u_i(\boldsymbol{M}) = \xi^{\frac{\gamma}{\gamma-1}} \left(\sum_{M \in \boldsymbol{M}} |M|^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma} |M_l|^{\frac{\gamma}{\gamma-1}} \right) - \xi \sum_{j \in N} \bar{g}_j \quad \forall i \in M_l. \tag{14}$$

The payoff function coincides with equation (1) if $|M_k| = 1$ for all $k \neq l$.

Example 2*. When the logarithmic payoff model (Example 2) is generalized to the present setting, members of a coalition $M_l \in \boldsymbol{M}$ choose $g_i = 1/(|M_l|\xi)$ and obtain the following reduced-form payoff (see Appendix C.2 for details):

$$u_i(\boldsymbol{M}) = -\ln(\xi) - \ln(|M_l|) - |\boldsymbol{M}| \quad \forall i \in M_l. \tag{15}$$

This payoff function coincides with equation (2) if $|M_k| = 1$ for all $k \neq l$.

Assumptions, Definitions, and Theorem 3.1 from the single-coalition setting extend directly to the setting with multiple coexisting coalitions: we simply replace coalitions M with coalition structures $\boldsymbol{M}$, sets of coalitions $\mathcal{M}$ with sets of coalition structures $\boldsymbol{\mathcal{M}}$, and the set of all coalitions $\mathcal{N}$ with the set of all coalition structures $\boldsymbol{\mathcal{N}}$.

Definition 5.1 (p-domination)**.** Given a belief $p : \boldsymbol{\mathcal{M}} \to (0, 1]$, a coalition structure $\boldsymbol{M}$ is p-dominated if there exists $i \in N$ such that $\mathbb{E}_p u_i(\cdot) > u_i(\boldsymbol{M})$.

Definition 5.2 (Self-enforcing stable set)**.** A nonempty subset $\boldsymbol{\mathcal{M}} \subset \boldsymbol{\mathcal{N}}$ of coalition structures is a self-enforcing stable set if there exists a belief $p : \boldsymbol{\mathcal{M}} \to (0, 1]$ such that:

$$\boldsymbol{M} \in \boldsymbol{\mathcal{M}} \iff \boldsymbol{M} \text{ is not } p\text{-dominated.}$$

Theorem 5.1. *A subset $\boldsymbol{\mathcal{M}} \subset \boldsymbol{\mathcal{N}}$ is a self-enforcing stable set if and only if there exists a Pareto efficient coalition structure $\boldsymbol{M}$ such that $\boldsymbol{\mathcal{M}} = \tilde{\boldsymbol{M}} := \{\boldsymbol{M}' \in \boldsymbol{\mathcal{N}} \mid \boldsymbol{M}' \sim \boldsymbol{M}\}$. There always exists a self-enforcing stable set.*

Self-enforcing stable sets remain effectively singleton and Pareto efficient. In light of Theorem 5.1, we take the liberty of referring to a coalition structure $\boldsymbol{M} \in \tilde{\boldsymbol{M}}$ as a self-enforcing stable coalition structure, rather than referring to the effectively singleton set.

5.2 One-shot stable sets

For comparison, we also characterize one-shot stable sets in the multi-coalition setting. When multiple coalitions can coexist, a unilateral deviation by a single player can induce several coalition structures. Suppose, without loss of generality, that the current coalition structure is $\boldsymbol{M} = \{M_1, \ldots, M_L\}$ and player i belongs to coalition M_l for some $l \in \{1, 2, \ldots, L\}$. If player i leaves M_l, there are L possible outcomes, depending on his or her destination. One possibility is that player i forms a singleton coalition, yielding $\boldsymbol{M}' = \{M_1, \ldots, M_l \setminus \{i\}, \ldots, M_L, \{i\}\}$. Alternatively, player i may join another coalition M_k with $k \neq l$, in which case the resulting coalition structure is $\boldsymbol{M}' = \{M_1, \ldots, M_l \setminus \{i\}, \ldots, M_k \cup \{i\}, \ldots M_L\}$. If $M_l \setminus \{i\}$ is empty, it is removed from the partition. We denote the set of all such potential coalition structures resulting from player i's deviation by $\boldsymbol{M}^o(i) \subset \boldsymbol{\mathcal{N}}$. With this notation, we can extend the definition of one-shot stability based on neighbor-domination.

Definition 5.3 (Neighbor-domination)**.** Given a coalition structure $\boldsymbol{M} \in \boldsymbol{\mathcal{N}}$, let $\boldsymbol{M}^o : N \rightrightarrows \boldsymbol{\mathcal{N}}$ be the correspondence that yields, for each player $i \in N$, the set of all coalition structures emerging when that player changes membership status. A coalition structure $\boldsymbol{M}$ is neighbor-dominated if

$$\exists i \in N,\ \exists \boldsymbol{M}' \in \boldsymbol{M}^o(i) \text{ such that } u_i(\boldsymbol{M}') \geq u_i(\boldsymbol{M}),$$

where we require strict inequality if player i belongs to a (weakly) smaller coalition under $\boldsymbol{M}'$ than under $\boldsymbol{M}$.[27]

Definition 5.4 (One-shot stability)**.** A coalition structure $\boldsymbol{M} \in \boldsymbol{\mathcal{N}}$ is one-shot stable if it is not neighbor-dominated.

The following proposition provides a complete characterization of the one-shot stable sets for Examples 1* and 2*.

[27] To prevent domination cycles, we adopt the following tie-breaking rule for unilateral moves: a player will switch to a different coalition if the move results in a strictly larger coalition and a weakly higher payoff. However, a move to a coalition of the same or smaller size requires a strictly positive payoff gain. This assumption is consistent with the convention used in our single-coalition analysis.

Proposition 5.1. *In Examples 1* and 2*, one-shot stable coalition structures take the following form. For Example 1* with $\gamma = 2$, a coalition structure $\boldsymbol{M}$ is one-shot stable if and only if every coalition in $\boldsymbol{M}$ has two or three members. In all other cases, $\boldsymbol{M}$ is one-shot stable if and only if it contains at most one singleton coalition and all remaining coalitions have exactly two members.*

Just as in the single-coalition setting, a one-shot stable coalition structure $\boldsymbol{M}$ in the multi-coalition setting is tightly constrained. For both Example 1* with $\gamma > 2$ and Example 2*, the structure is particularly simple: if n is even, every coalition in $\boldsymbol{M}$ has exactly two members; if n is odd, $\boldsymbol{M}$ includes one singleton coalition and all other coalitions have two members. For Example 1* with $\gamma = 2$, every coalition in $\boldsymbol{M}$ contains either two or three members. These results highlight three key features of one-shot stability. First, large coalitions remain unstable, as allowing for multiple coexisting coalitions does not increase the size of stable coalitions. This limitation is a direct consequence of myopic expectations: each coalition must independently withstand its members' short-sighted incentives for unilateral deviation. Second, the equilibrium structure is highly fragmented: negotiations among n players result in at least $\lfloor n/m \rfloor$ coexisting coalitions, where m is two or three. Third, coexisting coalitions must be of similar size, with size differences not exceeding one. When coalitions with asymmetric sizes coexist, players in the larger coalition have an incentive to switch to the smaller one, myopically expecting that everyone else stays put. Applied to climate negotiations with 200 countries in Example 2*, one-shot stability in the multi-coalition setting predicts that the outcome would be 100 two-country coalitions.

5.3 Self-enforcing stable sets

In contrast to one-shot stability, self-enforcing stable sets produce markedly different (and arguably more plausible) predictions for the same examples.

Proposition 5.2.
a) Possible full efficiency: The grand coalition, $\boldsymbol{N} \in \boldsymbol{\mathcal{N}}$, is self-enforcing stable in both Examples 1 and 2*.*
b) Limited fragmentation: Any self-enforcing stable coalition structure $\boldsymbol{M}$ satisfies

1. *$|\boldsymbol{M}| < n - z_\gamma(n) + 1$ in Example 1*, where $z_\gamma(n)$ is the positive root of equation (10).*
2. *$|\boldsymbol{M}| < \ln(n) + 1$ in Example 2*.*

c) Asymmetric coalition size: If $\boldsymbol{M}$ is self-enforcing stable, any two distinct coalitions $M, M' \in \boldsymbol{M}$ with $|M'| \geq |M|$ must satisfy:

1. *$|M'| > \frac{1}{\phi_\gamma}|M|$ in Example 1*, where $\phi_\gamma \in (0,1)$ is the unique root solving $\frac{\gamma}{\gamma-1} = (\phi+1)^{\frac{\gamma}{\gamma-1}} - \phi^{\frac{\gamma}{\gamma-1}}$.*

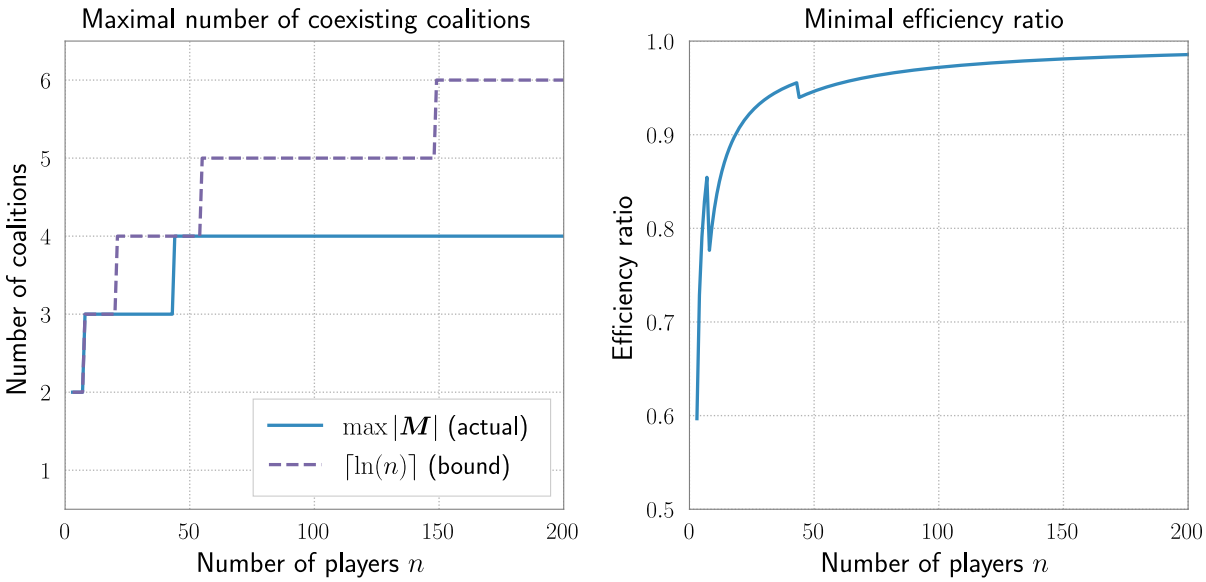

Figure 4: Maximal number of coalitions for stable coalition structures (left) and minimal equilibrium efficiency ratio (right), both based on Example 2*.

2. $|M'| > (e-1)|M|$ *in Example 2*, where* e *is Euler's number.*

Part a) points out that full cooperation, i.e., the grand coalition, remains a self-enforcing stable outcome of the game. The remainder of Proposition 5.2 characterizes the structure of less efficient self-enforcing stable outcomes (constrained efficient outcomes). In the present context, inefficiency arises from fragmentation: members of coexisting coalitions fail to internalize the externalities they impose on other coalitions. Part b) of the proposition derives an upper bound on the equilibrium number of coexisting coalitions, $|\boldsymbol{M}|$. In Example 2*, for instance, a stable coalition structure cannot contain more than $\lceil \ln(n) \rceil$ coalitions—at most four coalitions for $n = 50$, five for $n = 100$, and six even for $n = 400$. We note that (i) any individual free rider is counted as a "coalition" in this setting, so our results strictly limit the number of players who can free ride in any stable outcome; and (ii) the theoretical upper bound, $\lceil \ln(n) \rceil$, is not always tight; the left panel of Figure 4 contrasts it with the numerically computed maximal number of coalitions, which is often lower.

This result is driven by a general mechanism that extends beyond these examples. For a fragmented coalition structure to persist, some player must have a strong enough incentive to resist eliminating the fragmentation. For instance, when a few players form a small coalition and the remaining players form a much larger one, the smaller group benefits from this fragmentation and would oppose a merger into the grand coalition. This benefit diminishes, however, when the larger coalition splits into smaller coalitions and the structure becomes more fragmented. With sufficient fragmentation, no player would benefit from maintaining it, and therefore such a highly fragmented structure cannot be perceived as a long-term stable outcome under rational foresight. Equilibrium fragmentation, and its associated inefficiency, can arise only when it delivers sufficiently large benefits to some players.

