Sober optimism and the formation of international environmental agreements*

Larry Karp† Hiroaki Sakamoto‡

September 19, 2018

Abstract

We analyze a dynamic model of international environmental agreements (IEAs) where countries cannot make long-term commitments or use sanctions or rewards to induce cooperation. Countries can communicate with each other to build endogenous beliefs about the random consequences of (re)opening negotiation. If countries are patient, an effective agreement can be reached after a succession of short-lived ineffective agreements. This eventual success requires "sober optimism": the understanding that cooperation is possible but not easy to achieve. Beliefs are important and negotiations matter. Our results help explain heterogeneous outcomes and provide a counterweight to prevailing pessimistic views about the prospects for IEAs.

Keywords: Environmental agreements; Climate change; Dynamic game
JEL classification: C72; C73; D62; H41; Q54

*We distributed an earlier version of this paper under the title 'International environmental agreements without commitment'. We received valuable comments from participants at SURED 2018 and the Society of Economic Dynamics 2018 meeting, and from: Eugen Kovac, Robert Schmidt, and Seung Han Yoo. The usual disclaimer applies.

†Department of Agricultural and Resource Economics, University of California, Berkeley, United States. (karp@berkeley.edu)

‡Faculty of Law, Politics, and Economics, Chiba University. 1-33 Yayoicho Inage-ku, Chiba, Japan. (hsakamoto@chiba-u.jp)

1 Introduction

A negotiated solution to a global collective action problem, such as protection of the earth's climate, may depend on the negotiating parties' belief about the probability of success. If parties enter negotiations virtually certain that they will succeed, or that they will fail, they are unlikely to make the compromises necessary to achieve success. Their chance of success may be higher if they begin with "sober optimism", recognizing that the process will be difficult, the outcome uncertain, and that a successful agreement might result only after a sequence of failures. We examine the importance of beliefs in the formation of International Environmental Agreements (IEAs). Many different types of beliefs on the spectrum between extreme optimism and extreme pessimism are consistent with market fundamentals and the rules of negotiation. The actual beliefs arise from the political environment and from pre-negotiation conversations.

A two-stage participation game, with industrial organization antecedents, forms the basis for much of the theory of IEAs, and also for our model (d'Aspremont et al. (1983) and Palfrey and Rosenthal (1984)). In the first stage, parties to the negotiation make a binary decision, choosing whether to join the agreement or remain as outsiders.[1] In the next stage, those who joined the agreement choose an action, such as the reduction of greenhouse gas emissions, to maximize members' joint welfare. The free-riding non-members benefit from the members' provision of the public good. Countries' sovereignty, the lack of a supra-national enforcement agency, and the difficulty of making commitments about future behavior, all justify the use of this non-cooperative setting to study IEAs.[2]

Early applications of this game to the IEA setting, relying on parametric examples, conclude that large and effective IEAs do not emerge in equilibrium, especially when the potential gains from cooperation are large (Hoel, 1992; Carraro and Siniscalco, 1993; Barrett, 1994). These papers explain the actual difficulty of building a successful IEA. However, countries sometimes manage to form coalitions. Mitchell (2018) lists 1270 multilateral environmental agreements, including 512 amendments and 224 protocols, for the period from 1800 to 2018. Some agreements attract many members and have been important in mitigating

[1] We assume throughout that countries use pure strategies. Dixit and Olson (2000) and Hong and Karp (2012, 2014) study mixed strategy equilibria.

[2] A distinct strand of literature studies IEAs using concepts of cooperative game theory such as core (Chander and Tulkens, 1995, 1997; Germain et al., 2003) or farsightedness (Ray and Vohra, 2001; Osmani and Tol, 2009; Diamantoudi and Sartzetakis, 2015, 2018). Finus (2001), Wagner (2001), Barrett (2005), and de Zeeuw (2015) survey the literature.

trans-boundary pollution (Young, 2011). Member countries usually comply even if the IEA has no explicit sanctioning mechanism and despite international law's limited authority (Breitmeier et al., 2006). Kolstad and Toman (2005) describe the discrepancy between the pessimistic theory and the limited but real successes as the "paradox of international agreements".

Merely relaxing the parametric assumptions of the earlier models might reverse their pessimistic conclusions (Karp and Simon, 2013). However, those conclusions continue to hold sway in the profession. For example, reasoning from one of the simplest models, Nordhaus (2015) concludes that trade sanctions might be needed to enable nations to solve the problem of climate change. Earlier papers that study the role of trade sanctions, social norms, monetary transfers, or replacing convex technology with increasing-returns-to-scale include Barrett (1997, 2001, 2006); Hoel and Schneider (1997); and Carraro et al. (2006).

Several papers imbed the two-stage participation game into a repeated game. Consistent with the Folk Theorem, countries may be willing to remain in a large IEA if they are patient and believe that their defection triggers a low-membership equilibrium (Barrett, 2003). These large agreements are self-enforcing, and require no explicit commitment, but the deviation strategies that support them may be implausible. Battaglini and Harstad (2016) study a repeated game in which signatories can commit to the number of periods during which an IEA is binding. This commitment ability enables countries to solve an investment-holdup problem, potentially resulting in a large and long-lived IEA. Kovac and Schmidt (2017) demonstrate that even in the absence of commitment or the holdup problem, large IEAs are possible when deviation triggers a costly delay of reaching a long-term agreement. (Section 5 discusses these issues.)

Our dynamic model requires neither implausible out-of-equilibrium behavior nor long-term commitment (e.g. about the length of the agreement). There are no side-payments or (e.g trade) sanctions, and the abatement technology is standard. Like most of the literature, we use a symmetric-agent game. Even if in equilibrium there is a unique number of coalition members in the one-shot game, the equilibrium does not pin down the members' identity. This apparently vacuous multiplicity generates an incentive for countries to continue to work towards a large and successful agreement.

Reflecting real-world limitations in commitment ability, and historical examples (e.g. Canada's abrogation of the Kyoto Protocol), we recognize that signatories can review and reject any previously-signed agreement. Countries

adhere to an agreement only when it serves their national self-interest. For this reason, all agreements are "interim". Abandoning any interim agreement triggers a new round of negotiation, resulting in a new interim stable (non-cooperative Nash) agreement. Stable interim agreements are either "failures" or "successes". The failures have low membership and produce low welfare gains, just as in the standard one-shot models. The successes, in contrast, have (relatively) high membership and produce high welfare. Members of a failed agreement disband it at the earliest opportunity. By triggering a new round of negotiation, they might be free-riders in a future agreement, either a failed or a successful one; at worst, they become members of a subsequent short-lived failed agreement.

Although both the failed and the successful agreements are non-cooperative Nash equilibria to a participation game, and thus stable, only the successful agreements are sufficiently attractive to maintain members' *permanent* adherence. We denote these equilibria as "sustainable" (not merely stable).

To understand why the existence of such equilibria requires sober optimism, consider a subgame that begins with an interim sustainable agreement. Members of that agreement recognize that if they abandon it, thereby triggering a new round of negotiation, they might become free-riders in a subsequent agreement. If they are extremely optimistic about the near-term emergence of another successful agreement, the incentive to deviate from the existing agreement is high, making the original agreement non-sustainable. Thus, the existence of such agreements requires that countries are not "too optimistic" about the chance of successful negotiations. Now consider an out-of-equilibrium subgame that begins with an interim agreement that is neither a "failure" nor a "success", but something in between. Members are willing to abandon this agreement only if they are sufficiently optimistic about reaching a successful agreement in the near-term. Therefore, the possibility of reaching a successful agreement requires that countries are sufficiently optimistic about its prospects. In short, these successful (= sustainable) equilibria require sober optimism.

Our model has the flavor of real-world negotiations: they might be painstakingly long and their outcome uncertain (Benedick, 1998; Oberthur and Ott, 1999).[3] Negotiations might not be successful, but *ex ante* they are not a waste of time. The meta equilibrium in this game includes beliefs, summarized by an endogenous probability distribution over the size of the next-period IEA and

[3]Benedick (1998) documents that during the negotiation process that eventually resulted in the Montreal Protocol, events took a variety of surprising turns and some of the important agreements were shaped by chance.

the identity of its members. The negotiation process constrains but does not uniquely determine these beliefs.

Our results provide a counterweight to the literature suggesting that IEAs require special circumstances to succeed, and otherwise are doomed to be small and ineffective. This pessimistic view can be self-fulfilling, because beliefs affect outcomes. Beliefs can be influenced by the political environment and pre-game conversations among negotiators. By recognizing the stochastic relation between negotiations' fundamentals and their outcomes, our paper can explain observed heterogeneity. More importantly, it might shift the narrative about the prospects for successful IEAs, thereby improving those prospects.

Our major results use a reduced form model for the stage game payoffs. Under assumptions previously used in climate economics, we show that this repeated game is isomorphic to a dynamic model that incorporates stock pollutants such as CO2. The results therefore apply to climate negotiations.

2 The model

We specify the payoff, review the one-period game, and then describe the dynamic game. As in most of the literature, players can form a single coalition at a time. The model is described by a list $\langle \delta, N, (u_i)_{i\in N}\rangle$ where $\delta \in (0,1)$ is the discount factor, $N := \{1, 2, \ldots, n\}$ is the set of all players with cardinality $n \geq 4$, and $u_i : \mathcal{N} \to \mathbb{R}$ is the single-period reduced-form period payoff function of player i, where $\mathcal{N}$ is the set of all subsets of N. In every period, players decide whether to join a coalition. Their decisions in period t result in a coalition $M_t \in \mathcal{N}$. Player i's discounted present-value payoff from period t onward is

$$\sum_{s=t}^{\infty} \delta^{s-t} u_i(M_s).$$

Section 4 establishes an isomorphism between this model and one with stock pollutants, making the results relevant for climate economics. The reduced form payoff in a period depends only the coalition in that period. Two examples illustrate this dependence.

Example 1. Player $i's$ payoff function is

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

with $\gamma > 1$, $c > 0$, and g_i player i's pollution-generating input. The first term equals the net private benefit from consuming g_i and the second term represents the damage from aggregate pollution. Without pollution damage, player i would choose $g_i = \bar{g}_i > 0$. Members of a coalition jointly maximize their aggregate payoff; non-members individually maximize their own welfare. With either simultaneous or sequential moves the reduced form payoff function is

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

for each $M \in \mathcal{N}$. The reduced form payoff functions are symmetric across players even though the original payoff functions are not.

Example 2. A more familiar model uses the payoff function

$$g_i - c \sum_{j \in N} g_j, \tag{2}$$

with $1/n < c < 1$. Again, $g_i \in [0, 1]$, is player i's pollution level. For each $M \in \mathcal{N}$, the reduced-form payoff function is

$$u_i(M) = \begin{cases} -c(n - |M|) & \forall i \in M \text{ if } |M| \geq 1/c \\ 1 - c(n - |M|) & \forall i \notin M \text{ if } |M| \geq 1/c \\ 1 - cn & \forall i \in N \text{ if } |M| < 1/c. \end{cases} \tag{3}$$

.