To quantify the inefficiency that can arise in equilibrium, we define an efficiency measure for coalition structures. Let $\boldsymbol{M}_0$ denote the least efficient coalition structure where all players are free riding ($|\boldsymbol{M}_0| = n$). The efficiency measure of coalition structure $\boldsymbol{M}$ is then defined as the ratio $\frac{\sum_i \left(u_i(\boldsymbol{M}) - u_i(\boldsymbol{M}_0)\right)}{\sum_i \left(u_i(\boldsymbol{N}) - u_i(\boldsymbol{M}_0)\right)}$. This is a utilitarian measure that captures the welfare gain achieved by $\boldsymbol{M}$ relative to the maximized welfare gain achieved by the grand coalition.[28] We compute this efficiency ratio for every stable coalition structure in Example 2*. The right panel of Figure 4 plots, for each n, the minimal equilibrium efficiency ratio, illustrating the worst-case (in)efficiency for each given number of players. In this example, the equilibrium efficiency ratio generally increases with the number of players. Notably, it remains above 0.9 for $n \geq 20$, indicating that inefficiency due to fragmentation is quite limited.

Part c) of the proposition analyzes the composition of stable coalition structures. In contrast to the predictions of one-shot stability, asymmetry among coexisting coalitions is not only possible but also necessary for self-enforcing stability. Any stable fragmented structure requires a player who benefits from the existing fragmentation—and such a player exists only under a sufficient size asymmetry. If a coalition structure contains two coalitions of similar size, members of neither coalitions can effectively free ride on the other; merging them would constitute a Pareto improvement.[29] Accordingly, players do not expect similarly-sized coexisting coalitions to arise in any stable negotiation outcome. In Example 2*, for instance, any two coexisting coalitions in a stable structure must differ in size by a factor of at least $e - 1 \approx 1.72$.

The inequalities established in part c) of Proposition 5.2 are only necessary conditions, providing a minimum degree of asymmetry required for self-enforcing stability. Actual equilibrium coalition structures can, and often do, exhibit even greater asymmetry. Table 3 lists the stable coalition structures of Example 2* for selected values of n, where we use $\{|M_1|, |M_2|, \ldots, |M_L|\}$ as the numeric representation of a coalition structure $\boldsymbol{M} = \{M_1, M_2, \ldots, M_L\}$. Because players in Example 2* are symmetric, if $\boldsymbol{M}$ is self-enforcing stable, any coalition structure with the same numeric representation as $\boldsymbol{M}$ is also self-enforcing stable. For $n = 5$, only $\{5\}$ (grand coalition) and $\{1, 4\}$ are stable coalition structures, meaning at most a single player can free ride. For $n = 200$, numerous coalition structures are stable, but even in the least efficient structure, $\{1, 2, 5, 192\}$, over 95% of the players belong to a single coalition, while the remaining players form small coexisting coalitions. This theoretical prediction—a near-universal agreement coexisting with a small, fragmented fringe—aligns remarkably well with observed patterns of international coop-

[28] This efficiency measure is invariant under affine utility transformations, but not under general monotone transformations (analogous to the Arrow-Pratt measures of risk aversion).

[29] Although members of the merged coalition face higher abatement burdens, the benefits from reduced pollution more than offset the additional abatement costs. Note that this collective rationale for merging also exists under one-shot stability, where members of two-player coalitions would all benefit if they could merge into a four-player coalition. However, the resulting coalition is not stable because short-sighted players would have an incentive to unilaterally defect, expecting to free ride on the remaining three-player coalition.

Table 3: Set of all self-enforcing stable coalition structures in Example 2*

n	$\{\|M_1\|, \|M_2\|, \ldots, \|M_L\|\}$
3	{3}, {1, 2}*
4	{4}, {1, 3}*
5	{5}, {1, 4}*
10	{10}, {1, 9}, {2, 8}, {3, 7}, {1, 2, 7}*
20	{20}, {1, 19}, {2, 18}, {3, 17}, {1, 2, 17}*, {4, 16}, {1, 3, 16}, {5, 15}, {1, 4, 15}, {6, 14}, {1, 5, 14}, {2, 4, 14}, {7, 13}, {1, 6, 13}, {2, 5, 13}
35	{35}, {1, 34}, {2, 33}, {3, 32}, {1, 2, 32}*, {4, 31}, {1, 3, 31}, {5, 30}, {1, 4, 30}, {6, 29}, {1, 5, 29}, {2, 4, 29}, {7, 28}, {1, 6, 28}, {2, 5, 28}, {8, 27}, {1, 7, 27}, {2, 6, 27}, {9, 26}, {1, 8, 26}, {2, 7, 26}, {3, 6, 26}, {10, 25}, {1, 9, 25}, {2, 8, 25}, {3, 7, 25}, {11, 24}, {1, 10, 24}, {2, 9, 24}, {3, 8, 24}, {4, 7, 24}, {12, 23}, {1, 11, 23}, {2, 10, 23}, {3, 9, 23}, {4, 8, 23}, {1, 12, 22}, {2, 11, 22}, {4, 9, 22}
50	{50}, {1, 49}, {2, 48}, {3, 47}, {1, 2, 47}, {4, 46}, {1, 3, 46}, {5, 45}, {1, 4, 45}, {6, 44}, {1, 5, 44}, {2, 4, 44}, {7, 43}, {1, 6, 43}, {2, 5, 43}, {8, 42}, {1, 7, 42}, {2, 6, 42}, {1, 2, 5, 42}*, {9, 41}, {1, 8, 41}, {2, 7, 41}, {3, 6, 41}, {10, 40}, {1, 9, 40}, {2, 8, 40}, {3, 7, 40}, {11, 39}, {1, 10, 39}, {2, 9, 39}, {3, 8, 39}, {4, 7, 39}, {12, 38}, {1, 11, 38}, {2, 10, 38}, {3, 9, 38}, {4, 8, 38}, {13, 37}, {1, 12, 37}, {2, 11, 37}, {3, 10, 37}, {4, 9, 37}, {14, 36}, {1, 13, 36}, {2, 12, 36}, {3, 11, 36}, {4, 10, 36}, {5, 9, 36}, {15, 35}, {1, 14, 35}, {2, 13, 35}, {3, 12, 35}, {4, 11, 35}, {5, 10, 35}, {16, 34}, {1, 15, 34}, {2, 14, 34}, {3, 13, 34}, {4, 12, 34}, {5, 11, 34}, {17, 33}, {1, 16, 33}, {2, 15, 33}, {3, 14, 33}, {4, 13, 33}, {5, 12, 33}, {6, 11, 33}, {18, 32}, {1, 17, 32}, {2, 16, 32}, {3, 15, 32}, {4, 14, 32}, {5, 13, 32}, {6, 12, 32}
200	{200}, {1, 199}, {2, 198}, {3, 197}, {1, 2, 197}, …, {1, 2, 5, 192}*, …, {2, 25, 47, 126}

Note: * indicates that the coalition structure is the least efficient one for given n.

eration. The Paris Agreement, for instance, features broad participation alongside a few holdouts, demonstrating the empirical relevance of our theory even when multiple coalitions can form. Overall, the pattern across all cases in Table 3 reveals that, apart from the grand coalition, stable structures invariably feature substantial asymmetry, characterized by one dominant coalition and at most a few small coexisting groups.

These results echo the findings of Ray and Vohra (2001), who analyze coalition formation in a comparable public goods setting. Their framework, like ours, predicts possible full efficiency, limited fragmentation, and asymmetric coalition structures in equilibrium. However, the underlying logic is fundamentally different. To illustrate the distinction, consider Example 2* with $n = 4$ players. Table 3 shows that both the grand coalition $\{4\}$ and the asymmetric structure $\{1, 3\}$ are self-enforcing stable. By contrast, the solution concept of

Ray and Vohra (2001), when applied to the same example, predicts that the grand coalition is the unique equilibrium. The key difference is commitment. In their framework, once a player chooses to be a free rider and forms a singleton coalition, that player is permanently locked out, making the grand coalition unattainable thereafter. This commitment power makes threats of subsequent inefficient fragmentation credible, where the remaining three players form $\{1, 2\}$ to yield the final state of $\{1, 1, 2\}$. Without commitment, such a threat is empty because all players would prefer to renegotiate to the grand coalition and, therefore, understand that the suggested fragmentation would not be stable. Consequently, in our framework, where participation is non-binding, inefficient outcomes like $\{1, 1, 2\}$ cannot serve as credible deterrents. The logic of (external) stability only eliminates Pareto dominated coalition structures, allowing a structure like $\{1, 3\}$ to persist as a stable outcome.

6 Conclusions

This paper develops a theory of self-enforcing agreements based on a new solution concept: the self-enforcing stable set. The concept synthesizes the unilateral deviation logic of non-cooperative participation games with the farsighted consistency requirement of cooperative stability. A central methodological innovation is to endogenize players' beliefs about the eventual outcomes of negotiations. By treating beliefs as part of equilibrium, we obtain a simple but powerful implication: negotiations can settle only once expectations concentrate on an effectively unique endpoint. This consistency requirement on equilibrium beliefs rules out optimistic free riding, in which defectors rely on others remaining in unstable configurations. As a result, negotiations can converge only after all constrained Pareto improvements have been exhausted, with any remaining inefficiency reflecting the absence of enforceable transfers. The mechanism is general and does not rely on model-specific punishment paths, irreversible commitments, or bargaining protocols.

Applied to canonical models of international environmental agreements, the theory yields predictions that differ starkly from standard myopic participation games. Whereas one-shot stability typically implies small and highly fragmented coalitions, self-enforcing stable sets predict large participation rates that increase with the number of potential signatories. When multiple coalitions coexist, only limited and asymmetric fragmentation can be stable, yielding a theoretical explanation for core-periphery patterns observed in practice. In particular, quadratic abatement costs imply participation rates exceeding 70 percent, and a widely adopted logarithmic specification yields near-universal participation, closely matching observed outcomes under the Paris Agreement. A second-order refinement further sharpens these predictions by selecting the least efficient coalitions among multiple constrained-efficient outcomes.

In the real world, negotiation tactics often appear short-sighted; withdrawal threats, exemption demands, or aggressive bargaining for special treatment are not uncommon. Within

our framework, such behavior may arise from strategic positioning over equilibrium selection. When multiple constrained-efficient outcomes are self-enforcing, parties have incentives to steer negotiations toward their preferred endpoint, including one that grants them a free-riding position or implicit compensation. Identifying exactly which actors can successfully secure such concessions requires a richer model of bargaining power or negotiation protocols, dimensions we deliberately abstract from in the present framework.

References

Aumann, R. J. and R. B. Myerson (1988): "Endogenous formation of links between players and of coalitions: an application of the Shapley value," in *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, ed. by A. E. Roth, Cambridge University Press, chap. 12, 175–191.

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

——— (2005): "The theory of international environmental agreements," in *Handbook of Environmental Economics*, ed. by K.-G. Maler and J. R. Vincent, Elsevier, vol. 3, chap. 28, 1457–1516.

Battaglini, M. and B. Harstad (2016): "Participation and duration of environmental agreements," *Journal of Political Economy*, 124, 160–204.

——— (2020): "The political economy of weak treaties," *Journal of Political Economy*, 128, 544–590.

Breitmeier, H., O. R. Young, and M. Zurn (2006): *Analyzing International Environmental Regimes: From Case Studies to Database*, MIT Press.

Böhringer, C., J. C. Carbone, and T. F. Rutherford (2016): "The strategic value of carbon tariffs," *American Economic Journal: Economic Policy*, 8, 28–51.

Carattini, S., S. Levin, and A. Tavoni (2019): "Cooperation in the climate commons," *Review of Environmental Economics and Policy*, 13, 227–247.

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

Chander, P. and H. Tulkens (1995): "A core-theoretic solution for the design of cooperative agreements on transfrontier pollution," *International Tax and Public Finance*, 2, 279–293.

——— (1997): "The core of an economy with multilateral environmental externalities," *International Journal of Game Theory*, 26, 397–401.

Chatterjee, K., B. Dutta, D. Ray, and K. Sengupta (1993): "A noncooperative theory of coalitional bargaining," *Review of Economic Studies*, 60, 463–477.

Chwe, M. S.-Y. (1994): "Farsighted coalitional stability," *Journal of Economic Theory*, 63, 299–325.

Coase, R. H. (1960): "The problem of social cost," *Journal of Law and Economics*, 3, 1–44.

d'Aspremont, C., A. Jacquemin, J. J. Gabszewicz, and J. A. Wymark (1983): "On the stability of collusive price leadership," *Canadian Journal of Economics*, 16, 17–25.

Diamantoudi, E. and E. S. Sartzetakis (2015): "International environmental agreements: coordinated action under foresight," *Economic Theory*, 59, 527–546.

——— (2018): "International environmental agreements: the role of foresight," *Environmental and Resource Economics*, 71, 241–257.

Diamantoudi, E. and L. Xue (2007): "Coalitions, agreements and efficiency," *Journal of Economic Theory*, 136, 105–125.

Dixit, A. and M. Olson (2000): "Does voluntary participation undermine the Coase Theorem?" *Journal of Public Economics*, 76, 309–335.

Dutta, B. and R. Vohra (2017): "Rational expectations and farsighted stability," *Theoretical Economics*, 12, 1191–1227.

Dutta, P. K. and R. Radner (2025): "Climate payments: A Coase theorem," *Journal of Economic Theory*, 228, 106032.

Farrokhi, F. and A. Lashkaripour (2025): "Can trade policy mitigate climate change?" *Econometrica*, 93, 1561–1599.

Finus, M. (2001): *Game Theory and International Environmental Cooperation*, Edward Elgar.

Gerber, A. and P. C. Wichardt (2009): "Providing public goods in the absence of strong institutions," *Journal of Public Economics*, 93, 429–439.

Germain, M., P. Toint, H. Tulkens, and A. de Zeeuw (2003): "Transfers to sustain dynamic core-theoretic cooperation in international stock pollutant control," *Journal of Economic Dynamics & Control*, 28, 79–99.

Gersbach, H. and R. Winkler (2011): "International emission permit markets with refunding," *European Economic Review*, 55, 759–773.

Golosov, M., J. Hassler, P. Krusell, and A. Tsyvinski (2014): "Optimal taxes on fossil fuel in general equilibrium," *Econometrica*, 82, 41–88.

Harsanyi, J. C. (1974): "An equilibrium-point interpretation of stable sets and a proposed alternative definition," *Management Science*, 20, 1472–1495.

Harstad, B. (2016): "The dynamics of climate agreements," *Journal of the European Economic Association*, 14, 719–752.

——— (2023a): "Pledge-and-review bargaining," *Journal of Economic Theory*, 207, 105574.

——— (2023b): "Pledge-and-review bargaining: From Kyoto to Paris," *Economic Journal*, 133, 1181–1216.

——— (2024): "On international cooperation," in *Handbook of the Economics of Climate Change*, Elsevier, vol. 1, 249–295.

Harstad, B. and A. S. Kessler (2025): "Present bias in politics and self-committing treaties," *Journal of Public Economics*, 246, 105372.

Harstad, B., F. Lancia, and A. Russo (2019): "Compliance technology and self-enforcing agreements," *Journal of the European Economic Association*, 17, 1–29.

Hart, S. and M. Kurz (1983): "Endogenous formation of coalitions," *Econometrica*, 51, 1047–1064.

Hassler, J., P. Krusell, and A. A. Smith, Jr. (2016): "Environmental macroeconomics," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and H. Uhlig, Elsevier, vol. 2, chap. 24, 1893–2008.

Hoel, M. (1992): "International environmental conventions: the case of uniform reductions of emissions," *Environmental and Resource Economics*, 2, 141–159.

Iverson, T. (2025): "Tiered climate clubs: Global abatement without global agreement," SSRN Working Paper No. 4769577.

Karp, L. S. and H. Sakamoto (2021): "Sober optimism and the formation of international environmental agreements," *Journal of Economic Theory*, 197, 105321.

Kerr, S., S. Lippert, and E. Y. Lou (2025): "Transfers in climate action teams," *Economic Theory*, 80, 595–618.

Kortum, S. S. and D. Weisbach (2024): "Optimal unilateral carbon policy," SSRN Working Paper No. 3958930.

Myerson, R. B. (1977): "Graphs and cooperation in games," *Mathematics of Operations Research*, 2, 225–229.

Nordhaus, W. D. (1977): "Economic growth and climate: the carbon dioxide problem," *American Economic Review*, 57, 341–346.

——— (2015): "Climate clubs: overcoming free-riding in International Climate Policy," *American Economic Review*, 105, 1339–1370.

Okada, A. (2023): "Dynamic bargaining with voluntary participation and externalities," *Economic Theory*, 75, 427–452.

Palfrey, T. R. and H. Rosenthal (1984): "Participation and the provision of discrete public goods: a strategic analysis," *Journal of Public Economics*, 24, 171–193.

Ray, D. (2007): *A Game-Theoric Perspective on Coalition Formation*, Oxford University Press.

Ray, D. and R. Vohra (1997): "Equilibrium binding agreements," *Journal of Economic Theory*, 26, 286–336.

——— (1999): "A theory of endogenous coalition structures," *Games and Economic Behavior*, 26, 286–336.

——— (2001): "Coalitional power and public goods," *Journal of Political Economy*, 109, 1355–1384.

——— (2015a): "Coalitin Formation," in *Handbook of Game Theory with Economic Applications*, ed. by P. Young and S. Zamir, Elsevier, vol. 4, chap. 5, 239–326.

——— (2015b): "The farsighted stable set," *Econometrica*, 83, 977–1011.

——— (2019): "Maximality in the farsighted stable set," *Econometrica*, 87, 1763–1779.

Rubinstein, A. (1982): "Perfect equilibrium in a bargaining model," *Econometrica*, 50, 97–109.

UNFCCC (2016): *Paris Agreement*, The United Nations Framework Convention on Climate Change, 21st Conference of the Parties, Paris.

von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton University Press.

Vosooghi, S., M. Arvaniti, and F. van der Ploeg (2024): "Climate coalitions with sophisticated policy makers," SSRN Working Paper No. 4769577.

Young, O. R. (2011): "Effectiveness of international environmental regimes: existing knowledge, cutting-edge themes, and research strategies," *Proceedings of the National Academy of Sciences*, 108, 19853–19860.

A Sequential bargaining perspective

This section provides an alternative perspective on our solution concept using a sequential bargaining game. Following Rubinstein (1982) and Chatterjee et al. (1993), consider a sequential bargaining process in which a randomly selected player proposes a coalition. If the proposal is accepted unanimously, the game ends. Otherwise, a new proposer is chosen by nature, and the game continues. Let π_i be the (stationary) probability that player $i \in N$ is selected by nature. These probabilities need not be uniform; we require only that $\pi_i > 0$ for all i, ensuring that every player has a positive chance to propose. The game has an infinite time horizon and no discounting. A strategy of player i is a pair (μ_i, σ_i), where $\mu_i \in \mathcal{N}$ is the coalition that i proposes when selected, and $\sigma_i : \mathcal{N} \to \{0, 1\}$ specifies i's response to others' proposals. Here, $\sigma_i(M) = 1$ indicates acceptance of coalition M, while $\sigma_i(M) = 0$ indicates rejection. If the game does not end in a finite number of steps, the players' payoffs are given by a disagreement point $(d_i)_{i \in N}$. We assume that the disagreement is the worst-case scenario; that is, $d_i \leq \min_{M \in \mathcal{N}} u_i(M)$ for all $i \in N$.

A stationary subgame perfect equilibrium of this game is characterized by a strategy profile $(\mu_i, \sigma_i)_{i \in N}$ such that for each $i \in N$, the choice $\mu_i \in \mathcal{N}$ solves

$$\max_{\mu \in \mathcal{N}} \left\{ \prod_{j \in N} \sigma_j(\mu) u_i(\mu) + \left(1 - \prod_{j \in N} \sigma_j(\mu) \right) \bar{v}_i \right\} \tag{16}$$

and for each $M' \in \mathcal{N}$, the acceptance decision $\sigma_i(M') \in \{0, 1\}$ is given by

$$\sigma_i(M') = \begin{cases} 1 & \text{if } u_i(M') \geq \bar{v}_i \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $\bar{v}_i$ denotes the continuation value when a proposal is rejected. If $\prod_{k \in N} \sigma_k(\mu_j) = 0$ for all j, no coalition forms and the continuation value becomes $\bar{v}_i = d_i$. Otherwise, $\bar{v}_i$ solves

$$\bar{v}_i = \sum_{j \in N} \left[\prod_{k \in N} \sigma_k(\mu_j) u_i(\mu_j) + \left(1 - \prod_{k \in N} \sigma_k(\mu_j) \right) \bar{v}_i \right] \pi_j$$

or equivalently,

$$\bar{v}_i = \sum_{j \in N} p(\mu_j) u_i(\mu_j), \quad \text{where} \quad p(\mu_j) := \frac{\prod_{k \in N} \sigma_k(\mu_j) \pi_j}{\sum_{l \in N} \prod_{k \in N} \sigma_k(\mu_l) \pi_l}. \tag{18}$$

Condition (16) requires that each player's proposal maximize the expected payoff, while condition (17) ensures that players reject proposals if and only if doing so makes them strictly better off. Note that, in equilibrium, the function p defined in (18) provides a micro-foundation of the equilibrium belief we introduced in the definition of self-enforcing

stable sets. It represents the equilibrium probability that a given coalition eventually forms when a proposal is rejected.[30]

The following proposition establishes the equivalence between the set of equilibrium coalitions in this game and the set of self-enforcing stable sets. Every self-enforcing stable set can be supported as an outcome of the sequential bargaining process. Conversely, each equilibrium of the sequential bargaining corresponds to a self-enforcing stable set.

Proposition A.1. *The following are equivalent:*

1. $\mathcal{M}$ *is a self-enforcing stable set;*
2. *there exists an equilibrium strategy profile* $(\mu_i, \sigma_i)_{i\in N}$ *of the sequential bargaining such that*

$$\mathcal{M} = \{M \in \mathcal{N} \mid M \sim \mu_j \text{ for some } j \in N \text{ and } \sigma_i(M) = 1 \text{ for all } i \in N\}. \tag{19}$$

Proof. Suppose that $\mathcal{M}$ is a self-enforcing stable set. By Theorem 3.1, we know that there exists a Pareto efficient coalition $M \in \mathcal{N}$ such that $\mathcal{M} = \{M' \in \mathcal{N} \mid M' \sim M\}$. Consider a strategy profile $(\mu_i, \sigma_i)_{i\in N}$ defined by $\mu_i \in \mathcal{M}$ for all $i \in N$ and

$$\sigma_i(M') = \begin{cases} 1 & \text{if } u_i(M') \geq u_i(M) \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in N$. Then equality (19) holds. We show that this strategy profile is consistent with (16) and (17). Since $u_i(\mu_j) = u_i(M)$ for all $j \in N$, we have $\bar{v}_i = u_i(M)$ for each $i \in N$, which immediately implies that σ_i is consistent with (17). Also, since M is a Pareto efficient coalition,

$$u_i(M') > \bar{v}_i \implies u_j(M') < \bar{v}_j \text{ for some } j \in N \implies \prod_{j\in N} \sigma_j(M') = 0,$$

implying that μ_i actually solves (16).

Conversely, suppose that $(\mu_i, \sigma_i)_{i\in N}$ is an equilibrium strategy profile of the sequential bargaining. We show that $\mathcal{M}$ defined by (19) is an effectively singleton set containing a Pareto efficient coalition, which by Theorem 3.1 implies that $\mathcal{M}$ is a self-enforcing stable set. Notice first that $\mathcal{M}$ is nonempty because the disagreement point gives the worst payoff and, thus, at least one proposal is unanimously accepted in equilibrium. Moreover, $\mathcal{M}$ cannot contain distinct coalitions because otherwise there would exist $M, M' \in \mathcal{M}$ such that $u_i(M') > \bar{v}_i > u_i(M)$ for some $i \in N$, which contradicts the fact that $\sigma_i(M) = 1$. It follows that $\mathcal{M} = \{M' \in \mathcal{M} \mid M' \sim M\}$ for some $M \in \mathcal{N}$ and therefore $\mu_i \sim M$ for all

[30] If the same coalition is proposed by more than one player, say by j and $j' \neq j$, then the probability assigned to the coalition is the sum of the individual probabilities: $p(\mu_j) + p(\mu_{j'})$.

$i \in N$. If this M is not a Pareto efficient coalition, there would exist another coalition $M' \in \mathcal{N}$ that Pareto dominates M (and it thus Pareto dominates μ_i for all $i \in N$), which implies that $\sigma_i(M') = 1$ for all $i \in N$ and $u_{i'}(M') > \bar{v}_{i'}$ for some $i' \in N$. But this means that $\mu_{i'}$ does not solve (16), a contradiction. Therefore, the set $\mathcal{M}$ defined by (19) is a self-enforcing stable set. □

B Heterogeneous beliefs

In the main text we *assume* that in equilibrium, all players share a common belief over the set of self-enforcing stable coalitions. Here, we relax this assumption and show how the common belief can be interpreted as an outcome rather than a premise.