2.1 Static one-shot game

The one-shot game is a building block for the dynamic game.[4] We adopt the tie-breaking assumption that players join a coalition whenever they are indifferent between joining and not joining. A Nash equilibrium coalition $M \in \mathcal{N}$ in this

[4]The non-cooperative simultaneous move game emphasizes the final stage of a round of negotiation where each player makes a participation decision with the understanding that the other players will not react to her decision. This model is more plausible in a dynamic setting where the end of negotiation period is explicit and agents have opportunities to respond in the future to a current defection.

participation game satisfies[5]

$$i \in M \quad \text{if and only if} \quad u_i\left(M \cup \{i\}\right) \geq u_i\left(M \setminus \{i\}\right). \tag{6}$$

Following the literature, we say that a coalition is *stable* if and only if it satisfies (6). The 'only if' part in (6) implies that M is *internally stable* (no member wants to leave), and the 'if' part implies that it is *externally stable* (non-members do not want to join). We use m_* to denote the number of countries in a stable coalition to the one-shot game. For the two Examples above, m_* is unique.

Remark 1. For Example 1, there exists a unique equilibrium size $m_* \geq 2$; m_* is independent of c and is weakly decreasing in γ with $\lim_{\gamma \to 1} m_* = n$ and $\lim_{\gamma \to \infty} m_* = 2$. Furthermore, $m_* = 3$ for $\gamma = 2$ and $m_* = 2$ for all $\gamma > 2$.

Remark 2. For Example 2, there exists a unique equilibrium size $m_* \geq 2$, the solution to

$$m_* = \lceil 1/c \rceil,$$

where $\lceil 1/c \rceil$ (the ceiling function) is the smallest integer weakly greater than $1/c$.

For Example 2, larger marginal damages (higher c) lower the equilibrium coalition size and increase the benefit of cooperation. This relation is sometimes taken to imply that equilibrium cooperation is low precisely when it is most valuable. However, that conclusion depends on parametric assumptions. By Remark 1, the equilibrium coalition size in Example 1 falls with γ, but the relation between the benefit of cooperation and γ is non-monotonic. Here, the relation between the benefit of cooperation and the equilibrium level of cooperation is non-monotonic. Assertions that equilibrium cooperation is low when cooperation is most valuable are, in general, unwarranted.

In the symmetric setting Condition (6) does not pin down the identity of coalition members. Denote $\mathcal{M} \subset \mathcal{N}$ as the set of equilibrium outcomes:

$$\mathcal{M} := \{M \in \mathcal{N} \mid M \text{ satisfies (6)}\}. \tag{7}$$

[5]The 'only if' part is equivalent to

$$u_i\left(M\right) \geq u_i\left(M \setminus \{i\}\right) \quad \forall i \in M \tag{4}$$

and the 'if' part is equivalent to

$$u_i\left(M \cup i\right) < u_i\left(M\right) \quad \forall i \notin M. \tag{5}$$

In the examples above, $\mathcal{M}$ contains $C_{m_*}^n := \binom{n}{m_*}$ different stable coalitions, each with m_* members. This indeterminacy is innocuous in the one-shot model, but it is important in the dynamic setting; there we need to describe players' beliefs about the negotiation outcome.[6]

Provided that $\mathcal{M}$ is not a singleton, the outcome of the negotiation is uncertain prior to the negotiation process. By assumption, players know that *some* stable coalition in $\mathcal{M}$ will emerge, but they are not sure which one. We describe players' beliefs using the probability distribution $\pi = (\pi_M)_{M\in\mathcal{M}}$, where $\pi_M \in [0,1]$ equals the probability that M is the outcome of the stage game. The distribution π might be purely subjective, reflecting a common belief about the equilibrium outcome. Alternatively, we can view π as a randomization device that players collectively agree to use to promote coordination.

We refer to π a common belief without specifying its micro-foundations. Players who share a common belief π evaluate their ex-ante payoff as

$$\mathbb{E}_\pi\left[u_i(\tilde{M})\right] := \sum_{M\in\mathcal{M}} u_i(M)\pi_M,$$

where $\mathcal{M} \subset \mathcal{N}$ is defined by (7).

2.2 Dynamic setting

The dynamic game contains many periods, each of which has two stages. We assume that countries cannot commit to a coalition for more than a single period.[7] The state variable at the beginning of a period is the coalition inherited from the previous period, M_{-1}; the initial condition is $M_{-0} = \varnothing$, the null coalition. In the first stage of a period players decide whether to reopen the negotiation process. If every player chooses to stay with the existing coalition, they receive the payoffs associated with that coalition for a period and then move to the next period. If any player deviates from the existing coalition in the first stage, that coalition dissolves and players move to the second stage where a stable coalition

[6]The assumption of symmetric agents makes obvious the indeterminacy of the identity of members, and the resulting multiplicity of equilibria. Parties to actual international negotiations are not, of course, symmetric. However, that asymmetry does not, in general, pin down the identity of members. Our qualitative results would also hold with asymmetric agents, provided that the asymmetry is not 'too extreme', although the precise formulae would change.

[7]Introducing commitment ability and allowing members of a coalition to endogenously choose the duration of the agreement as in Battaglini and Harstad (2016), does not change our results. For a small coalition, the duration is always set to the shortest possible length (i.e., only one period). For a sufficiently large coalition, members make it as long-term as possible.

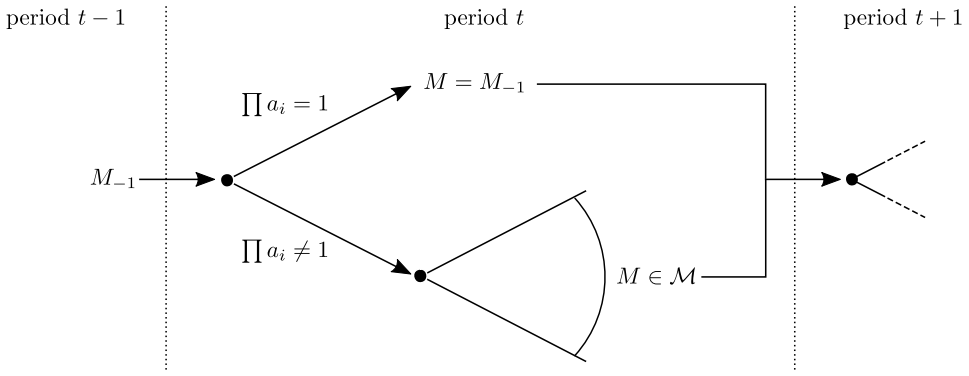

Figure 1: The timing of the game.

is randomly selected using the probability distribution π.[8] Players receive the payoff associated with that coalition for a period, and then move to the next period. The abandonment of a previously negotiated agreement in the first stage does not affect the probability distribution of the negotiated coalition in the second stage.[9] Figure 1 shows the timing of the game.

We consider only Markov perfect equilibria, where each player's first-stage strategy is a function $a_i : \mathcal{N} \to \{0, 1\}$ that determines their first stage action. Given an existing coalition $M_{-1} \in \mathcal{N}$, $a_i(M_{-1}) = 1$ means that player i wants to stick to M_{-1} and $a_i(M_{-1}) = 0$ means that she wants to reopen the negotiation process. If $\prod_{j \in N} a_j(M_{-1}) = 1$, then players retain the existing coalition M_{-1}, and the game moves to the next period. Otherwise, the game moves to the second stage, where a new coalition $M \in \mathcal{N}$ emerges from the participation game. Players have rational expectations; they understand that the probability distribution π governs the second stage outcome, conditional on defection from the existing coalition. Every coalition in the support of π is stable (a Nash equilibrium).

Denote $V_i(M_{-1})$ as player i's equilibrium value of entering a period with the existing coalition M_{-1}. Generalizing equation (6), $M \in \mathcal{N}$ is stable, i.e., it is a

[8]We ignore discounting between the first and second stage of a period and other costs of reopening negotiations. Those costs would make countries less willing to abandon either a large or a small existing coalition, so their equilibrium effect is uncertain. We see no reason to think that they would alter our qualitative results.

[9]The plausible relation between past coalitions and current beliefs is ambiguous. If the last abandoned coalition was M, should players then think that M is more likely or less likely to emerge at the next round? Our assumption that prior coalitions have no effect on current beliefs is neutral with regard to this question and it makes the model tractable.

Nash equilibrium of the second-stage participation game, if and only if[10]

$$i \in M \iff u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \geq u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}). \quad (8)$$

In the first stage of a period, each player compares the payoffs associated with two scenarios, and decides whether to stick with the inherited coalition M_{-1}. If all players stick with M_{-1}, i's payoff is $u_i(M_{-1}) + \delta V_i(M_{-1})$. If any player abandons M_{-1}, thus moving to the second stage, they know that they will end up with one of the coalitions satisfying (8). Unless such a coalition is unique, it is viewed as a random variable, $\tilde{M}$, with distribution π, the common belief. Player i's payoff depends on her first-stage action only if all other players stick with the inherited coalition. We use the tie-breaking assumption that players who are indifferent between actions stick with the current coalition. Player i's expected payoff of abandoning M_{-1} and reopening the negotiation process, is

$$\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] := \sum_{M \in \mathcal{M}} \left(u_i(M) + \delta V_i(M) \right) \pi_M,$$

where

$$\mathcal{M} := \{ M \in \mathcal{N} \mid M \text{ satisfies (8)} \}. \quad (9)$$

Hence, player i will stick with M_{-1} if and only if

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi [u_i(\tilde{M}) + \delta V_i(\tilde{M})],$$

which determines the policy function a_i. The policy function, together with the common belief π, determines the value function. The value function in turn affects the equilibrium belief via (9). Therefore, at equilibrium, the common belief $(\pi_M)_{M \in \mathcal{M}}$ and the policy function $(a_i)_{i \in N}$ are simultaneously determined.