For each player $i \in N$, let p_i denote a probability measure on $\mathcal{N}$ representing player i's belief. We denote the set of coalitions with positive support under p_i by $\mathcal{M}_{p_i}$, and write $\mathbb{E}_{p_i} u_j(\cdot) := \sum_{M \in \mathcal{M}_{p_i}} p_i(M) u_j(M)$, the expected payoff of player j under the belief of player i. Given the set of all players' beliefs, the intersection $\mathcal{M} = \cap_{i \in N} \mathcal{M}_{p_i}$ specifies the set of coalitions that receive strictly positive probability from all players. For this game, we define self-enforcing stable sets as follows:

Definition B.1 (Self-enforcing stable set)**.** A nonempty subset $\mathcal{M} \subset \mathcal{N}$ is a self-enforcing stable set if there exists a belief profile $(p_i)_{i \in N}$ such that $\mathcal{M} = \cap_{i \in N} \mathcal{M}_{p_i}$ and

$$M \in \mathcal{M}_{p_j} \iff u_i(M) \geq \mathbb{E}_{p_j} u_i(\cdot) \quad \forall i \in N \tag{20}$$

for each $j \in N$.

This definition naturally extends Definition 2.7 introduced in the main text. The right-arrow in (20) requires internal stability: a coalition can only be believed to be stable by player j if no one has an incentive to deviate *under the belief of player j*. The left-arrow (20) corresponds to external stability: if player j believes that every player weakly prefers a particular coalition to the expected outcome of defection, then that coalition must be considered stable by player j. In equilibrium, both the internal and external stability conditions must hold for all players. This implies that even if we allow players to begin a negotiation process with different beliefs, they must converge to a common belief in any stable outcome of the negotiation.

Proposition B.1. *If $\mathcal{M}$ is a self-enforcing stable set as defined in Definition B.1, the corresponding equilibrium belief profile $(p_i)_{i \in N}$ must satisfy $\mathcal{M}_{p_i} = \mathcal{M}$ for all $i \in N$. Consequently, the equilibrium outcomes are equivalent to those derived under the common-belief assumption in Definition 2.7.*

Proof. Consider an arbitrary player $j \in N$. Following the same logic as in the common-belief case, the internal stability condition requires that $\mathcal{M}_{p_j}$ cannot contain distinct coalitions

(see Lemma 3.1). Then, as shown in Theorem 3.1, the external stability condition implies that $\mathcal{M}_{p_j}$ must be the indifference class $\tilde{M}_j$ for some Pareto efficient coalition M_j. This Pareto efficient coalition M_j may vary across players. For a self-enforcing stable set to exist, however, the intersection $\cap_{j\in N}\mathcal{M}_{p_j}$ must be nonempty. This is possible only if the indifference class that each player believes to be stable is identical, i.e., $M_j \sim M_k$ for all $j, k \in N$. Therefore, in equilibrium, all players must share a common belief. □

What plays an important role here is the requirement that the intersection $\cap_{i\in N}\mathcal{M}_{p_i}$ must be nonempty in equilibrium. On its own, this condition does not imply belief convergence; it merely ensures that all players assign positive probability to at least one common outcome. This is arguably the minimal requirement for a belief profile to constitute an equilibrium, as any outcome of the negotiation would otherwise contradict the belief of at least one player. When this basic requirement is combined with the internal and external stability conditions, it implies that players' beliefs must converge. We emphasize that our approach does not characterize how such convergence occurs. The details of that process are likely context-dependent and would differ across negotiation environments.

C Details governing payoffs in the examples

C.1 Payoff structures in Examples 1 and 2

Example 1. Consider any coalition $M \in \mathcal{N}$. A free-riding player $i \notin M$ accounts for only private damages from emissions and solves

$$\max_{g_i}\left\{-\frac{1}{\gamma}(\bar{g}_i - g_i)^\gamma - \xi\sum_{j\in N} g_j\right\},$$

leading to the first order condition

$$(\bar{g}_i - g_i)^{\gamma-1} - \xi = 0 \implies g_i = \bar{g}_i - \xi^{\frac{1}{\gamma-1}},$$

which satisfies $0 \le g_i \le \bar{g}_i$ as we assume $\bar{g}_i \ge n^{\frac{1}{\gamma-1}}\xi^{\frac{1}{\gamma-1}}$. On the other hand, a member $i \in M$ of the coalition internalizes the damages to the other members and solves

$$\max_{g_i}\left\{-\frac{1}{\gamma}(\bar{g}_i - g_i)^\gamma - \sum_{k\in M}\xi\sum_{j\in N} g_j\right\},$$

leading to the first order condition

$$(\bar{g}_i - g_i)^{\gamma-1} - |M|\xi = 0 \implies g_i = \bar{g}_i - (|M|\xi)^{\frac{1}{\gamma-1}} \ge 0. \tag{21}$$

Hence, the $|M|$ coalition members each emit $\bar{g}_i - (|M|\xi)^{\frac{1}{\gamma-1}}$ and the $n-|M|$ free riders each emit $\bar{g}_i - \xi^{\frac{1}{\gamma-1}}$. Inserting this result back into the payoffs of the free riders implies

$$\begin{aligned}
u^{\text{free}}(M) &= -\frac{1}{\gamma}\left(\bar{g}_i - \left(\bar{g}_i - \xi^{\frac{1}{\gamma-1}}\right)\right)^{\gamma} - \xi\Bigg(\sum_{j\in N\setminus M}\left(\bar{g}_j - \xi^{\frac{1}{\gamma-1}}\right) + \sum_{j\in M}\left(\bar{g}_j - (|M|\xi)^{\frac{1}{\gamma-1}}\right)\Bigg) \\
&= -\frac{1}{\gamma}\left(\xi^{\frac{1}{\gamma-1}}\right)^{\gamma} - \xi\sum_{j\in N}\bar{g}_j + \xi\left((n-|M|)\xi^{\frac{1}{\gamma-1}} + |M||M|^{\frac{1}{\gamma-1}}\xi^{\frac{1}{\gamma-1}}\right) \\
&= -\frac{1}{\gamma}\xi^{\frac{\gamma}{\gamma-1}} - \xi\sum_{j\in N}\bar{g}_j + \left(n-|M|+|M|^{\frac{\gamma}{\gamma-1}}\right)\xi^{\frac{\gamma}{\gamma-1}}
\end{aligned}$$

and similarly for members of the coalition

$$\begin{aligned}
u^{\text{member}}(M) &= -\frac{1}{\gamma}\left(\bar{g}_i - \left(\bar{g}_i - (|M|\xi)^{\frac{1}{\gamma-1}}\right)\right)^{\gamma} - \xi\sum_{j\in N}\bar{g}_j + \left(n-|M|+|M|^{\frac{\gamma}{\gamma-1}}\right)\xi^{\frac{\gamma}{\gamma-1}} \\
&= -\frac{1}{\gamma}|M|^{\frac{\gamma}{\gamma-1}}\xi^{\frac{\gamma}{\gamma-1}} - \xi\sum_{j\in N}\bar{g}_j + \left(n-|M|+|M|^{\frac{\gamma}{\gamma-1}}\right)\xi^{\frac{\gamma}{\gamma-1}}
\end{aligned}$$

delivering equation (1).

Example 2. In the second example, the free-riders face

$$\max_{g_i}\left\{\ln(g_i) - \xi\sum_{j\in N} g_j\right\} \implies g_i = \frac{1}{\xi}$$

and the members of the coalition face

$$\max_{g_i}\left\{\ln(g_i) - \sum_{k\in M}\xi\sum_{j\in N} g_j\right\} \implies g_i = \frac{1}{|M|\xi}. \tag{22}$$

Inserting these emissions back into the payoff functions yields

$$\begin{aligned}
u^{\text{free}}(M) &= -\ln(\xi) - \xi\left(\sum_{j\in N\setminus M}\frac{1}{\xi} + \sum_{j\in M}\frac{1}{|M|\xi}\right) \\
&= -\ln(\xi) - n - 1 + |M|
\end{aligned}$$

and

$$u^{\text{member}}(M) = -\ln(|M|\xi) - (n-|M|) - 1 = -\ln(\xi) - \ln(|M|) - n - 1 + |M|,$$

delivering equation (2).

C.2 Payoff structures in Examples 1* and 2*

Example 1*. Following the derivation of equation (21), members of a coalition $M_l \in \boldsymbol{M}$ face the first order condition

$$g_i = \bar{g}_i - (|M_l|\xi)^{\frac{1}{\gamma-1}}.$$

Inserting these results back into the payoff function of a member $i \in M_l$ yields

$$\begin{aligned} u_i(\boldsymbol{M}) &= -\frac{1}{\gamma}\left(\bar{g}_i - \left(\bar{g}_i - (|M_l|\xi)^{\frac{1}{\gamma-1}}\right)\right)^{\gamma} - \xi\left(\sum_{M_k\in\boldsymbol{M}}\sum_{j\in M_k}\left(\bar{g}_j - (|M_k|\xi)^{\frac{1}{\gamma-1}}\right)\right) \\ &= -\frac{1}{\gamma}|M_l|^{\frac{\gamma}{\gamma-1}}\xi^{\frac{\gamma}{\gamma-1}} - \xi\sum_{j\in N}\bar{g}_j + \xi\sum_{M_k\in\boldsymbol{M}}|M_k|(|M_k|\xi)^{\frac{1}{\gamma-1}} \\ &= \xi^{\frac{\gamma}{\gamma-1}}\left(\sum_{M_k\in\boldsymbol{M}}|M_k|^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma}|M_l|^{\frac{\gamma}{\gamma-1}}\right) - \xi\sum_{j\in N}\bar{g}_j, \end{aligned}$$

delivering equation (14) of Examples 1*.

Example 2*. Following the derivation of equation (22), members of a coalition $M_l \in \boldsymbol{M}$ face the first order condition

$$g_i = \frac{1}{|M_l|\xi}.$$

Inserting these results back into the payoff function of a member $i \in M_l$ yields

$$\begin{aligned} u_i(\boldsymbol{M}) &= -\ln(|M_l|\xi) - \xi\left(\sum_{M_k\in\boldsymbol{M}}\sum_{j\in M_k}\frac{1}{|M_k|\xi}\right) \\ &= -\ln(\xi) - \ln(|M_l|) - \xi\sum_{M_k\in\boldsymbol{M}}\frac{1}{\xi} = -\ln(\xi) - \ln(|M_l|) - |\boldsymbol{M}|, \end{aligned}$$

delivering equation (15) of Example 2*.

D Proofs

D.1 Proof of Remark 1

Proof. Let $\mathcal{M}$ be a vNM stable set. We shall show that $\mathcal{M}$ is empty. We start with three observations. The grand coalition, $\{1,2,3,4\}$, Pareto dominates any coalition with 2 or fewer members.[31] A coalition with three members, e.g., $\{1,2,3\}$, Pareto dominates the scenario where the "coalition" is a singleton.[32] A coalition with two members is only dominated by

[31] The members of the grand coalition have payoff level $u_i^{\text{member of 4}} = -\ln(\xi) - \ln(4) - 1$, which is larger than the payoff level of the free riders on a coalition of two $u_i^{\text{free on 2}} = -\ln(\xi) - 2 - 1$.

[32] $u_i^{\text{member of 3}} = -\ln(\xi) - \ln(3) - 1 - 1 > -\ln(\xi) - 3 - 1 = u_i^{\text{free}}$ where all u_i^{free} is the payoff when all players are free riding.

$\mathcal{M}$ if the grand coalition is in $\mathcal{M}$.[33]

Suppose a coalition of three is part of the solution set, e.g., $\{1,2,3\} \in \mathcal{M}$. Since $\mathcal{M}$ is a vNM stable set, $\{1,2,3\}$ cannot be dominated by $\mathcal{M}$. As a result $\{1,2\} \notin \mathcal{M}$ because $u_3(\{1,2\}) = -\ln(\xi) - 3 > -\ln(\xi) - \ln(3) - 2 = u_3(\{1,2,3\})$. Then we must have $\{1,2,3,4\} \in \mathcal{M}$ because otherwise $\{1,2\}$ would not be dominated by $\mathcal{M}$. But this is a contradiction because $\{1,2,3,4\}$ is dominated by $\{1,2,3\}$. The same argument applies to other coalitions with three members.

Now suppose $\{1,2,3,4\} \in \mathcal{M}$. Since $\mathcal{M}$ is a vNM stable set, the grand coalition cannot be dominated by $\mathcal{M}$. However, $\{1,2,3,4\}$ would be dominated by $\mathcal{M}$ if $\{1,2,3\} \in \mathcal{M}$ because $u_4(\{1,2,3\}) = -\ln(\xi) - 2 > -\ln(\xi) - \ln(4) - 1 = u_4(\{1,2,3,4\})$. Hence, $\{1,2,3\} \notin \mathcal{M}$, implying that $\{1,2,3\}$ must be dominated by $\mathcal{M}$. Because a coalition of three Pareto dominates the case of singleton coalitions, it follows that $\{1,2\} \in \mathcal{M}$, $\{1,3\} \in \mathcal{M}$, or $\{2,3\} \in \mathcal{M}$. But these coalitions would then be dominated by $\mathcal{M}$ because $\{1,2,3,4\} \in \mathcal{M}$, a contradiction.

Given $\{1,2,3,4\} \notin \mathcal{M}$, it has to be dominated by $\mathcal{M}$. However, the grand coalition is only dominated by a coalition of three and we have just shown that a coalition of three cannot be part of $\mathcal{M}$. Thus, the solution set is empty. □

D.2 Proof of Lemma 3.1

Proof. Let $\mathcal{M}$ be a self-enforcing stable set. Suppose, by way of contradiction, that $\mathcal{M}$ contains distinct coalitions. Then there exists a player $i \in N$ and coalitions $M', M'' \in \mathcal{M}$ such that $u_i(M') = \min_{M \in \mathcal{M}} u_i(M) < u_i(M'')$. It follows that

$$u_i(M') < p(M')u_i(M') + \sum_{M \in \mathcal{M} \setminus \{M'\}} p(M)u_i(M) = \mathbb{E}_p u_i(\cdot),$$

which, together with the right arrow in condition (8), implies $M' \notin \mathcal{M}$, a contradiction. □

D.3 Proof of Theorem 3.1

Proof. Sufficiency ("if"): Suppose that $\mathcal{M} = \tilde{M}$ for some Pareto efficient coalition M. Then

$$M' \in \mathcal{M} \implies u_i(M') = u_i(M) = \mathbb{E}_p u_i(\cdot) \quad \forall i \in N$$

[33] The first statement established that the grand coalition dominates a coalition of two. A member defecting from a coalition of two would reduce payoff from $u_i^{\text{member of 2}} = -\ln(\xi) - \ln(2) - 2 - 1$ to the lower value $u_i^{\text{free}} = -\ln(\xi) - 3 - 1$, thus, a coalition of two cannot be dominated by everyone free riding. A free rider joining a coalition of two would reduce his or her payoff from $u_i^{\text{free on 2}} = -\ln(\xi) - 2 - 1$ to $u_i^{\text{member of 3}} = -\ln(\xi) - \ln(3) - 1 - 1$. Thus, a coalition of two cannot be dominated by a coalition of three.

whereas Pareto efficiency, together with the fact that any coalition not in $\mathcal{M}$ has to be distinct from M, implies

$$M' \notin \mathcal{M} \implies \exists\, i \in N \text{ s.th. } u_i(M') < u_i(M) = \mathbb{E}_p u_i(\cdot).$$

Therefore, $\mathcal{M}$ is a self-enforcing stable set.

Necessity ("only if"): Let $\mathcal{M}$ be an arbitrary self-enforcing stable set and pick $M \in \mathcal{M}$. By Lemma 3.1, we know that $\mathcal{M}$ cannot contain distinct coalitions, which implies that

$$\mathcal{M} \subset \{M' \in \mathcal{N} \mid M' \sim M\}.$$

Since $u_i(M) = \mathbb{E}_p u_i(\cdot)$ for all $i \in N$, any coalition that is not distinct from M must be included in $\mathcal{M}$ (by the left arrow in condition (8)). Hence,

$$\mathcal{M} = \{M' \in \mathcal{N} \mid M' \sim M\}.$$

It remains to show that M is Pareto efficient. Suppose that M is not Pareto efficient. Then there exists a coalition M' that is distinct from M and $u_i(M') \geq u_i(M) = \mathbb{E}_p u_i(\cdot)$ for all $i \in N$. Since $\mathcal{M}$ is a self-enforcing stable set, this inequality implies that M' must be included in $\mathcal{M}$, contradicting the fact that $\mathcal{M}$ cannot contain distinct coalitions. □

D.4 Proof of Proposition 3.1

Proof. The proof is divided into three parts.

Part 1: *Establishing equation* (9)
By Theorem 3.1, we only have to identify the set of Pareto efficient coalitions. A necessary condition for a coalition to be Pareto efficient is that the coalition is not Pareto dominated by the grand coalition. This condition also turns out to be sufficient (see below). Members are always better off in the grand coalition than in any other coalition. Thus, for coalition M not to be Pareto dominated, the free riders on coalition M must be strictly better off than the members of the grand coalition, which leads to the condition

$$\frac{\gamma - 1}{\gamma} n^{\frac{\gamma}{\gamma - 1}} - n < |M|^{\frac{\gamma}{\gamma - 1}} - |M| - \frac{1}{\gamma}. \tag{23}$$

Since the right side is increasing in $|M|$, this inequality is equivalent to $|M| > z_\gamma(n)$, where $z_\gamma(n)$ is the positive root defined in the proposition (see below for existence and uniqueness of such $z_\gamma(n)$).

To complete the argument, we show that every inefficient coalition is Pareto dominated by the grand coalition. First, a coalition cannot be Pareto dominated by another coalition

of the same or smaller size, because free riders always receive higher payoffs than members and both members and free riders are better off when the coalition is larger. This means that any inefficient coalition must be Pareto dominated by a larger coalition. A coalition is Pareto dominated by a larger coalition if and only if free riders in the smaller coalition are weakly better off as members of the larger coalition. Since members' payoffs are highest in the grand coalition, it follows that any coalition Pareto dominated by a larger coalition must also be Pareto dominated by the grand coalition. This establishes that equation (9) characterizes the set of self-enforcing stable sets.

Part 2: *Participation sensitivity to* γ

To analyze the properties of the threshold value $z_\gamma(n)$, we define the function:

$$F(z,\theta,n) := \underbrace{\frac{1}{\theta}(n^\theta - 1)}_{=:g(\theta,n)} - n - z^\theta + z + 1$$

for $z \geq 0$, $\theta > 0$, and $n > 1$ (all treated as real numbers). The threshold $z_\gamma(n)$ discussed in the main text solves $F(z,\theta,n) = 0$ for given parameters $\theta = \frac{\gamma}{\gamma-1} \in (1,2]$ and $n \geq 3$. To understand how this root changes with the parameters, we first characterize the function $F(z,\theta,n)$ itself. The partial derivatives of $F(z,\theta,n)$ are as follows:

$$F_z(z,\theta,n) = 1 - \theta z^{\theta-1}, \qquad F_\theta(z,\theta,n) = \underbrace{\frac{1}{\theta}n^\theta \ln(n) - \frac{1}{\theta^2}(n^\theta - 1)}_{g_\theta(\theta,n)} - z^\theta \ln(z).$$

$$F_{zz}(z,\theta,n) = -\theta(\theta-1)z^{\theta-2}$$

$$F_{\theta\theta}(z,\theta,n) = \underbrace{\frac{1}{\theta}n^\theta \ln(n)\ln(n) - \frac{2}{\theta^2}n^\theta \ln(n) + \frac{2}{\theta^3}(n^\theta - 1)}_{g_{\theta\theta}(\theta,n)} - z^\theta \ln(z)\ln(z)$$

$$F_{z\theta}(z,\theta,n) = F_{\theta z}(z,\theta,n) = -(\theta \ln(z) + 1)z^{\theta-1}$$

The function $g(\theta,n)$ has the integral representation

$$g(\theta,n) = \frac{1}{\theta}(n^\theta - 1) = \ln(n)\int_0^1 n^{\theta t}dt$$

because $\int_0^1 n^{\theta t}dt = \frac{n^\theta - 1}{\ln(n^\theta)}$. One implication of this form is that

$$\frac{\partial^k g(\theta,n)}{\partial\theta^k} = (\ln(n))^{k+1}\int_0^1 t^k n^{\theta t}dt > 0 \quad \forall k = 0,1,2,\ldots,$$

where $\frac{\partial^0 g(\theta,n)}{\partial\theta^0} := g(\theta,n)$, so all partial derivatives of $g(\theta,n)$ with respect to θ as well as $g(\theta,n)$ itself are strictly positive. Furthermore, the integral representation implies that

$\ln(\partial^k g(\theta,n)/\partial\theta^k)$ is strictly convex with respect to θ:

$$\frac{\partial^{k+2} g(\theta,n)}{\partial\theta^{k+2}}\frac{\partial^k g(\theta,n)}{\partial\theta^k} > \frac{\partial^{k+1} g(\theta,n)}{\partial\theta^{k+1}}\frac{\partial^{k+1} g(\theta,n)}{\partial\theta^{k+1}} \quad \forall k = 0,1,2,\ldots \tag{24}$$

because its integral form, $(\int_0^1 t^{k+2}n^{\theta t}dt)(\int_0^1 t^k n^{\theta t}dt) > (\int_0^1 t^{k+1}n^{\theta t}dt)^2$, is equivalent to

$$\left(\int_0^1 (t^{\frac{k+2}{2}})^2 n^{\theta t}dt\right)^{\frac{1}{2}}\left(\int_0^1 (t^{\frac{k}{2}})^2 n^{\theta t}dt\right)^{\frac{1}{2}} > \int_0^1 t^{\frac{k+2}{2}}t^{\frac{k}{2}}n^{\theta t}dt,$$

which follows from the Cauchy-Schwarz inequality. Note that this inequality holds for $k=0$, meaning that $\ln(g(\theta,n))$ is also strictly convex in θ.

The following lemma provides a general insight into how $F(z,\theta,n)$ changes with θ, playing a critical role in our subsequent proof. We note that z in the lemma need not be a root of $F(z,\theta,n)=0$. Also, θ can be equal to (or even less than) 1, as this is required for the boundary analysis in a later step of the proof.

Lemma D.1. *Let $n \geq 12$, $z > 1$, and $\theta \in (0,2]$. If $F_\theta(z,\theta,n) = 0$, then $F_{\theta\theta}(z,\theta,n) < 0$.*

Proof. We first use the condition $F_\theta(z,\theta,n)=0$ to eliminate z. The condition $F_\theta(z,\theta,n)=0$ is equivalent to $g_\theta(\theta,n) = z^\theta \ln(z)$, and thus to $\ln(g_\theta(\theta,n)) = \theta\ln(z) + \ln(\ln(z))$. Using these expressions, we can rewrite the desired conclusion, $F_{\theta\theta}(z,\theta,n)<0$, solely in terms of θ and n. The inequality $F_{\theta\theta}(z,\theta,n)<0$ is equivalent to $g_{\theta\theta}(\theta,n) < z^\theta \ln(z)\ln(z)$, and thus

$$\begin{aligned}
F_{\theta\theta}(z,\theta,n) < 0 &\iff \frac{g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)} < \ln(z) \\
&\iff \theta\frac{g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)} + \ln\left(\frac{g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)}\right) < \underbrace{\theta\ln(z) + \ln(\ln(z))}_{\ln(g_\theta(\theta,n))} \\
&\iff \underbrace{\ln(g_\theta(\theta,n)) - \theta\frac{g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)} - \ln\left(\frac{g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)}\right)}_{=:G(\theta,n)} > 0.
\end{aligned}$$

Hence, the proof is complete if we can show $G(\theta,n) > 0$ for all $\theta \in (0,2]$ and for all $n \geq 12$.

We proceed in two steps. First, observe that $G(\theta,n)$ is a strictly decreasing function of θ because the partial derivative is negative:

$$G_\theta(\theta,n) = -\underbrace{\frac{g_{\theta\theta\theta}(\theta,n)g_\theta(\theta,n) - g_{\theta\theta}(\theta,n)g_{\theta\theta}(\theta,n)}{g_\theta(\theta,n)g_\theta(\theta,n)}}_{>0 \text{ because of (24) for } k=1}\left(\theta + \frac{g_\theta(\theta,n)}{g_{\theta\theta}(\theta,n)}\right) < 0,$$

which means that, for each n, $G(\theta,n) > 0$ for $\theta \in (0,2]$ if and only if $G(2,n) > 0$. Second, we show that $G(2,n) > 0$ for all $n \geq 12$. To verify this inequality, differentiate $G(2,n)$ with

respect to n and observe

$$\frac{\partial G(2,n)}{\partial n} > 0 \iff n^2\ln(n) > \frac{\left(\frac{1}{2}n^2\ln(n) - \frac{1}{4}(n^2-1)\right)\left(n^2\ln(n)\ln(n) - \frac{1}{2}n^2\ln(n) + \frac{1}{4}(n^2-1)\right)}{\frac{1}{2}n^2\ln(n)\ln(n) - \frac{1}{2}n^2\ln(n) + \frac{1}{4}(n^2-1)}$$
$$\iff n - \frac{1}{n} > 2\ln(n).$$

Since this inequality holds for all $n \geq 2$, $G(2,n)$ is monotonically increasing for all $n \geq 2$. Moreover, direct calculations confirm that $G(2,11) < -0.005 < 0 < 0.008 < G(2,12)$. Combining the monotonicity with this numerical check, we conclude

$$G(\theta,n) \geq G(2,n) \geq G(2,12) > 0 \quad \forall n \geq 12, \forall \theta \in (0,2],$$

as claimed. □

We now analyze the root of $F(z,\theta,n) = 0$. We aim to show that for $n \geq 12$, this root is decreasing in $\theta \in (1,2]$, which implies it is increasing in $\gamma = \frac{\theta}{\theta-1} \in [2,\infty)$. For any given $n > 1$ (treated as a real number) and $\theta > 1$, let $z(\theta,n)$ be the positive real root of $F(z,\theta,n) = 0$. The following lemma establishes that this root exists and is unique.