Definition 2.1. A list $(\pi, (a_i)_{i \in N})$ is an equilibrium of model $\langle \delta, N, (u_i)_{i \in N} \rangle$ if and only if there exist value functions $(V_i)_{i \in N}$ such that:
a) the support $\mathcal{M}$ of the common belief π is given by

$$\mathcal{M} = \{ M \in \mathcal{N} \mid M \text{ satisfies (8) given } (V_i)_{i \in N} \}; \quad (10)$$

[10]Equations (6) and (8) have the same interpretation. Taking as given the actions of other players at the second stage, no agent wants to change its membership status.

b) the policy functions $(a_i)_{i\in N}$ satisfy

$$a_i(M_{-1}) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ \left[u_i(M_{-1}) + \delta V_i(M_{-1})\right] a_i + \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] (1 - a_i) \Bigg\}; \tag{11}$$

c) the value functions $(V_i)_{i\in N}$ solve

$$V_i(M_{-1}) = \begin{cases} u_i(M_{-1}) + \delta V_i(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] & \text{otherwise.} \end{cases} \tag{12}$$

Condition (10) requires that the equilibrium common belief be rationalizable in the sense that every coalition in its support is stable and every coalition outside the support is not stable under the belief.[11] Condition (11) states that player i chooses $a_i = 1$ whenever she would like to use the preceding coalition, even if she knows it will be blocked by other players. This condition follows from our tie-breaking assumption, and it rules out uninteresting equilibria where a player chooses $a_i = 0$ simply because another player chooses $a_j = 0$.[12]

To simplify the analysis, we emphasize a class of equilibria where π treats players symmetrically, in the sense that the probability of forming a coalition of a particular size is independent of the identity of its members:[13]

Definition 2.2. A common belief π is symmetric if

$$|M| = |M'| \implies \pi_M = \pi_{M'},$$

Symmetric beliefs are reasonable when players are symmetric. Moreover, if π is interpreted as a randomization device used to facilitate coordination, players would not unanimously agree to use the device unless it treats them impartially.

[11] To see that the latter requirement is necessary, let M be a coalition not included in the support of the equilibrium belief. As a thought experiment, however, players can ask themselves what happens if M emerges as a candidate coalition during the negotiation process. If M satisfies the stability condition (8), players realize that the negotiation process can actually result in M, invalidating the original belief which excludes M from its support.

[12] Condition (12) implies that even non-members can trigger the abandonment of the inherited coalition. The modification where non-members do not have this veto power would not change the equilibrium in the presence of a free-rider problem. There, if members of a coalition want to stick with the coalition, so do non-members.

[13] The assumption of symmetric beliefs is common in multistage participation games, e.g. where investment precedes the participation decision (Barrett, 2006).

3 Results

We show that if agents are impatient, every stable coalition to the dynamic game has m_* members, just as in the static game. These coalitions are repeatedly formed and subsequently abandoned, and they do little to solve the collective action problem. However, if agents are patient, stable coalitions have either m_* members or more members. The small coalitions are abandoned in the next period, but the larger coalitions, once formed, are never abandoned: they are sustainable. There are no equilibrium structures with coalitions having three or more sizes. We discuss equilibrium selection when agents are patient.

To characterize the equilibrium, we rely upon the following assumption, consistent with the essential aspects of the examples above.

Assumption 1. The reduced-form payoff functions are symmetric across players and there exists an integer $m_* \in \{2, 3, \ldots, n-2\}$ such that

$$u_i(M) > u_i(M \cup \{i\}) \quad \forall i \in N \setminus M \iff |M| \geq m_* \tag{13}$$

and

$$u_i(M) \geq u_i(M \setminus \{i\}) \quad \forall i \in M \iff |M| \leq m_*. \tag{14}$$

Moreover, for any $M \in \mathcal{N}$ such that $|M| \geq m_* - 1$,

a) $|M| < |M'|$ implies $u_i(M) \leq u_i(M')$ for all $i \in M \cap M'$ and the second inequality is strict if $|M| \geq m_*$;

b) $|M| < |M'|$ implies $u_i(M) < u_i(M')$ for all $i \notin M \cup M'$;

c) $|M| < |M'|$ implies $\sum_{i \in N} u_i(M) < \sum_{i \in N} u_i(M')$;

d) $u_i(M) \leq u_j(M)$ for all $i \in M$ and $j \notin M$ and the inequality is strict if $|M| \geq m_*$.

Conditions (13) and (14) imply that the size of any stable coalition to the one-shot game, m_*, is unique. Properties a) and b) mean that a larger coalition is preferable both for coalition members and non-members, and property c) requires that the aggregate period payoff increases in the coalition size. Property d) implies that the economy suffers from a free-rider problem. In view of the assumed symmetry of the reduced-form payoff functions, we often use u_{in}^m and u_{out}^m to denote the period payoffs of members and non-members, respectively, when the size of current coalition is m.

3.1 Equilibrium with a single coalition size

Here we present a pessimistic result showing that even in the dynamic setting all equilibria might have only m_* members, just as in the static model. This result uses the following notation. For each $m \in \{1, 2, \ldots, n\}$, define the average payoff

$$\bar{u}^m := \frac{m}{n} u_{in}^m + \left(1 - \frac{m}{n}\right) u_{out}^m$$

and observe that under Assumption 1-c) and -d)

$$u_{in}^m < \bar{u}^m < u_{out}^m \quad \forall m \geq m_* - 1.$$

Because the aggregate period payoff strictly increases in $m \geq m_* - 1$, so does the average payoff $\bar{u}^m$. We denote l^* as the smallest coalition for which insiders' payoff is no less than the average payoff when the coalition has m_* members. That is $l^* > m_*$ is defined by

$$u_{in}^{l^*} \geq \bar{u}^{m_*} > u_{in}^{l^*-1};$$

l^* exists and is unique under Assumption 1 because $u_{in}^n = \bar{u}^n > \bar{u}^{m_*} > u_{in}^{m_*}$ and u_{in}^m is strictly increasing in $m \geq m_*$.

Proposition 3.1. *Under Assumption 1, the strategy profile $(a_i)_{i \in N}$ defined by*

$$a_i(M_{-1}) = \begin{cases} 1 & \text{if } |M_{-1}| \geq l^* \text{ and } i \in M_{-1} \\ 1 & \text{if } |M_{-1}| \geq m_* \text{ and } i \notin M_{-1} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

together with the symmetric common belief π defined by

$$\pi_M = 1/C_{m_*}^n \quad \forall M \in \mathcal{M},$$

where

$$\mathcal{M} := \{M \in \mathcal{N} \,|\, |M| = m_*\},$$

constitutes an equilibrium if and only if

$$\delta < \delta_{l^*} := \frac{u_{out}^{l^*-1} - u_{in}^{l^*}}{u_{out}^{l^*-1} - \bar{u}^{m_*}} \in (0, 1].$$

In the equilibrium described in Proposition 3.1, players believe that reopening

the negotiation process always results in a coalition of size m_*. This belief is rationalizable because under it the second-stage participation game yields only coalitions of size m_*. In the first stage of each period, players collectively choose to stay with the coalition they inherit from the preceding period if and only if it is larger than or equal to $l^* > m_*$. If the dynamic game begins at $t = 0$ with a coalition smaller than l^*, players for $t \geq 1$ inherit a coalition of size m_*. Players abandon the inherited coalition and start over every period. The coalition size remains constant at m_*, but the identity of members changes. This equilibrium exists if and only if players are sufficiently impatient ($\delta < \delta_{l^*}$).[14]

The equilibrium values of m_*, l^* and δ_{l^*} are highly nonlinear discontinuous functions of model parameters. Appendix B discusses these functions for $n = 15$. In Example 1 $m_* \in \{2, 3\}$ for $\gamma > 1.2$, a range that includes the quadratic case, $\gamma = 2$, used in many papers. For $\gamma > 1.2$ the pessimistic outcome, where all stable coalitions have m_* members, exists only if $\delta < 0.6$. Thus, although the dynamic model produces the pessimistic static result in some circumstances, a moderate level of patience implies that the support of any equilbrium belief *must* contain larger coalitions. In Example 2, small changes in c can lead to large changes in δ_{l^*}. For a given δ, a small change in c can cause the nature of the equilibrium to change. Thus, for both Examples the dynamic and static versions of the model may have quite different implications.

3.2 Equilibria with multiple coalition sizes

We say that a coaltion is *sustainable* if it is both stable and, once formed, permanent. The requirement of stability means that the coalition can be formed during the second-stage negotiation: it can therefore be reached even if the preceding coalition was smaller. Sustainability means that members are willing to remain in the coalition even though by doing so they give up the possibility of free riding. Members make this tradeoff only if they are sufficiently patient, i.e., if the discount factor is large.[15] Proposition 3.2 characterizes equilibria for large δ, where there are both small and large stable coalitions. Only the large coalitions are sustainable. This proposition defines the endogenous probability

[14]This type of equilibrium does not exist for a larger δ because of condition (10), which requires that every stable coalition is included in the support of the equilibrium common belief. When the discount factor is large enough, coalitions of size l^* are stable, invalidating the belief that the negotiation process always results in a coalition of size m_*.

[15]There are no sustainable equilibria if agents are very impatient. In this case, every stable coalition has m_* members, as in Proposition 3.1. However, these coalitions are not permanent: in each period, they disband and a new one forms.

that negotiation results in a large (and sustainable) coalition; it makes the term "sober optimism" precise. We then show that there do not exist equilibria with coalitions having three or more different sizes.

Proposition 3.2. *For each $m^* \geq \max\{l^*, m_* + 2\}$, (a) and (b) are equivalent:*

a) There exists a symmetric common belief π with

$$\mathcal{M} = \{M \in \mathcal{N} \,|\, |M| \in \{m_*, m^*\}\},$$

and integer k^* with $m_* \leq k^* \leq m^*$ for which the strategy profile $(a_i)_{i \in N}$ defined by

$$a_i(M_{-1}) = \begin{cases} 1 & \text{if } |M_{-1}| \geq m^* \text{ for all } i \\ 1 & \text{if } |M_{-1}| \geq k^* \text{ and } i \notin M_{-1} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

constitutes an equilibrium.

b) The discount factor δ is greater than

$$\delta_{m^*} := \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \max\{\bar{u}^{m_*}, u_{in}^{m^*-1}\}} \in (0, 1]. \tag{17}$$

The common belief associated with this equilibrium is given by

$$\pi_M = \begin{cases} \pi^{m^*}/C_{m^*}^n & \text{if } |M| = m^* \\ \left(1 - \pi^{m^*}\right)/C_{m_*}^n & \text{if } |M| = m_* \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where $\pi^{m^*} \in (0, 1)$ can be any value in the interval

$$\Pi_\delta^{m^*} := \left(\max\left\{0, \frac{(1-\delta)\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}{\bar{u}^{m^*} - \bar{u}^{m_*} - \delta\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}\right\}, \frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}\right] \subset (0, 1). \tag{19}$$

For a given discount factor, two forces constrain m^*, the size of the large coalition. A stable coalition of the second-stage game cannot be too large, or members would want to defect and free ride for a period. However, m^* cannot be too small, because otherwise members of a coalition with k^* countries would

not be willing defect, in the hope of obtaining m^*. In general, the equilibrium value of m^* is not unique. For a given δ any integer in the set

$$\{m \in \mathbb{N} \mid \max\{l^*, m_* + 2\} \leq m \leq n, \delta > \delta_{m^*}\} \tag{20}$$

can be an equilibrium value of m^*. Define $\bar{m}^*$ as the largest element in this set, i.e., the largest stable coalition; $\bar{m}^*$ is an increasing function of δ.

Unlike the equilibrium presented in Proposition 3.1, the common belief in Proposition 3.2 is not uniquely determined, although π^{m^*}, the probability of drawing a coalition with m^* members, must lie in the interval given by (19). If π^{m^*} is too large, members want to deviate from a large coalition because of the high probability that they will be free-riders to a future large coalition; in that case, m^* would not be stable. Thus, large coalitions cannot be too easy to reproduce, once abandoned. However, the equilibrium π^{m^*} must be large enough so that players want to abandon any coalition smaller than m^*. Restriction (19) provides a precise meaning to "sober optimism".

The next proposition shows that there is no symmetric equilibrium with three or more distinct stable coalition sizes. The proof of this surprising result proceeds by showing that if such equilibria did exist, then (at least) two types of stable coalitions have more than m_* members. Moreover, both of these coalitions are sustainable, but defection by any member renders them no longer sustainable. This conclusion implies that there exists a sustainable coalition and another coalition at least as large that is not sustainable. We show that this (implausible) implication must be false, thus ruling out the possibility of stable coalitions with three or more sizes.

Proposition 3.3. *The support of any symmetric equilibrium belief cannot contain coalitions of three or more distinct sizes.*

3.2.1 Illustrating Proposition 3.2

We illustrate Proposition 3.2 using the two Examples above. For both, we obtain a simple formula for δ_{m^*}, the threshold discount factor above which there are both small and large stable coalitions.

Proposition 3.4. *In Example 1, for each* $m^* > \max\{l^*, m_* + 1\}$*, the equilibria*

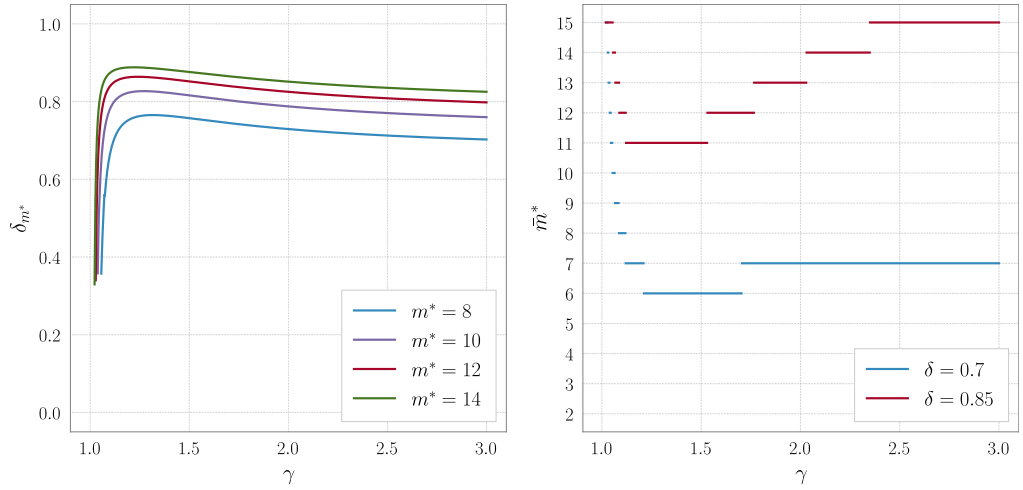

Figure 2: The threshold value δ_{m^*} of discount factor (left) and the largest size $\bar{m}^*$ of stable coalitions (right) in the model of Example 1, for $n = 15$.

described in Proposition 3.2 exist if and only if δ is greater than

$$\delta_{m^*} = \frac{\gamma(m^* - 1)^{\frac{\gamma}{\gamma-1}} - (\gamma - 1)((m^*)^{\frac{\gamma}{\gamma-1}} - 1)}{(m^* - 1)^{\frac{\gamma}{\gamma-1}} - 1}.$$

A larger sustainable coalition requires more patience: δ_{m^*} increases in m^*.

The left panel of Figure 2 shows that δ_{m^*} is non-monotonic in γ. As the value of γ increases from its lower bound, 1, (i.e., as the pollution abatement cost function becomes slightly convex), players must be much more patient to sustain large coalitions. This result is consistent with the analysis in the static setting, where the stable coalition size, m_*, falls with γ (Remark 1). However, for higher convexity, the threshold value δ_{m^*} falls with γ. In the dynamic setting, stronger convexity can make it easier (by requiring less patience) to achieve large sustainable coalitions. The right panel of Figure 2 shows this relation more clearly, graphing the largest equilibrium coalition, $\bar{m}^*$, as a function of γ for two values of δ. As γ increases, the value of $\bar{m}^*$ initially decreases, but then increases once the cost function becomes sufficiently convex. The grand coalition can be sustained when $\delta = 0.85$ and $\gamma > 2.3$.

Even in the dynamic setting, the equilibrium in Example 1 is independent of the marginal damage parameter, c. In Example 2, in contrast, the equilibrium depends on the marginal damage parameter in a striking manner.

Proposition 3.5. *In Example 2, for each $m^* > \max\{l^*, m_* + 1\}$, the equilibria*

described in Proposition 3.2 exist if and only if δ is greater than

$$\delta_{m^*} = 1 - c.$$

The value of δ_{m^*} is decreasing in c, but is independent of m^*.

We know from Remark 2 that in the static version of the same model, a large coalition requires small marginal damage. Proposition 3.5 shows that the opposite is true in the dynamic model: larger marginal damage decreases the patience needed to sustain a large coalition, and in that sense makes large coalitions – even the grand coalition – easier to achieve. For a given discount factor δ, it is possible to sustain the grand coalition when $c > 1 - \delta$.

The mechanism behind this result is quite simple, and depends on the stage-2 stability condition. In stage 2, a member's incentive to leave a coalition *falls* with c. If a member of a coalition of size $m^* > m_*$ leaves the coalition, she obtains the immediate net benefit of $u_{out}^{m^*-1} - u_{in}^{m^*} = 1 - c$ (the abatement cost the player avoids by leaving the coalition, minus the private benefit she receives from this abatement). Because $m^* > m_* \geq 1/c$, her defection does not influence the abatement levels of the other players in the current period. Coalitions of size m^* are not stable in the static setting because the short run benefit of defecting, $1 - c$, is positive, and there is no long run cost of defecting. In the dynamic setting, however, a player needs to take into account (at stage 2) the next-period consequence of a current deviation from a coalition with m^* members. That deviation causes players to enter the next period with a coalition of size $m^* - 1$. The remaining members disband this coalition, inflicting a long run cost on the erstwhile member who defected in the previous period. The next period round of negotiation might result in a small coalition, m_*. The cost of leaving the coalition depends on the discount factor. To discourage members from defecting, the discount factor needs to be large enough to counteract the immediate net benefit of leaving, which is $1 - c$.

3.2.2 Equilibrium beliefs

Characterizing the equilibrium belief is not straightforward even in these examples, but numerical illustrations paint a clear picture. For Example 1, we fix $\gamma = 2$ with $n = 15$ and depict in the left panel of Figure 3 the possible equilibrium combinations of m^* and π^{m^*} for four values of δ. Here, $m_* = 3$, $l^* = 5$, and $\delta_{l^*} = 0.588$. If $\delta < 0.588$ there exists only the small equilibrium characterized

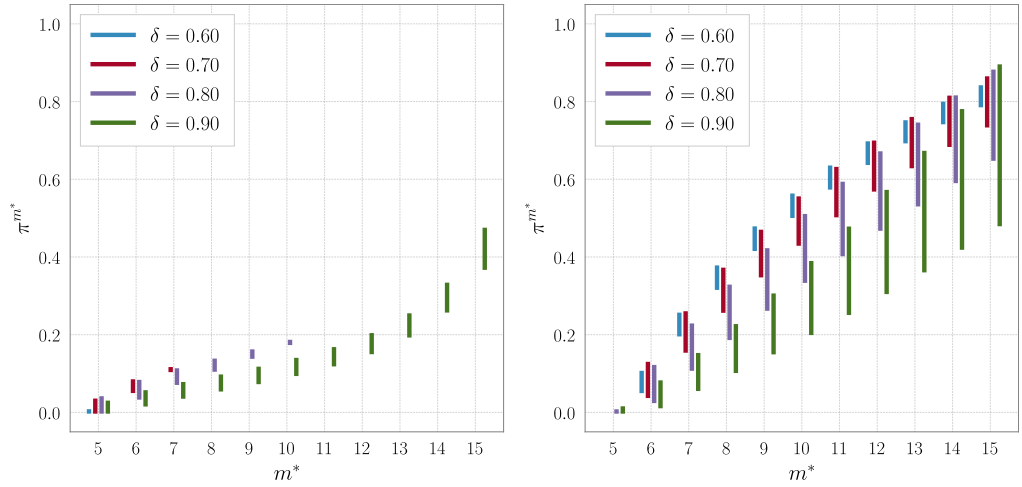

Figure 3: The equilibrium beliefs in Example 1 (left) and Example 2 (right). For each m^*, each bar represents the range $\Pi_\delta^{m^*}$ of possible values of π^{m^*} for different δ. The number of players is set to $n = 15$ for both cases. In Example 1 we set $\gamma = 2$ and in Example 2 we set $c = 0.475$. In both examples, $m_* = 3$ and $l^* = 5$.

by Proposition 3.1. Along the equilibrium path, coalitions of size $m_* = 3$ are repeatedly formed and then abandoned.

When δ is greater than 0.588, however, larger coalitions can emerge as sustainable outcomes. When $\delta = 0.6$, for example, the stable set consists of both the small coalitions with $m_* = 3$ members and the large (sustainable) coalitions with $m^* = 5$ members. The common belief associated with $m^* = 5$ is $\pi^{m^*} \in \Pi_\delta^{m^*} = (0, 0.005]$. A new round of negotiation is believed to produce coalitions with five members with probability less than or equal to 0.005. Once a coalition with five members is formed, players will stick with it. Thus, for $\delta = 0.6$, even though the exact value of π^{m^*} is not pinned down, the size of the larger stable coalitions is uniquely determined as $m^* = 5$.

If the discount factor is larger, say $\delta = 0.7$, the size m^* of the larger coalitions may be either $m^* = 5$, 6, or 7, and the associated range of π^{m^*} is given by $(0, 0.032]$, $(0.053, 0.082]$, and $(0.107, 0.113]$, respectively. As δ gets closer to 1, even larger coalitions can be stable. In particular, the grand coalition can be in the support of equilibrium belief if δ is greater than 0.861.

The right panel of Figure 3 presents the result of a similar exercise for Example 2. Here we set $c = 0.475$, so $m_* = 3$, $l^* = 5$, and $\delta_{l^*} = 0.777$. If δ is smaller than 0.777, the equilibrium characterized by Proposition 3.1 exists, where all stable coalitions have $m_* = 3$ members. By Proposition 3.5, $\delta_{m^*} = 1 - c = 0.525$ for all $m^* > l^* = 5$. Therefore, any coalition of size $m^* \in \{6, \ldots, 15\}$ can be

stable if δ is greater than 0.525. However, the associated value of π^{m^*} varies with m^* and δ. For large m^*, the range $\Pi_\delta^{m^*}$ of possible value of π^{m^*} becomes wider as the discount factor gets closer to 1.