Lemma D.2. *For each $n > 1$ and $\theta > 1$, the equation $F(z,\theta,n) = 0$ has a unique positive real solution $z(\theta,n)$ in the open interval $(1,n)$. At this solution, the slope of $z \mapsto F(z,\theta,n)$ is strictly negative: $F_z(z(\theta,n),\theta,n) < 0$.*

Proof. For $z \geq 1$ and $\theta > 1$, the partial derivative with respect to z is $F_z(z,\theta,n) = 1 - \theta z^{\theta-1} \leq 1 - \theta < 0$. Thus, the function $z \mapsto F(z,\theta,n)$ is strictly decreasing on $[1,\infty)$. Evaluating the function at the boundaries of the interval $(1,n)$ gives

$$F(1,\theta,n) = g(\theta,n) - (n-1) > g(1,n) - (n-1) = 0$$

and

$$F(n,\theta,n) = -\frac{\theta-1}{\theta}(n^\theta - 1) < 0.$$

Since $z \mapsto F(z,\theta,n)$ is continuous, the intermediate value theorem then guarantees that a unique root exists in $(1,n)$.

To complete the proof, we show that no root exists in $[0,1]$. Since $F_{zz}(z,\theta,n) < 0$ for all $z > 0$, it is sufficient to show that both $F(0,\theta,n)$ and $F(1,\theta,n)$ are positive. Because $F(0,\theta,n) = g(\theta,n) - n + 1 > 0$ and we have already established that $F(1,\theta,n) > 0$, the function $F(z,\theta,n)$ must be strictly positive over the interval $[0,1]$. Therefore, the equation $F(z,\theta,n) = 0$ has no root in $[0,1]$. □

To characterize how $z(\theta,n)$ varies with θ, apply the implicit function theorem to $F(z(\theta,n),\theta,n) = 0$ to obtain

$$\frac{\partial z(\theta,n)}{\partial \theta} = -\frac{F_\theta(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)}.$$

As shown in Lemma D.2, the denominator is negative: $F_z(z(\theta,n),\theta,n) < 0$. Hence, $\frac{\partial z(\theta,n)}{\partial \theta} < 0$ if and only if $F_\theta(z(\theta,n),\theta,n) < 0$. Differentiating the equation once more yields

$$\begin{aligned}\frac{\partial^2 z(\theta,n)}{\partial \theta^2} = &-\frac{F_{\theta\theta}(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)} - \frac{F_{\theta z}(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)}\frac{\partial z(\theta,n)}{\partial \theta} \\ &+ \frac{F_{zz}(z(\theta,n),\theta,n)\frac{\partial z(\theta,n)}{\partial \theta} + F_{z\theta}(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)F_z(z(\theta,n),\theta,n)}F_\theta(z(\theta,n),\theta,n),\end{aligned}$$

which means

$$\frac{\partial z(\theta,n)}{\partial \theta} = 0 \implies \frac{\partial^2 z(\theta,n)}{\partial \theta^2} = -\frac{F_{\theta\theta}(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)}.$$

Given $F_z(z(\theta,n),\theta,n) < 0$, it follows that if the sign of $\frac{\partial z(\theta,n)}{\partial \theta}$ flips from negative to positive at some θ, then $F_{\theta\theta}(z(\theta,n),\theta,n) \geq 0$ must hold for such θ. By Lemma D.1, such a sign change is not possible for $n \geq 12$. Hence, if the sign of $\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{\partial \theta}$ is negative, the derivative remains negative for all $\theta \in (1,2]$. The following lemma formalizes this argument.

Lemma D.3. *For $n \geq 12$, $z(\theta,n)$ is strictly decreasing with $\theta \in (1,2]$.*

Proof. We aim to show $\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{\partial \theta} < 0$ for all $n \geq 12$. For each $n > 1$, let $z(1,n) := \lim_{\theta\searrow 1} z(\theta,n)$, which must solve

$$\lim_{\theta\searrow 1}\frac{\frac{1}{\theta}(n^\theta - 1) - (n-1)}{z^\theta - z} = 1,$$

and therefore $z(1,n)\ln(z(1,n)) = n\ln(n) - (n-1)$. This means that $z(1,n) > 1$ and $F_\theta(z(1,n),\theta,n)|_{\theta=1} = 0$. It then follows from Lemma D.1 that

$$\lim_{\theta\searrow 1} F_{\theta\theta}(z(\theta,n),\theta,n) = F_{\theta\theta}(z(1,n),1,n) < 0 \quad \forall n \geq 12.$$

To evaluate $\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{\partial \theta}$, we apply L'Hopital's rule because $\lim_{\theta\searrow 1} F_\theta(z(\theta,n),\theta,n) = 0$ and $\lim_{\theta\searrow 1} F_z(z(\theta,n),\theta,n) = 0$:

$$\begin{aligned}\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{\partial \theta} &= -\lim_{\theta\searrow 1}\frac{F_\theta(z(\theta,n),\theta,n)}{F_z(z(\theta,n),\theta,n)} \\ &= -\lim_{\theta\searrow 1}\frac{\frac{d}{d\theta}F_\theta(z(\theta,n),\theta,n)}{\frac{d}{d\theta}F_z(z(\theta,n),\theta,n)} \\ &= -\lim_{\theta\searrow 1}\frac{F_{\theta\theta}(z(\theta,n),\theta,n) + F_{z\theta}(z(\theta,n),\theta,n)\frac{\partial z(\theta,n)}{d\theta}}{F_{z\theta}(z(\theta,n),\theta,n) + F_{zz}(z(\theta,n),\theta,n)\frac{\partial z(\theta,n)}{d\theta}} \\ &= -\frac{F_{\theta\theta}(z(1,n),1,n) + F_{z\theta}(z(1,n),1,n)\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{d\theta}}{F_{z\theta}(z(1,n),1,n) + F_{zz}(z(1,n),1,n)\lim_{\theta\searrow 1}\frac{\partial z(\theta,n)}{d\theta}},\end{aligned}$$

where $F_{zz}(z(1,n),1,n) = 0$ and $F_{z\theta}(z(1,n),1,n) = -(\ln(z(1,n)) + 1)$. Solving the equation

for $\lim_{\theta \searrow 1} \frac{\partial z(\theta,n)}{\partial \theta}$ yields

$$\lim_{\theta \searrow 1} \frac{\partial z(\theta,n)}{\partial \theta} = \frac{1}{2} \frac{F_{\theta\theta}(z(1,n),1,n)}{\ln(z(1,n))+1} < 0 \quad \forall n \geq 12,$$

meaning that $z(\theta,n)$ is strictly decreasing in a neighborhood of $\theta = 1$. We know from Lemma D.1 that as long as $n \geq 12$, the sign of $\frac{\partial z(\theta,n)}{\partial \theta}$ cannot flip from negative to positive as θ increases from 1 to 2. Therefore, we conclude that $\frac{\partial z(\theta,n)}{\partial \theta} < 0$ for all $\theta \in (1,2]$ and $n \geq 12$. □

Let $\theta = \frac{\gamma}{\gamma-1}$. Since this transformation is strictly decreasing in γ (with θ declining from 2 to 1 as γ rises from 2 to ∞), Lemma D.3 immediately implies that as long as $n \geq 12$, the threshold value $z_\gamma(n) = z(\theta,n)|_{\theta=\frac{\gamma}{\gamma-1}}$ strictly increases with $\gamma \geq 2$.

Part 3: *Participation sensitivity to n*

To simplify the exposition, we continue using the transformation $\theta = \frac{\gamma}{\gamma-1}$. Since $\theta \in (1,2]$ for $\gamma \geq 2$, it is sufficient to show that both the threshold $z(\theta,n)$ and the minimal participation rate $\frac{z(\theta,n)}{n}$ are strictly increasing in $n \geq 3$ for all $\theta \in (1,2]$.

Differentiating $F(z(\theta,n),\theta,n) = 0$ with respect to n yields

$$\frac{\partial z(\theta,n)}{\partial n} = -\frac{n^{\theta-1}-1}{F_z(z(\theta,n),\theta,n)} > 0,$$

where we know from Lemma D.2 that the denominator is strictly negative. This implies that $z(\theta,n)$ is strictly increasing in n. Also, since

$$\frac{\partial (z(\theta,n)/n)}{\partial n} = \frac{1}{n^2} \frac{(\theta-1)(n-z(\theta,n)-1)}{\theta(z(\theta,n))^{\theta-1}-1},$$

the ratio $\frac{z(\theta,n)}{n}$ is increasing in n if and only if $n - z(\theta,n) - 1 > 0$. Observe

$$\frac{\partial (n-z(\theta,n)-1)}{\partial n} > 0 \iff \frac{\partial z(\theta,n)}{\partial n} < 1 \iff \frac{z(\theta,n)}{n} > \left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}}.$$

This means that, for a given θ, if both $n - z(\theta,n) - 1 \geq 0$ and $\frac{z(\theta,n)}{n} > \left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}}$ hold at some n, they must also hold (with the first inequality being strict) for all larger n. We verify that at $n = 3$, these inequalities simultaneously hold for any $\theta \in (1,2]$.

For $n = 3$, the two required inequalities, $3 - z(\theta,3) - 1 \geq 0$ and $\frac{z(\theta,3)}{3} > \left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}}$, are equivalent to $3\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}} < z(\theta,3) \leq 2$. Then, since $3\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}} > 3/e > 1$ and $F(z,\theta,3)$ is strictly decreasing in $z > 1$, this condition is equivalent to

$$F(z,\theta,3)\big|_{z=3\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}}} > \underbrace{F(z(\theta,3),\theta,n)}_{=0} \geq F(z,\theta,3)|_{z=2}. \tag{25}$$

Because $\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}} \leq 1/2 < 1$ and $\frac{1}{\theta}3^\theta \geq 3$, the leftmost term in condition (25) is strictly positive: $F(z,\theta,3)\big|_{z=3\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}}} = (1-\left(\frac{1}{\theta}\right)^{\frac{1}{\theta-1}})(\frac{1}{\theta}3^\theta - 3) + 1 - \frac{1}{\theta} > 0$. To verify that the rightmost term is non-positive, observe

$$F(z,\theta,3)|_{z=2} \leq 0 \iff \underbrace{\frac{1}{\theta}(3^\theta - 1)}_{g(\theta,3)} \leq 2^\theta \iff \ln(g(\theta,3)) - \theta\ln(2) \leq 0.$$

We know from inequality (24) that $\ln(g(\theta,3))$ is strictly convex in θ, hence so is $\ln(g(\theta,3)) - \theta\ln(2)$. Thus, $\ln(g(\theta,3)) - \theta\ln(2)$ must attain its maximum at an endpoint:

$$\ln(g(\theta,3)) - \theta\ln(2) \leq \max_{\theta\in\{1,2\}} \{\ln(g(\theta,3)) - \theta\ln(2)\} = 0 \quad \forall\theta \in [1,2],$$

implying $F(z,\theta,3)|_{z=2} \leq 0$ for all $\theta \in (1,2]$. Consequently, condition (25) holds for all $\theta \in (1,2]$, which in turn proves that the minimal membership share $\frac{z(\theta,n)}{n}$ increases with $n \geq 3$.