The one-shot setting predicts the same outcome, $m_* = 3$, in both of these examples. But in the dynamic setting, these models give significantly different predictions. For instance, in Example 1 with $n = 15$ and $\gamma = 2$, the symmetric equilibrium is always characterized by either Proposition 3.1 or Proposition 3.2 for a given value of δ. (For the same value of δ there cannot be equilibria with two sizes of coalitions and also an equilibrium with only the smaller size m_*.) Example 2, in contrast, allows the two types of equilibria to coexist when the discount factor lies in between 0.525 and 0.777. When δ is in this range, the players may end up with the equilibrium where the negotiation always yields a coalition with three members; however, they might end up with another equilibrium having a larger coalition, even the grand coalition. Moreover, in Example 1, while medium-sized coalitions might be stable outcomes even for moderate levels of δ, very large coalitions require that the discount factor is close to 1. Example 2, in contrast, predicts that all of the coalitions larger than 5 can be stable for a moderate discount factor.

3.3 Equilibrium selection

There may be many equilibria to the dynamic game, raising the issue of equilibrium selection. For games with small δ, the multiplicity is small. In the example depicted in the left panel of Figure 3, the equilibrium size of larger stable coalitions is unique, $m^* = 5$, whenever $\delta \in (0.588, 0.625)$. Here, multiplicity exists only because the value of π^{m^*} is not pinned down. For games with a larger δ, however, different equilibria may involve different sizes of stable coalitions. If δ is 0.7 in the same example, players may end up with an equilibrium where the larger stable coalitions has $m^* = 7$ members or they may find themselves trapped in another equilibrium where the stable coalitions cannot be larger than $m^* = 5$. Example 2 has even greater multiplicity of equilibria.

Multiplicity helps explain the fact that international environmental agreements sometimes fail and sometimes succeed in attracting a large number of members. The actual outcome may depend on self-fulfilling beliefs generated by the political climate. In a "soberly optimistic" environment, countries believe that large coalitions are possible, leading to a good outcome. If there is little political momentum to solve the problem, in contrast, countries believe that only

small coalitions are possible, making it impossible to achieve a larger coalition.

Here we consider two refinements that select the number of participants of the large coalition, m^*. The first refinement uses the assumption that an increase in the width of the interval $\Pi_\delta^{m^*}$ increases the plausibility of the corresponding value of m^*. The second uses Pareto efficiency.

To motivate the first refinement, suppose that a shock (e.g. an election result) shifts the common belief. This shift might cause the updated value of π^{m^*} to leave the admissible range, $\Pi_\delta^{m^*}$ given by (19), unless this interval is sufficiently wide. A narrower $\Pi_\delta^{m^*}$ requires more precise coordination of beliefs among otherwise uncoordinated players. This reasoning suggests that whenever multiple values for m^* are possible, the one associated with the largest interval $\Pi_\delta^{m^*}$ is most likely to materialize. This refinement selects m^* as a solution to

$$m^* \in \underset{m}{\operatorname{argmax}}\{\max_\pi \Pi_\delta^m - \inf_\pi \Pi_\delta^m\}. \tag{21}$$

The second refinement selects the Pareto Efficient equilibrium from the set of feasible equilibria. An agent's ex ante payoff equals her expected payoff before learning the result of the negotiation. Her ex ante flow payoff conditional on a coalition of size m emerging from the negotiation is $\bar{u}^m$. This ex ante conditional payoff increases in m. Therefore, a sufficient condition for the the unconditional ex ante payoff to increase in m^* is that the probability that negotiation produces a large coalition (m^* instead of m_*) also increases in m^*. Because the mapping from m^* to the probability π^{m^*} is a correspondence, not a function, the meaning of this sufficient condition is ambiguous. However, it seems reasonable to assume that if a larger m^* shifts up the interval $\Pi_\delta^{m^*}$, i.e. causes both its boundaries to increase, then the probability of m^* also increases. With this assumption, a sufficient condition for equilibria with larger m^* to Pareto dominate equilibria with smaller m^* is that a larger m^* shifts up $\Pi_\delta^{m^*}$.

Inspection of Figure 3 shows that for our numerical examples, both refinements select the largest feasible m^*: an increase in m^* causes the interval $\Pi_\delta^{m^*}$ to become wider and also to shift up. The next proposition provides evidence that these results hold more generally.

Proposition 3.6.

a) Under Assumption 1 there always exists $\delta^* \in (0,1)$ such that

$$\underset{m}{\operatorname{argmax}}\{\max_\pi \Pi_\delta^m - \inf_\pi \Pi_\delta^m\} = \{n\}$$

for any $\delta > \delta^*$.

b) For Example 2*, an equilibrium with larger* m^* *Pareto dominates any equilibrium with smaller* m^*.

Part (a) shows that when players are sufficiently patient, under our first refinement they keep reopening the negotiation process until they achieve the grand coalition. Part (b) shows that for Example 2, coalitions with larger m^* Pareto dominate, smaller coalitions. In all cases, the negotiation process may produce many short-lived agreements with small membership along the way.

4 Structural models

The discussion above uses a reduced-form model where the period payoff is a function of only the coalition in that period. This approach significantly simplifies the analysis while keeping the generality of the model fairly intact. However, the reduced-form focus limits our results' applicability because not every model has a reduced-form representation. In particular, the absence of stock variables may seem restrictive: climate change involves greenhouse gas stocks. Here we present an isomorphism, showing the features of a model with stock variables having a reduced-form representation.

4.1 The model

To establish the isomorphism, we define a structural (as distinct from reduced-form) model, one characterized by a list $\langle \delta, N, (\Phi_i)_{i\in N}, F, T\rangle$; as above, $\delta \in (0,1)$ is the discount factor and $N := \{1, 2, \ldots, n\}$ is the set of all players. The function $\Phi_i(\boldsymbol{g}_t, G_t)$ determines the period payoff; the vector $\boldsymbol{g}_t := (g_{1,t}, \ldots, g_{n,t})$ contains the players' emissions, which affect the evolution of the stock G_t, a public bad such as greenhouse gasses. The integer $T \leq \infty$ equals the number of periods. We are primarily interested in the case with $T = \infty$, but we also need to consider finite-period versions of the model in order to define limit equilibria. The equation of motion for G is

$$G_t = F(\boldsymbol{g}_t, G_{t-1})$$

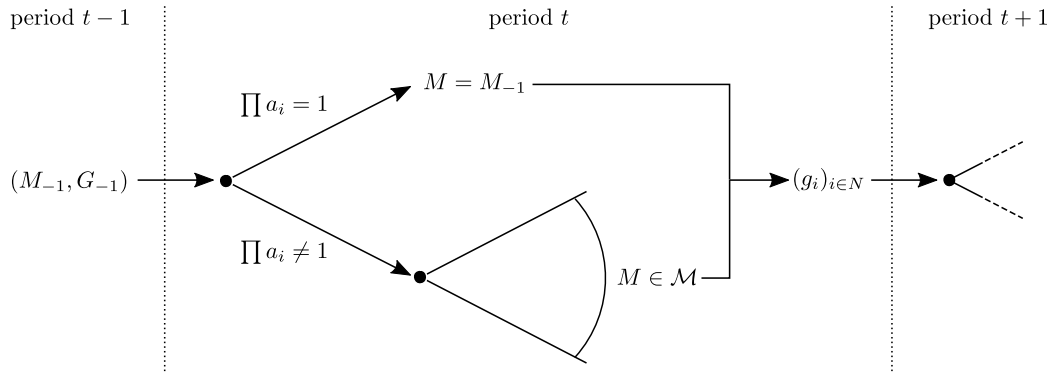

Figure 4: The timing of the structural game.

for some function F. Player i's discounted present-value payoff at $t \leq T$ is

$$\sum_{s=t}^{T} \delta^{s-t} \Phi_i(\boldsymbol{g}_s, G_s).$$

The game proceeds as in the preceding section, but now players choose their contribution to the public bad (emissions) in each period after a coalition forms. Members of a coalition jointly choose their g_i's to maximize their aggregate life-time payoff, and each non-member chooses g_i to maximize her individual life-time payoff. To simplify the notation, let $\tau \leq T$ denote the number of remaining periods. Each player's strategy is a pair of policy rules, a function $a_i^T(M_{-1}, G_{-1}, \tau) \in \{0, 1\}$ that determines whether an agent sticks with the existing coalition in the first stage, and a real-valued function $g_i^T(M, G_{-1}, \tau)$, that determines her contribution to G at the end of each period. Figure 4 depicts the timing of the game.

Let $V_i^T(M_{-1}, G_{-1}, \tau)$ be player i's continuation value when the economy has τ periods to go, conditional on the coalition M_{-1} and the level of the public bad G_{-1} inherited from the preceding period. We define $V_i^T(M_{-1}, G_{-1}, 0) := 0$. In the second stage of the period game, coalition $M \in \mathcal{N}$ is a Nash-equilibrium (i.e., stable) outcome if and only if

$$i \in M \iff \begin{array}{l} \hat{\Phi}_i^T(M \cup \{i\}, G_{-1}, \tau) + \delta \hat{V}_i^T(M \cup \{i\}, G_{-1}, \tau - 1) \\ \qquad \geq \hat{\Phi}_i^T(M \setminus \{i\}, G_{-1}, \tau) + \delta \hat{V}_i^T(M \setminus \{i\}, G_{-1}, \tau - 1), \end{array} \tag{22}$$

where

$$\hat{\Phi}_i^T(M, G_{-1}, \tau) := \Phi_i(\boldsymbol{g}^T(M, G_{-1}, \tau), F(\boldsymbol{g}^T(M, G_{-1}, \tau), G_{-1}))$$

and

$$\hat{V}^T(M, G_{-1}, \tau - 1) := V_i^T(M, F(\boldsymbol{g}^T(M, G_{-1}, \tau), G_{-1}), \tau - 1).$$

Condition (22) is the structural analogue of (8). With this notation, we can define the equilibrium of structural models for $T \leq \infty$.

Definition 4.1. A list $(\pi^T, (a_i^T)_{i\in N}, (g_i^T)_{i\in N})$ is an equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, T\rangle$ if there exist value functions $(V_i^T)_{i\in N}$ such that a) for each G_{-1} and τ, the support $\mathcal{M}^T(G_{-1}, \tau)$ of the common belief π^T is given by

$$\mathcal{M}^T(G_{-1}, \tau) = \{M \in \mathcal{N} \,|\, M \text{ satisfies (22) given } (V_i^T)_{i\in N},\ (g_i^T)_{i\in N}, \text{ and } G_{-1}\}; \tag{23}$$

b) the policy functions $(a_i^T)_{i\in N}$ satisfy

$$a_i^T(M_{-1}, G_{-1}, \tau) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ \left[\hat{\Phi}_i^T(M_{-1}, G_{-1}, \tau) + \delta \hat{V}_i^T(M_{-1}, G_{-1}, \tau - 1)\right] a_i + \mathbb{E}_{\pi^T}\left[\hat{\Phi}_i^T(\tilde{M}, G_{-1}, \tau) + \delta \hat{V}_i^T(\tilde{M}, G_{-1}, \tau - 1)\right](1 - a_i) \Bigg\}; \tag{24}$$

c) the policy functions $(g_i^T)_{i\in N}$ solve

$$\begin{aligned}(g_i^T(M, G_{-1}, \tau))_{i\in M} \in \ & \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^T(M, F(\boldsymbol{g}, G_{-1}), \tau - 1)\right\} \\ & \text{s.t.} \quad g_j = g_j^T(M, G_{-1}, \tau) \quad \forall j \notin M,\end{aligned}$$

$$\begin{aligned} g_i^T(M, G_{-1}, \tau) \in \ & \underset{g_i}{\operatorname{argmax}} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^T(M, F(\boldsymbol{g}, G_{-1}), \tau - 1)\right\} \\ & \text{s.t.} \quad g_j = g_j^T(M, G_{-1}, \tau) \quad \forall j \in N \setminus \{i\} \end{aligned} \qquad \forall i \notin M;$$

d) the value functions $(V_i^T)_{i\in N}$ solve

$$V_i^T(M_{-1}, G_{-1}, \tau) = \begin{cases} \hat{\Phi}_i^T(M_{-1}, G_{-1}, \tau) + \delta \hat{V}_i^T(M_{-1}, G_{-1}, \tau - 1) & \text{if } \prod_{j\in N} a_j(M_{-1}, G_{-1}, \tau) = 1 \\ \mathbb{E}_{\pi^T}\left[\hat{\Phi}_i^T(\tilde{M}, G_{-1}, \tau) + \delta \hat{V}_i^T(\tilde{M}, G_{-1}, \tau - 1)\right] & \text{otherwise.} \end{cases} \tag{25}$$

This definition is a straightforward extension of Definition 2.1. We are interested in the case where $T = \infty$, where the common belief, the policy functions, and the value functions are all independent of the number of remaining periods,

τ. Unlike the equilibria of reduced-form models, however, $\mathcal{M}$ and a might still depend on G_{-1}.[16]

Definition 4.2. An equilibrium $(\pi^\infty, (a_i^\infty)_{i\in N}, (g_i^\infty)_{i\in N})$ of the infinite time horizon structural model is called a limit equilibrium if for each $T < \infty$ there exists an equilibrium $(\pi^T, (a_i^T)_{i\in N}, (g_i^T)_{i\in N})$ of the T-period version of the same structural model such that $(\pi^\infty, (a_i^\infty)_{i\in N}, (g_i^\infty)_{i\in N})$ is a point-wise limit of $(\pi^T, (a_i^T)_{i\in N}, (g_i^T)_{i\in N})$ as $T \to \infty$.

4.2 Isomorphism

Structural models are isomorphic to reduced-form models if there exists a mapping between the two types of model such that a) any equilibrium of the reduced-form representation of an (infinite time horizon) structural model coincides with an equilibrium of the structural model, and b) any limit equilibrium of a structural model coincides with an equilibrium of the associated reduced-form model. The key assumption is *linearity-in-state*.

Assumption 2 (Linearity-in-state)**.** The per-period payoff function of structural models is given by

$$\Phi_i(\boldsymbol{g}_t, G_t) = \phi_i(\boldsymbol{g}_t) - cG_t$$

for some function $\phi_i(\cdot)$ and constant $c > 0$, and the equation of motion for G is

$$F(\boldsymbol{g}_t, G_{t-1}) = f(\boldsymbol{g}_t) + \sigma G_{t-1}$$

for some function $f(\cdot)$ and constant $\sigma \in [0, 1)$.

Battaglini and Harstad's (2016) model of international environmental agreements satisfies this assumption. Even when the underlying model does not seem to satisfy this assumption, it may be possible to transform it into a linear-in-state representation, as in Golosov et al.'s (2015) climate model and Traeger's (2015) generalization.

To make structural models consistent with reduced-form models, we also need the following assumption regarding the functions ϕ_i and f.

[16]The further generalization where the probability π_M as well as the support $\mathcal{M}$ depends on G_{-1} makes the analysis unmanageable, and we do not pursue it.

Assumption 3. For each integer $\tau \leq \infty$ and $M \in \mathcal{N}$, there exists a unique vector $\hat{\boldsymbol{g}}^\tau(M) = (\hat{g}_1^\tau(M), \ldots, \hat{g}_n^\tau(M))$ that solves

$$\max_{(g_i)_{i\in M}} \sum_{i\in M} \left\{ \phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma} f(\boldsymbol{g}) \right\} \text{ given } (\hat{g}_j^\tau(M))_{j\in N\setminus M}, \tag{26}$$

and

$$\max_{g_i} \left\{ \phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma} f(\boldsymbol{g}) \right\} \text{ given } (\hat{g}_j^\tau(M))_{j\in N\setminus\{i\}} \quad \forall i \notin M, \tag{27}$$

simultaneously.

Under Assumptions 2 and 3, we can define a mapping that transforms a structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$ into a reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$ where the period payoff function is

$$u_i^\infty(M) := \phi_i(\hat{\boldsymbol{g}}^\infty(M)) - c\frac{1}{1-\delta\sigma} f(\hat{\boldsymbol{g}}^\infty(M)) \quad \forall M \in \mathcal{N}.$$

Proposition 4.1. *Under Assumptions 2 and 3, if $(\pi, (a_i)_{i\in N})$ is an equilibrium of reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$, then $(\pi, (a_i)_{i\in N}, (\hat{g}_i^\infty)_{i\in N})$ is an equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$.*

This proposition ensures that any equilibrium of the reduced-form model associated with an (infinite time horizon) structural model coincides with an equilibrium of the structural model. It enables us to characterize an equilibrium of a structural model by characterizing an equilibrium of the associated reduced-form model.

Proposition 4.1 does not necessarily mean, however, that analyzing reduced-form models is sufficient. For $T = \infty$ there may exist an equilibrium of a structural model that cannot be characterized as an equilibrium of the associated reduced-form model. The next proposition addresses this concern, showing that the converse of Proposition 4.1 is true for limit equilibria.

Proposition 4.2. *Under Assumptions 2 and 3, if $(\pi^\infty, (a_i^\infty)_{i\in N}, (g_i^\infty)_{i\in N})$ is a limit equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$, then $(\pi^\infty, (a_i^\infty)_{i\in N})$ is an equilibrium of reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$.*

Under Assumptions 2 and 3, all of the limit equilibria of a structural model can be characterized by studying the associated reduced-form model.[17]

[17] In the linear-in-state model, we can replace the scalar G with a vector at the cost of only

4.3 Examples

Examples show how the isomorphism extends the applicability of our analysis in Section 3.

Example 3. A simplified version of Battaglini and Harstad's (2016) model is represented by $\langle \delta, N, (\Phi_i)_{i\in N}, F, T\rangle$ where

$$\Phi_i(\boldsymbol{g}, G) = -\frac{1}{2}(\bar{g}_i - g_i)^2 - cG$$

and

$$F(\boldsymbol{g}, G_{-1}) = \sigma G_{-1} + \sum_{i\in N} g_i.$$

With these functional forms, Assumptions 2 and 3 are both satisfied. In particular, using superscript τ to index the parameter τ, we have

$$\hat{g}_i^\tau(M) = \begin{cases} \bar{g}_i - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma}|M| & \forall i \in M \\ \bar{g}_i - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma} & \forall i \notin M. \end{cases}$$

for each $\tau \in \{1, 2, \ldots, \infty\}$ and

$$u_i^\infty(M) = \begin{cases} -\frac{c}{1-\delta\sigma}\left\{\sum_{i\in N}\bar{g}_i - \frac{c}{1-\delta\sigma}\left(|M|^2 - |M| + n - \frac{1}{2}|M|^2\right)\right\} & \forall i \in M \\ -\frac{c}{1-\delta\sigma}\left\{\sum_{i\in N}\bar{g}_i - \frac{c}{1-\delta\sigma}\left(|M|^2 - |M| + n - \frac{1}{2}\right)\right\} & \forall i \notin M \end{cases}$$

Propositions 4.1 and 4.2 then show that it suffices to analyze the equilibrium of the associated reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$. This model, after being transformed into the reduced-form model, produces Example 1 with $\gamma = 2$.

Example 4. The following climate-economy model provides a richer structure. The discounted present-value payoff of player i is

$$\sum_{s=t}^{\infty} \delta^{s-t} \ln(C_{i,t}),$$

where $C_{i,t}$ is consumption of player i at period t. Output $Y_{i,t}$ is divided into consumption $C_{i,t}$ and investment. Assuming full depreciation of capital, we can

additional notation. That generality is important for representing a climate system, but not for our purposes here.

write the end-of-period level of capital as

$$K_{i,t} = Y_{i,t} - C_{i,t}.$$

The production function is given by

$$Y_{i,t} = \Omega(G_t) A_{i,t-1} K_{i,t-1}^{\kappa} H_i(N_{i,t}^1, \ldots, N_{i,t}^L) \quad \text{with} \quad \sum_{l=1}^{L} N_{i,t}^l = 1,$$

where G_t is the stock of carbon (after absorbing the current emission), $A_{i,t-1}$ is the total factor productivity, and $N_{i,t}^l$ is the fraction of labor used for intermediate-good production sector l. Here, $\Omega(\cdot)$ and $H_i(\cdot)$ are some functions. The production process generates carbon dioxide as a byproduct, and the level $g_{i,t}$ of carbon emission depends on the labor allocation vector $(N_{i,t}^1, \ldots, N_{i,t}^L)$ via

$$g_{i,t} = E_i(N_{i,t}^1, \ldots, N_{i,t}^L)$$

for some function $E_i(\cdot)$. The equation of motion for carbon stock is

$$G_t = F(\boldsymbol{g}_t, G_{t-1})$$

for some function $F(\cdot)$.