To see the limit of $z(\theta,n)/n$ for $n \to \infty$, we note that $z(\theta,n)$ satisfies

$$\begin{aligned} F(z(\theta,n),\theta,n) = 0 &\iff \frac{1}{\theta}\left(n^\theta - 1\right) - n - (z(\theta,n))^\theta + z(\theta,n) + 1 = 0 \\ &\iff \left(\frac{z(\theta,n)}{n}\right)^\theta - n^{1-\theta}\frac{z(\theta,n)}{n} = \frac{1}{\theta} + \frac{\theta-1}{\theta}n^{-\theta} - n^{1-\theta} \end{aligned}$$

for all n. Since $z(\theta,n)/n < 1$ for all n, taking the limit $n \to \infty$ and observing that $n^{-\theta}$ and $n^{1-\theta}$ both converge to 0 yields

$$\left(\lim_{n\to\infty}\frac{z(\theta,n)}{n}\right)^\theta = \frac{1}{\theta} \iff \lim_{n\to\infty}\frac{z(\theta,n)}{n} = \left(\frac{1}{\theta}\right)^{\frac{1}{\theta}} = \left(\frac{\gamma-1}{\gamma}\right)^{\frac{\gamma-1}{\gamma}},$$

completing the proof. □

D.5 Proof of Proposition 3.2

Proof. Part a): The (one-shot) internal stability condition requires for all $i \in M$

$$u_i(M) \geq u_i(M\setminus\{i\}) \iff |M| - \ln(|M|) \geq |M| - 1 \iff |M| \leq e \approx 2.72,$$

whereas the (one-shot) external stability condition requires for all $i \notin M$

$$u_i(M) > u_i(M\cup\{i\}) \iff |M| > |M| + 1 - \ln(|M|+1) \iff |M| > e - 1 \approx 1.72.$$

The two conditions jointly imply that a coalition is one-shot stable if and only if $|M| = 2$.
Part b): A coalition M is not Pareto dominated by the grand coalition if and only if

$|M| > n - \ln(n)$. Based on the same argument as in the proof of Proposition 3.1, this necessary condition for Pareto efficiency is also sufficient. We give a rigorous proof for general symmetric free-rider games when we prove Theorem 4.1. □

D.6 Proof of Theorem 4.1

Proof. We prove Theorem 4.1 in six steps—part *a*) is established in step 3 and part *b*) in step 6. Assumption 1 implies that for any $M, M' \in \mathcal{N}$ (regardless of potential overlap),

$$|M'| > |M| \implies u_i(M') > u_j(M) \quad \forall i \in M', \forall j \in M, \tag{26}$$

$$|M'| > |M| \implies u_i(M') \geq u_j(M) \quad \forall i \notin M', \forall j \notin M, \tag{27}$$

$$u_i(M) \leq u_j(M) \quad \forall i \in M, \forall j \notin M \tag{28}$$

We refer to statements (26), (27), (28) as Inequalities 1 (strict), 2 (weak), 3 (weak), respectively.

Step 1: *The grand coalition, N, is Pareto efficient.*
For all $M \in \mathcal{N}$ with $M \neq N$, we have $|M| < |N|$. Thus, by Inequality 1, there exists $i \in N$ such that $u_i(N) > u_i(M)$ for all $M \in \mathcal{N} \setminus \{N\}$.

Step 2: *Apart from the grand coalition, a coalition $M \in \mathcal{N} \setminus \{N\}$ is Pareto efficient if and only if there exists $i \notin M$ such that $u_i(M) > u_i(N)$.*
Necessity ("only if"): In words, a free rider on any Pareto efficient non-grand coalition M has to be better off than the members of the grand coalition. Pareto efficiency of coalition $M \neq N$ is equivalent to the statement that M is not Pareto dominated by any coalition $\hat{M} \in \mathcal{N}$, namely,

$$\exists\, i \in N \text{ such that } u_i(M) > u_i(\hat{M}) \quad \text{or} \quad u_i(M) \geq u_i(\hat{M}) \,\forall\, i \in N \tag{29}$$

for all $\hat{M} \in \mathcal{N}$. In particular, for $\hat{M} = N$, the first condition in equation (29) must be satisfied because the second possibility is ruled out by Inequality 1 (and the fact that $|M| < |N|$). If the first condition is satisfied for some $i \in N$, then by Inequality 3 the same condition has to be satisfied for a free rider $j \notin M$. Thus, we established that the existence of $j \notin M$ such that $u_j(M) > u_j(N)$ is a necessary condition for the Pareto efficiency of any non-grand coalition M.

Sufficiency (" if"): We shall show that

$$\exists\, j \notin M \text{ such that } u_j(M) > u_j(N) \implies \begin{array}{l} \exists\, i \in N \text{ such that } u_i(M) > u_i(\hat{M}) \\ \text{or} \\ u_i(M) \geq u_i(\hat{M})\, \forall\, i \in N \end{array} \quad \forall \hat{M} \in \mathcal{N}. \tag{30}$$

For $\hat{M} = N$, the first condition on the right side immediately follows from the left side. For $\hat{M} = M$, the second condition on the right side is trivially satisfied. For $\hat{M} \notin \{N, M\}$, we distinguish three cases. (i) Let $|\hat{M}| > |M|$. Then there exists $i \in \hat{M} \setminus M$ and the right side of statement (30) follows from Inequality 1 and symmetry because $u_i(M) = u_j(M) > u_j(N) > u_i(\hat{M})$ given $\hat{M} \neq N$. The other two cases hold independently of the premise, i.e., of the left side of statement (30). (ii) Let $|\hat{M}| < |M|$. Then, by Inequality 1, the members of M are strictly better off than those of $\hat{M}$. If the intersection $M \cap \hat{M}$ is nonempty, this directly implies $u_i(M) > u_i(\hat{M})$ for $i \in \hat{M} \cap M$. If $M \cap \hat{M}$ is empty, Inequalities 3 and 1 imply $u_i(M) \geq u_k(M) > u_i(\hat{M})$ for $i \notin M$ (and thus $i \in \hat{M}$) and $k \in M$. (iii) Let $|\hat{M}| = |M|$. Then, by Inequality 3, the free riders are either strictly better off than the members, i.e., $u_i(M) > u_i(\hat{M})$ for $i \in \hat{M} \setminus M$, or free riders and members have equal payoffs so that $u_i(M) = u_i(\hat{M})$ for all $i \in N$.

Step 3: *There exists $m_\star \in \mathbb{N}$ such that M is Pareto efficient if and only if $|M| \geq m_\star$. By Theorem 3.1, this step proves part a) of the theorem.*
By step 2, we know that a non-grand coalition M is Pareto efficient if and only if there exists $i \notin M$ such that $u_i(M) > u_i(N)$. Assume M satisfies this condition and let $\hat{M} \in \mathcal{N}, \hat{M} \neq N$ with $|\hat{M}| \geq |M|$. Then $u_j(\hat{M}) \geq u_i(M) > u_i(N) = u_j(N)$ for all $j \notin \hat{M}$ and $i \notin M$, i.e., the condition is also satisfied for all coalitions of same or larger size. The first (weak) inequality follows from Inequality 2 if $|\hat{M}| > |M|$ or payoff symmetry if $|\hat{M}| = |M|$, the second (strict) inequality holds by assumption about the nature of M, and the final equality by symmetry of the payoff functions (all players are members of the grand coalition). Let $\mathcal{A} := \{M \mid \exists i \notin M \text{ such that } u_i(M) > u_i(N)\}$ and set $m_\star := \min\{|M| \mid M \in \mathcal{A}\}$ if $\mathcal{A}$ is nonempty and $m_\star := n$ otherwise. Steps 1 and 2 have shown that the union $\mathcal{A} \cup \{N\}$ characterizes the set of all Pareto efficient coalitions. We have now shown that a coalition is in this set if and only if it has at least $m_\star$ members.

Step 4: *If $|M| = m_\star$, there is no other Pareto efficient coalition M' such that*

$$u_i(M') \geq u_i(M) \quad \forall i \in E(M, M'),$$

where at least one of the inequalities is strict, meaning that M is not dominated by M'.
If $m_\star = n$, the statement follows from step 3 because then the grand coalition is the only

Pareto efficient coalition. Let $m_\star < n$ and M' be a Pareto efficient coalition. We show that

$$\exists i \in E(M, M') \text{ such that } u_i(M') < u_i(M) \tag{31}$$

or

$$u_i(M') \leq u_i(M) \quad \forall i \in E(M, M'), \tag{32}$$

where $E(M, M') = (M' \setminus M) \cup (M \setminus M')$. If $M' = N$, condition (32) holds by definition of $m_\star$ and payoff symmetry. Condition (32) also holds if $M' \sim M$. Let $M' \neq N$ with $M' \not\sim M$. We know from part *a*) of the theorem, i.e., step 3 of the proof, that $n > |M'| \geq m_\star$. We distinguish two cases. (i) If $|M'| = m_\star$, both $M' \setminus M$ and $M' \setminus M$ are nonempty, and by Inequality 3

$$u_i(M') \leq u_i(M) \quad \forall i \in M' \setminus M$$

and

$$u_i(M') \geq u_i(M) \quad \forall i \in M \setminus M'.$$

Since players are symmetric, either both inequalities hold with equality or both of them hold with strict inequality. Since M' is distinct from M, the first case cannot be true. Thus, both of these must hold with strict inequality, which implies condition (31). (ii) If $n > |M'| > m_\star$, then $M' \setminus M$ is nonempty (because $|M| = m_\star$) and

$$u_i(M) \geq u_i(N) > u_i(M') \quad \forall i \in M' \setminus M,$$

where the first inequality follows from step 2 of the proof and the second follows from Inequality 1 (note that $M' \neq N$). Therefore, condition (31) is satisfied.

Step 5: *If $|M| > m_\star$, there is a coalition $M' \subset M$ with $|M'| = m_\star$ such that*

$$u_i(M') \geq u_i(M) \quad \forall i \in M \setminus M',$$

where at least one of the inequalities is strict, meaning that M is dominated by M'.
If $M = N$, take any M' with $|M'| = m_\star$ and the statement follows directly from steps 2 and 3 of the proof; since M' is a Pareto efficient non-grand coalition (step 3), free riders on M' must be better off than the members of the grand coalition (step 2). Let $M \neq N$. Then $n > |M| > m_\star$. We can find a coalition $M' \subset M$ with $|M'| = m_\star$ and $M \setminus M'$ being nonempty. Since M' is Pareto efficient (step 3), we have

$$u_i(M') \geq u_i(N) > u_i(M) \quad \forall i \in M \setminus M',$$

where the first inequality follows from step 2 of the proof and the second from Inequality 1.

Step 6: *A set $\mathscr{A}$ of self-enforcing stable sets is second-order stable if and only if*

$$\mathscr{A} = \left\{ \tilde{M} \subset \mathcal{N} \mid |M| = m_\star \right\},$$

which proves part b) of the theorem.
Let $\mathscr{A}$ be a second-order stable collection of self-enforcing stable sets. By step 4 of the proof, the set $\mathscr{A}$ must include every $\tilde{M}$ such that $|M| = m_\star$. Then, by step 5 of the proof, every other self-enforcing stable set is $\mathscr{A}$-dominated and, therefore, cannot be included in $\mathscr{A}$. □

D.7 Proof of Theorem 5.1

Proof. The proofs of Lemma 3.1 and Theorem 3.1 carry through step by step merely replacing coalitions M by coalition structures $\boldsymbol{M}$ (and similarly the primed variations and those carrying a tilde), replacing the sets $\mathcal{M}$ by the sets $\boldsymbol{\mathcal{M}}$, and replacing the set of all coalitions $\mathcal{N}$ by the set of all coalition structures $\boldsymbol{\mathcal{N}}$. These establish as an intermediate step (Lemma 3.1) that a self-enforcing stable set of coalition structures cannot contain distinct coalition structures and subsequently the Theorem's statement characterizing the set of self-enforcing equilibria. □

D.8 Proof of Proposition 5.1

We prove the following lemma, from which the statement of Proposition 5.1 directly follows.

Lemma D.4. *In Examples 1* and 2*, a coalition structure $\boldsymbol{M}$ is one-shot stable if and only if a) any coalition in $\boldsymbol{M}$ has at most m_* members, where $m_* = 3$ for Example 1* with $\gamma = 2$, and $m_* = 2$ otherwise; b) whenever $\boldsymbol{M}$ contains a singleton coalition, all other coalitions have exactly m_* members; and c) for Example 1* with $\gamma = 2$, no coalition in $\boldsymbol{M}$ is singleton.*

Proof. Necessity ("only if"): We first note that m_* is the integer characterizing the size of one-shot stable coalitions in the single-coalition setting: a coalition is robust against insiders' unilateral defections if and only if it has at most m_* members, and it is robust against outsiders' unilateral entries if and only if it has at least m_* members. See the proof of Proposition 3.2 for Example 2. For Example 1, see the appendix of Karp and Sakamoto (2021).

Now, let $\boldsymbol{M}$ be a one-shot stable coalition structure in the multi-coalition setting. Since players are allowed to leave existing coalitions and unilaterally form singleton coalitions, every coalition in $\boldsymbol{M}$ must be immune to insiders' unilateral defections, just as in the single-coalition setting. This deviation test is unaffected by the presence of other coexisting coalitions. As shown in the reduced-form payoff functions (14) and (15), the change in a

player's payoff from switching coalitions depends only on the sizes of the origin and destination coalitions.[34] This immediately implies condition a) of the lemma: any coalition in $\boldsymbol{M}$ must have at most m_* members. Also, when there is a singleton coalition in $\boldsymbol{M}$, the player in that singleton coalition must not be better off joining any other coalitions. It follows that all other coalitions must be immune to unilateral entry by a free rider as well, requiring that those coalitions must have at least m_* members. Combined with condition a), this proves condition b) of the lemma: whenever $\boldsymbol{M}$ contains a singleton coalition, all other coalitions in $\boldsymbol{M}$ must have exactly m_* members. An immediate implication is that $\boldsymbol{M}$ can contain at most one singleton coalition. To see that condition c) is implied for Example 1* with $\gamma = 2$, suppose to the contrary that $\boldsymbol{M}$ contains a singleton coalition. Then, by condition b), all other coalitions in $\boldsymbol{M}$ have size $m_* = 3$. But members of these three-player coalitions have an incentive to leave and merge with the singleton coalition, forming a coalition of size two; the payoff gain from such a switch is strictly positive: $\xi^2\left(\frac{1}{2}2^2 - 1^2 - \frac{1}{2}3^2 + 2^2\right) = \frac{\xi^2}{2} > 0$. This contradicts the assumed stability of $\boldsymbol{M}$.