We can simplify this structural model provided that

$$H_i^*(g_i) := \max_{N_i^1, \ldots, N_i^L} \left\{ H_i(N_i^1, \ldots, N_i^L) \,\middle|\, E_i(N_i^1, \ldots, N_i^L) \leq g_i \right\}$$

is well defined for each $g_i > 0$. The solution of this maximization problem determines the labor allocation vector that maximizes production without exceeding carbon emissions g_i. Then, without loss of generality, we may simplify the production function as a function of emission level:

$$Y_{i,t} = \Omega(G_t) A_{i,t-1} K_{i,t-1}^{\kappa} H_i^*(g_{i,t}).$$

Moreover, denoting $s_{i,t} := K_{i,t}/Y_{i,t}$ as the savings rate, we can write

$$\begin{aligned}\sum_{v=t}^{\infty} \delta^{v-t} \ln(C_{i,v}) = \frac{\kappa}{1-\delta\kappa} \ln(K_{i,t-1}) + \frac{1}{1-\delta\kappa} \sum_{v=t}^{\infty} \delta^{v-t} \ln\left(A_{i,v-1}\right) \\ + \sum_{v=t}^{\infty} \delta^{v-t} \left(\ln(1-s_{i,v}) + \frac{\delta\kappa}{1-\delta\kappa} \ln(s_{i,v}) \right) \\ + \frac{1}{1-\delta\kappa} \sum_{v=t}^{\infty} \delta^{v-t} \left\{ \ln\left(H_i^*(g_{i,v})\Omega(G_v)\right) \right\}.\end{aligned}$$

The first and the second terms on the right-hand side are both predetermined at the beginning of period t. Moreover, since the third and the fourth terms are additive and separable, the optimal choice of savings rate can be immediately computed as $s = \delta\kappa$, irrespective of the values of $(G_v, g_{i,v})_{v=t}^{\infty}$. Consequently, we can treat the third term as a constant, with respect to the emissions choice. It follows that the normalized discounted payoff of player i can be written as

$$\sum_{v=t}^{\infty} \delta^{v-t} \ln\left(H_i^*(g_{i,v})\Omega(G_v)\right).$$

Therefore, this model is represented by a structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty \rangle$ where $\Phi_i(\boldsymbol{g}, G) = \ln\left(H_i^*(g_i)\Omega(G)\right)$.

In the economics literature, the climate system is often modeled as a linear system, which suggests specifying $F(\boldsymbol{g}, G) = \sigma G + \sum_{i\in N} g_i$. Also, in this type of model, it is reasonable to specify $\Omega(G) = e^{-cG}$ for some $c > 0$ (Hassler et al., 2016). Hence, as long as $\phi_i(\boldsymbol{g}) = \ln\left(H_i^*(g_i)\right)$ is consistent with Assumption 3, we can apply Propositions 4.1 and 4.2. This model nests Example 3.

5 Discussion

The possibility that patient agents achieve a good outcome in repeated games is familiar. Folk theorems show that when players are sufficiently patient, any efficient outcome can be supported as a subgame-perfect Nash equilibrium in infinitely repeated games (Friedman, 1971; Fudenberg and Maskin, 1986). Players are discouraged from deviating when they think that deviation triggers a punishment phase. During that phase no player, including those responsible for imposing the punishment, benefits from unilateral deviation. The folk theorems suggest that sustaining cooperative outcomes in a repeated game setting requires

some form of punishment. In Battaglini and Harstad (2016) defecting from an equilibrium coalition triggers the replacement of a long-term agreement (which circumvents a hold-up problem) by a short-term agreement (which suffers from the hold-up problem). The punishment in Kovac and Schmidt (2017), costly delay of an agreement, is explicit. Battaglini and Harstad (2016), unlike Kovac and Schmidt (2017), assume that countries are able to commit to long-term agreements.

In our model, countries understand that if in the second stage of a period they leave an agreement with threshold size m^*, remaining members will abrogate the agreement in the next period. That abandonment makes the original defector worse off, and thus plays a role similar to the punishment in previous models. Barrett (2003) notes that abandoning an agreement to punish a defector is not credible if it harms those who carry out the punishment. There are no self-harming punishments in our model, where abandoning an agreement implies neither the end of negotiation (as in the grim-trigger strategy) nor a retaliation against non-compliance (as in the getting-even strategy). Members abandon any coalition smaller than m^* not to punish others or free-ride but to make a fresh start by renegotiating. Abandoning these small coalitions makes erstwhile members better off.

Farrell and Maskin (1989) note that standard trigger strategies are not credible if players anticipate future renegotiation. After the punishment phase is triggered, players will realize that it is in their interest to renegotiate to achieve cooperation. The possibility of renegotiation undermines the credibility of self-harming punishment, calling into question the plausibility of the equilibrium that hinges on the punishment. Barrett (1999, 2002, 2003) and Finus and Rundshagen (1998) emphasize the importance of renegotiation in the IEA setting. Many IEAs stipulate regular member meetings, making renegotiation an integral part of the agreement. For this reason, they argue that self-enforcing international agreements must be renegotiation proof.

We agree that renegotiation is an integral part of the process of forming and sustaining IEAs, but we question whether transitory equilibria must satisfy Farrell and Maskin's (1989) definition of renegotiation proofness. That definition requires that the equilibria cannot be Pareto ranked. In our setting, for a sufficiently large discount factor, the stable set in the second-stage participation game consists of small and large coalitions. Large coalitions Pareto dominate small coalitions. Rational agents might believe that the second-stage

participation game could result in coalitions with m_* members even though they unanimously prefer a coalition with m^* members; this belief is rational because switching from a small stable coalition to a large one is a nontrivial move. Members of a small coalition may propose that some outsiders join to make a larger stable coalition. But the outsiders, although aware that joining is Pareto improving, would rather wait for *other* outsiders to join the coalition (unless of course $m^* = n$, where there is no 'other outsider' to relent first). The possibility of eventually achieving a large stable coalition makes it even more attractive to remain as an outsider. Moreover, once possible changes of coalition members are on the table in the second stage, even the current members of the small coalition would have an incentive to defect at that stage, hoping that others would fill their seats. That additional stage creates an additional level of uncertainty. In short, we think that opportunities for renegotiation are integral to a model of IEAs, but we reject the conclusion that *transitory* equilibria cannot be Pareto dominated. Players abandon the Pareto–dominated transitory equilibria at the earliest opportunity, the next period.

The uncertainty inherent in the negotiation process is a central motivation of our analysis. The fundamental source of uncertainty, the multiplicity of equilibria, stems from the very nature of the problem. In the IEA participation game, (for $m^* < n$) there always exists an alternative equilibrium outcome that makes at least one player better off. This alternative is due to the fact that the game is a multi-player variant of the game of chicken, where players make threats to induce others to back down. In the real-world negotiation process of IEAs, it seems, countries actually play a game of chicken. During the climate negotiation at the Hague in 2000, for instance, countries waited so long for the others to relent that at the last minute of the negotiation, when a compromise was expected, there was little time left even to understand what others were suggesting (Bodansky, 2001). After having failed to build an effective agreement in the Hague the United States rejected the existing agreement, the Kyoto Protocol. The departure of the United States, along with the fact that other large emitters like China and India had no obligations under the protocol, led to a renewed negotiation process, which eventually yielded a new agreement at Paris in 2015. It remains uncertain whether the Paris agreement will be sustainable (a long-term stable agreement) and effective.

Our definition of equilibrium is closely related to Aumann's (1974; 1987) correlated equilibrium. Negotiations usually follow a pre-negotiation phase where

countries share a basic sense of what might be possible once a higher-level negotiation begins. Because the final outcome of negotiation is still contingent upon how things unfold later in the process, the pre-negotiation phase naturally yields a state-contingent correlated strategy of Aumann (1987), which we call a common belief. The equilibrium common belief π can be seen as a correlated equilibrium of the second-stage participation game. However, the equilibrium conditions we impose on π are stronger than Aumann (1987) requires. In particular, we rule out the possibility that the communication channel is noisy or that the 'mediator' can communicate separately and confidentially with each country. Moreover, we require an equilibrium correlation device to include all of the Nash equilibrium outcomes in its support. These additional restrictions make our analysis conservative. One might obtain a larger equilbrium set by relaxing these restrictions.

In contrast to previous studies, our analysis highlights the critical role of communication. A pre-negotiation phase of international agreements works as a communication channel through which countries build a common belief (a correlation device) to coordinate their actions. In the static setting pre-play communication can influence the outcome if players can commit themselves to binding contracts or if a mediator transforms the game into one of incomplete information (Myerson, 1994).[18] Neither of these possibilities is plausible in international negotiations. Countries can always renege on their promise, and agreements are usually negotiated openly among the countries involved. We show that in a dynamic setting, even when no commitment is allowed and no mediator is available, pre-play communication can decisively affect the outcome by influencing the common belief. Players need to share the belief that a large stable coalition is possible but cannot be taken for granted: their optimism must be sober, not giddy. The belief must be accompanied by a high cut-off size for the first-stage of the period game, so that players do not settle for too little.

6 Conclusion

We provide a dynamic model of agreements among sovereign nations, in which countries abandon any agreement when doing so is in their self-interest. This possibility reflects the reality of international relations, where countries cannot

[18]Forges (1990) shows that a system of direct communication plays an important role even in the absence of mediator only if communication between any pair of players is not observable by the other players.

credibly commit to agreements. Countries can sign a long-term agreement to provide public goods, but they can, and sometimes do, break promises. Any successful international agreement should be based on the understanding that it cannot be credibly implemented solely based on promises. Agreements, like punishments, need to be credible if they are to improve upon the status quo.

The difficulty inherent in the relationship among sovereign countries makes it tempting to paint a bleak picture of international agreements. We find, however, that countries can cooperate even in the presence of a free-rider problem. If they are fairly patient, (re)opening the negotiation process might yield a large coalition. In the next period small coalitions are abandoned in an attempt to make a fresh start, and large coalitions are sustained; members of the large coalition remain compliant. This result is based on a general reduced-form model and does not require explicit sanctions or direct money transfers. There is no delay of the agreement or assumed punishment phase. Our conclusions explain the "paradox of international agreements": some negotiations achieve meaningful results, even though the circumstances might appear to doom them to failure. We also provide conditions under which the reduced-form model is isomorphic to one with a pollution stock, so our results are applicable to climate treaties.

The simple idea underlying our analysis is worth re-stating here. The exact outcome of the IEA negotiation process is inherently uncertain due to the multiplicity of equilibria. This uncertainty opens the possibility that countries continue cooperating once they reach a sufficiently good agreement. The emergence of a good agreement requires that countries set the bar sufficiently high and also believe that it is possible to clear the hurdle. Excessive optimism would undermine a large existing agreement by making members think that defection is cheap. Meaningful cooperation among sovereign countries requires sober optimism: the understanding that cooperation is possible but not easy to achieve.

References

Aumann, R. J. (1974): "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, 1, 67–96.

——— (1987): "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, 55, 1–18.

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

——— (1997): "The strategy of trade sunctions in international environmental agreements," *Resource and Energy Economics*, 19, 345–361.

——— (1999): "A theory of full international cooperation," *Journal of Theoretical Politics*, 11, 519–541.