Sufficiency ("if"): Let $\boldsymbol{M}$ be a coalition structure that satisfies conditions a), b), and c) of the lemma. Then, as discussed above, every coalition in $\boldsymbol{M}$ is immune to both insiders' unilateral defections to singleton status and outsiders' unilateral entries from singleton status. It remains to show that members of any non-singleton coalition in $\boldsymbol{M}$ have no incentive to switch to other existing coalition in $\boldsymbol{M}$. The proof makes use of two key properties. First, as an immediate consequence of the three conditions, the size difference between any two coexisting coalitions in $\boldsymbol{M}$ is at most one.[35] Second, our tie-breaking rule (Definition 5.3) states that a player will defect to a larger coalition if it provides a weakly higher payoff, but will only defect to a same-sized or smaller coalition if the payoff gain is strictly positive.

Consider player i in coalition $M_l \in \boldsymbol{M}$ with $|M_l| \geq 2$, who contemplates switching to another coalition $M_k \in \boldsymbol{M}$. This deviation transforms M_l and M_k into $M_l \setminus \{i\}$ and $M_k \cup \{i\}$, respectively, while leaving all other coalitions unchanged. In Example 1*, the player is weakly better off from this switch if and only if

$$\xi^{\frac{\gamma}{\gamma-1}}\left(\sum_{M\in\boldsymbol{M}\setminus\{M_l,M_k\}}|M|^{\frac{\gamma}{\gamma-1}} + (|M_l|-1)^{\frac{\gamma}{\gamma-1}} + (|M_k|+1)^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma}(|M_k|+1)^{\frac{\gamma}{\gamma-1}}\right) - \xi\sum_{j\in N}\bar{g}_j$$
$$\geq \xi^{\frac{\gamma}{\gamma-1}}\left(\sum_{M\in\boldsymbol{M}}|M|^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma}|M_l|^{\frac{\gamma}{\gamma-1}}\right) - \xi\sum_{j\in N}\bar{g}_j.$$

[34] More precisely, the payoff difference also depends on whether the move creates a new coalition (when the destination coalition was originally empty) or dissolves an existing one (when the origin coalition had only one member), but these considerations are independent of the rest of the coalition structure and apply equally to the deviation test in the single-coalition setting.

[35] If a singleton exists, condition b) requires all other coalitions to have exactly $m^* \in \{2, 3\}$ members; when $m^* = 2$, the gap is 1, whereas when $m^* = 3$, condition c) rules out the singleton. If no singleton exists, all coalitions have size at least 2 and at most $m^* \leq 3$. Consequently, any two coexisting coalitions differ in size by at most one.

After canceling terms, this inequality simplifies to:

$$h(|M_k|+1) \geq h(|M_l|), \quad \text{where} \quad h(z) := \frac{\gamma-1}{\gamma} z^{\frac{\gamma}{\gamma-1}} - (z-1)^{\frac{\gamma}{\gamma-1}}.$$

A strict payoff improvement requires $h(|M_k|+1) > h(|M_l|)$. This means that the player has an incentive to switch if either i) $h(|M_k|+1) \geq h(|M_l|)$ and $|M_k \cup \{i\}| > |M_l|$ (i.e., the player weakly prefers to form a strictly larger coalition), or ii) $h(|M_k|+1) > h(|M_l|)$ and $|M_k \cup \{i\}| \leq |M_l|$ (i.e., the player strictly prefers to form a weakly smaller coalition). Note that function $h(z)$ is strictly decreasing in z for any $z > \frac{1}{1-\left(\frac{\gamma-1}{\gamma}\right)^{\gamma-1}} \in (1/e, 2]$. Because $|M_l| \geq 2$, it follows from this monotonicity of $h(z)$ that case i) never holds. Case ii) requires $|M_k|+1 < |M_l|$, but this is impossible because the size difference among coexisting coalitions in $\boldsymbol{M}$ cannot exceed one.

In Example 2*, since the number of coexisting coalitions $|\boldsymbol{M}|$ is unchanged, switching from coalition M_l to M_k makes player i weakly better off if and only if

$$-\ln(\xi) - \ln(|M_k|+1) - |\boldsymbol{M}| \geq -\ln(\xi) - \ln(|M_l|) - |\boldsymbol{M}|$$

or $|M_k|+1 \leq |M_l|$. The player is strictly better off if and only if $|M_k|+1 < |M_l|$. If the player ends up in a strictly larger coalition ($|M_k \cup \{i\}| > |M_l|$), he or she is strictly worse off. On the other hand, a strictly profitable deviation to a same-sized or smaller coalition requires $|M_k|+1 < |M_l|$, which is impossible because $|M_k|$ and $|M_l|$ can differ by at most one. Therefore, starting from $\boldsymbol{M}$, no player has an incentive to switch from a non-singleton coalition to another coalition and this completes the proof. □

D.9 Proof of Proposition 5.2

Proof. Part a): If one coalition structure Pareto dominates another, it must yield a strictly higher aggregate payoff. In both examples, the grand coalition structure produces the highest aggregate payoff. Therefore, the grand coalition structure is not Pareto dominated by any alternative coalition structure.

Part b): Suppose that a coalition structure $\boldsymbol{M} \neq \boldsymbol{N}$ is part of a self-enforcing stable set. By Theorem 5.1, $\boldsymbol{M}$ must be Pareto efficient. In particular, $\boldsymbol{M}$ is not Pareto dominated by the grand coalition, and thus

$$\exists i \in N \text{ such that } u_i(\boldsymbol{N}) < u_i(\boldsymbol{M}). \tag{33}$$

For Example 2*, it follows that $-\ln(n) - 1 < -\ln(m_1) - |\boldsymbol{M}|$, where $m_1 := \min_{M \in \boldsymbol{M}} |M|$ is the size of the smallest coalition in $\boldsymbol{M}$. Since $m_1 \geq 1$, it must hold that $-\ln(n) - 1 < -|\boldsymbol{M}|$ and, thus, $|\boldsymbol{M}| < \ln(n) + 1$ as claimed. For Example 1*, we write $\boldsymbol{M} = \{M_1, M_2, \ldots, M_{|\boldsymbol{M}|}\}$ and assume, without loss of generality, that $|M_1| \leq |M_2| \leq \ldots \leq |M_{|\boldsymbol{M}|}|$. We start with two

observations. First, players in the smallest coalition, M_1, receive the largest payoff because the payoff function is

$$u_i(\boldsymbol{M}) = \xi^{\frac{\gamma}{\gamma-1}} \left(\sum_{M \in \boldsymbol{M}} |M|^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma} |M_l|^{\frac{\gamma}{\gamma-1}} \right) - \xi \sum_{j \in N} \bar{g}_j \quad \forall i \in M_l, \tag{34}$$

where only the second term in the parentheses varies between coalitions. Second, reallocating a member of a smaller coalition to a larger one (while keeping the number of coexisting coalitions) always improves the payoff of the players in the smallest coalition, because

$$1 < |M_l| < |M_{|\boldsymbol{M}|}| \implies |M_l|^{\frac{\gamma}{\gamma-1}} + |M_{|\boldsymbol{M}|}|^{\frac{\gamma}{\gamma-1}} < (|M_l| - 1)^{\frac{\gamma}{\gamma-1}} + (|M_{|\boldsymbol{M}|}| + 1)^{\frac{\gamma}{\gamma-1}}.$$

Such a reallocation increases the first term in the parentheses in (34) and, if $l = 1$, also decreases the second term. These observations imply that the payoff of the smallest coalition is maximized when players are concentrated in the largest coalition as much as possible, i.e., when $1 = |M_1| = |M_2| = \cdots = |M_{|\boldsymbol{M}|-1}| < |M_{|\boldsymbol{M}|}| = n - |\boldsymbol{M}| + 1$. This yields the highest possible payoff of players under the requirement that $|\boldsymbol{M}|$ coalitions coexist. Hence,

$$u_i(\boldsymbol{M}) \leq \xi^{\frac{\gamma}{\gamma-1}} \left((n - |\boldsymbol{M}| + 1)^{\frac{\gamma}{\gamma-1}} + |\boldsymbol{M}| - 1 - \frac{1}{\gamma} \right) - \xi \sum_{j \in N} \bar{g}_j \quad \forall i \in N,$$

where the right side is the payoff of members of M_1 when every coalition except for $M_{|\boldsymbol{M}|}$ is singleton. It then follows from equation (33) that

$$\begin{aligned} n^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma} n^{\frac{\gamma}{\gamma-1}} &= \xi^{-\frac{\gamma}{\gamma-1}} \left(u_i(\boldsymbol{N}) + \xi \sum_{j \in N} \bar{g}_j \right) \\ &< \xi^{-\frac{\gamma}{\gamma-1}} \left(u_i(\boldsymbol{M}) + \xi \sum_{j \in N} \bar{g}_j \right) \\ &\leq (n - |\boldsymbol{M}| + 1)^{\frac{\gamma}{\gamma-1}} + |\boldsymbol{M}| - 1 - \frac{1}{\gamma}, \end{aligned}$$

from which we conclude

$$\frac{\gamma - 1}{\gamma} n^{\frac{\gamma}{\gamma-1}} - n < (n - |\boldsymbol{M}| + 1)^{\frac{\gamma}{\gamma-1}} - (n - |\boldsymbol{M}| + 1) - \frac{1}{\gamma}. \tag{35}$$

Note that the right side of this inequality is equated with the left side when $(n - |\boldsymbol{M}| + 1)$ are all replaced by $z_\gamma(n)$, the unique root of equation (10). Because the right side of inequality (35) is increasing in $(n - |\boldsymbol{M}| + 1)$, this implies $n - |\boldsymbol{M}| + 1 > z_\gamma(n)$ for any self-enforcing stable coalition structure $\boldsymbol{M}$.

Part c): Let $\boldsymbol{M}$ be in a self-enforcing stable set with $M, M'' \in \boldsymbol{M}$ such that $M' \neq M$ and $|M'| \geq |M|$. We now consider an alternative coalition structure $\hat{\boldsymbol{M}}$ which coincides with $\boldsymbol{M}$

except that coalitions M and M' are merged into a single coalition. Such a merger makes all players not part of M or M' better off, i.e., $u_j(\hat{\boldsymbol{M}}) > u_j(\boldsymbol{M})$ for all $j \notin M \cup M'$; see equations (14) and (15). Then, since $\boldsymbol{M}$ is Pareto efficient by Theorem 5.1, there must exist $i \in M \cup M'$ such that $u_i(\hat{\boldsymbol{M}}) < u_i(\boldsymbol{M})$. Because players' payoffs are larger in a smaller coalition (and symmetric), this implies that

$$u_i(\hat{\boldsymbol{M}}) < u_i(\boldsymbol{M}) \quad \forall i \in M. \tag{36}$$

For Example 2*, this condition translates into $-\ln(|M| + |M'|) - |\hat{\boldsymbol{M}}| < -\ln(|M|) - |\boldsymbol{M}|$. Since $|\hat{\boldsymbol{M}}| = |\boldsymbol{M}| - 1$, this inequality implies $|M'| > (e-1)|M|$ as claimed. For Example 1*, it follows from (36) that

$$(|M| + |M'|)^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma}(|M| + |M'|)^{\frac{\gamma}{\gamma-1}} < |M|^{\frac{\gamma}{\gamma-1}} + |M'|^{\frac{\gamma}{\gamma-1}} - \frac{1}{\gamma}|M|^{\frac{\gamma}{\gamma-1}},$$

or equivalently,

$$\frac{\gamma}{\gamma-1} > \left(\frac{|M|}{|M'|} + 1\right)^{\frac{\gamma}{\gamma-1}} - \left(\frac{|M|}{|M'|}\right)^{\frac{\gamma}{\gamma-1}}.$$

The right side of this inequality is increasing in $|M|/|M'|$, mapping onto $(1, 2^{\frac{\gamma}{\gamma-1}} - 1]$ for $|M|/|M'| \in (0, 1]$. Because $\frac{\gamma}{\gamma-1} \in (1, 2]$, there exists a unique root $\phi_\gamma \in (0, 1)$ to the equation

$$\frac{\gamma}{\gamma-1} = (\phi + 1)^{\frac{\gamma}{\gamma-1}} - \phi^{\frac{\gamma}{\gamma-1}}.$$

and it must hold $|M|/|M'| < \phi_\gamma$. □