——— (2001): "International cooperation for sale," *European Economic Review*, 45, 1835–1850.

——— (2002): "Consensus treaties," *Journal of Institutional and Theoretical Economics*, 158, 529–547.

——— (2003): *Environment and Statecraft: The strategy of Environmental Treaty-Making*, Oxford University Press.

——— (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.

——— (2006): "Climate treaties and 'breakthrough' technologies," *American Economic Review: Papers and Proceedings*, 96, 22–25.

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

Benedick, R. E. (1998): *Ozone Diplomacy: New Directions in Safeguarding the Planet*, Harvard University Press.

Bodansky, D. (2001): "Bonn voyage: Kyoto's uncertain revival," *The National Interest*, 65, 45–55.

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

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

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.

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.

de Zeeuw, A. (2015): "International environmental agreements," *Annual Review of Resource Economics*, 7, 151–168.

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," Forthcoming in *Environmental and Resource Economics*.

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

Farrell, J. and E. Maskin (1989): "Renegotiation in repeated games," *Games and Economic Behavior*, 1, 327–360.

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

Finus, M. and B. Rundshagen (1998): "Renegotition-proof equilibria in a global emission game when players are impatient," *Environmental and Resource Economics*, 12, 275–306.

Forges, F. (1990): "Universal Mechanisms," *Econometrica*, 58, 1341–1364.

Friedman, J. W. (1971): "A non-cooperative equilibrium for supergames," *Review of Economic Studies*, 38, 1–12.

Fudenberg, D. and E. Maskin (1986): "The folk theorem in repeated games with discounting or with incomplete information," *Econometrica*, 54, 533–554.

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.

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

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.

Hoel, M. and K. Schneider (1997): "Incentives to participate in an international environmental agreements," *Environmental and Resource Economics*, 9, 153–170.

Hong, F. and L. S. Karp (2012): "International environmental agreements with mixed strategies and investment," *Journal of Public Economics*, 96, 685–697.

——— (2014): "International environmental agreements with endogenous or exogenous risk," *Journal of the Association of Environmental and Resource Economists*, 1, 365–394.

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

Kolstad, C. D. and M. Toman (2005): "The economics of climate policy," in *Handbook of Environmental Economics*, ed. by K.-G. Maler and J. R. Vincent, Elsevier, vol. 3, chap. 30, 1561–1618.

Kovac, E. and R. C. Schmidt (2017): "A simple dynamic climate cooperation model," BDPEMS Working Paper No. 2015-17.

Mitchell, R. B. (2018): "International Environmental Agreements Database Project (Version 2017.1)," URL: http://iea.uoregon.edu.

Myerson, R. B. (1994): "Communication, correlated equilibria and incentive compatibility," in *Handbook of Game Theory with Economic Applications*, ed. by R. J. Aumann and S. Hart, Elsevier, vol. 2, chap. 24, 827–847.

Nordhaus, W. D. (2015): "Climate Clubs: overcoming free-riding in International Climate Policy," *American Economic Review*, 105, 1339–1370.

Oberthur, S. and H. E. Ott (1999): *The Kyoto Protocol: International Climate Policy for the 21st Century*, Springer.

Osmani, D. and R. Tol (2009): "Toward farsightedly stable international environmental agreements," *Journal of Public Economic Theory*, 11, 455–492.

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

Traeger, C. P. (2015): "Analytic integrated assessment and uncertainty," DOI: 10.2139/ssrn.2667972.

Wagner, U. J. (2001): "The design of stable international environmental agreements: economic theory and political economy," *Journal of Economic Surveys*, 15, 377–411.

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 Proofs

Remark 2 is well known, but we provide its proof to make the paper self-contained. Remark 1, to the best of our knowledge, is original. This Remark is related to Proposition 1 in Karp and Simon (2013), which shows how the curvature of marginal costs affects the largest and the smallest stable coalition. We impose more structure here, leading to uniqueness and monotonicity results.

A.1 Proof of Remark 1

To simplify the notation, we define $\theta := \gamma/(\gamma - 1)$. Note that θ is strictly decreasing in $\gamma \in (1, \infty)$ with $\lim_{\gamma \to 1} \theta = \infty$ and $\lim_{\gamma \to \infty} \theta = 1$. We now show that for each θ, there exists a unique integer m_* such that a coalition M is stable if and only if $|M| = m_*$. The integer m_* is given by

$$m_* = \min\{n, \lfloor x(\theta) \rfloor\}, \tag{A.1}$$

where $\lfloor x(\theta) \rfloor$ (the floor function) is the greatest integer weakly smaller than $x(\theta)$. Here $x(\theta)$ is the unique root of $\Gamma(x, \theta) = 0$, where $\Gamma(x, \theta)$ is defined as

$$\Gamma(x, \theta) := \theta \frac{(x-1)^\theta}{x^\theta - 1} - 1 \quad \forall x \in (1, \infty).$$

We characterize m_* by characterizing $x(\theta)$. In particular, we prove that $x(\theta)$ is strictly increasing in $\theta \in (1, \infty)$ with $\lim_{\theta \to 1} x(\theta) \in (2, 3)$ and $\lim_{\theta \to \infty} x(\theta) = \infty$. A key step in the proof is to show that $x(\theta)$ is a bijective mapping and therefore is monotonic.

With this roadmap in mind, we first prove the following lemma.

Lemma A.1. *For each $\theta \in (1, \infty)$, $\Gamma(x, \theta)$ is strictly increasing in x and there exists unique $x \in (1, \infty)$ such that $\Gamma(x, \theta) = 0$.*

Proof. Fix $\theta \in (1, \infty)$. Then we have

$$\frac{\partial \Gamma(x, \theta)}{\partial x} = \theta^2 \frac{(x-1)^{\theta-1}}{(x^\theta - 1)^2}(x^{\theta-1} - 1) > 0 \quad \forall x \in (1, \infty),$$

which means that $\Gamma(x, \theta)$ is strictly increasing in x. Using L'Hospital's Rule we have

$$\lim_{x \to 1} \Gamma(x, \theta) = -1 < 0 < \theta - 1 = \lim_{x \to \infty} \Gamma(x, \theta).$$

Because $\Gamma(x,\theta)$ is continuous and strictly increasing in x, there exists unique $x \in (1,\infty)$ such that $\Gamma(x,\theta) = 0$. □

Using Lemma A.1, we can implicitly define a function $x(\theta)$ by

$$\Gamma(x(\theta),\theta) = 0 \quad \forall \theta \in (1,\infty).$$

To characterize $x(\theta)$ as a function of θ, it is useful to define another function $H(x,\theta)$ as

$$H(x,\theta) := 1 - \ln(\theta) + \ln(x^\theta - 1) - \frac{x^\theta}{x^\theta - 1}\ln(x^\theta)$$

for each $x > 1$ and $\theta > 1$. The next lemma characterizes $H(x,\theta)$ and provides its connection to $\Gamma(x,\theta)$.

Lemma A.2. *The function $H(x,\theta)$ has the following properties:*

(i) for each $\theta \in (1,\infty)$,

$$\left.\frac{\partial \Gamma(x,\theta)}{\partial \theta}\right|_{x=x(\theta)} \gtreqless 0 \iff H(x(\theta),\theta) \gtreqless 0; \tag{A.2}$$

(ii) for each $x \in (1,\infty)$,

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x,\theta)}{\partial \theta} = H(x,1); \tag{A.3}$$

(iii) $H(x,1)$ is strictly increasing in x and $H(x,1) = 0$ has a unique root, x_1, which satisfies $2 < x_1 < 3$;

(iv) Given $x \in (1,\infty)$, $H(x,\theta)$ is strictly decreasing in θ; if $x \in (1,x_1]$, we have $H(x,\theta) < 0$ for all $\theta \in (1,\infty)$; if $x \in (x_1,\infty)$, on the other hand, there is unique $\theta \in (1,\infty)$ such that $H(x,\theta) = 0$.

Proof. (i) We have

$$\frac{\partial \Gamma(x,\theta)}{\partial \theta} = \frac{(x-1)^\theta}{x^\theta - 1}\left(1 + \ln((x-1)^\theta) - \frac{x^\theta}{x^\theta - 1}\ln(x^\theta)\right) \tag{A.4}$$

and

$$\Gamma(x(\theta),\theta) = 0 \iff (x(\theta)-1)^\theta = ((x(\theta))^\theta - 1)/\theta. \tag{A.5}$$

Using (A.5) and the definition of $H(x,\theta)$ we conclude

$$\left.\frac{\partial \Gamma(x,\theta)}{\partial \theta}\right|_{x=x(\theta)} = \frac{(x(\theta)-1)^\theta}{(x(\theta))^\theta - 1} H(x(\theta),\theta),$$

which proves (A.2).

(ii) Using (A.4), we have

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x,\theta)}{\partial \theta} = 1 + \ln(x-1) - \frac{x}{x-1}\ln(x) = H(x,1),$$

which proves (A.3).

(iii) Next we note that

$$\frac{\partial H(x,1)}{\partial x} = \frac{\ln(x)}{(x-1)^2} > 0,$$

so $H(x,1)$ is strictly increasing. Also, it is easy to see that $\lim_{x\to 1} H(x,1) = -\infty$, $\lim_{x\to\infty} H(x,1) = 1$, and

$$H(2,1) = \ln(e/4) < 0 < \ln(2e/3^{3/2}) = H(3,1).$$

Therefore, the equation $H(x,1) = 0$ has a unique root x_1 in the interval $(2,3)$.

(iv) Given $x \in (1,\infty)$,

$$\frac{\partial H(x,\theta)}{\partial \theta} = -\frac{1}{\theta}\left(1 - \frac{x^\theta}{(x^\theta-1)^2}\ln(x^\theta)\ln(x^\theta)\right) < 0,$$

so $H(x,\theta)$ is strictly decreasing in θ. Because $\lim_{\theta\to\infty} H(x,\theta) = -\infty$ and $\lim_{\theta\to 1} H(x,\theta) = H(x,1)$, and because $H(x,1)$ is strictly increasing in x, it follows that when $x \in (1,x_1]$, we have $H(x,\theta) < 0$ for all $\theta \in (1,\infty)$ and when $x \in (x_1,\infty)$, the equation $H(x,\theta) = 0$ has a unique root with respect to θ. □

Equipped with Lemma A.2, we can characterize $\Gamma(x,\theta)$ as a function of θ. The next lemma shows that the equilibrium number of members cannot be less than x_1.

Lemma A.3. *If $x \in (1,x_1]$, there is no $\theta \in (1,\infty)$ such that $\Gamma(x,\theta) = 0$.*

Proof. Fix $x \in (1,x_1]$ and suppose to the contrary that there exists $\theta \in (1,\infty)$ such that $\Gamma(x,\theta) = 0$. We establish the Lemma by falsifying this hypothesis. Let θ_x be the smallest θ satisfying $\Gamma(x,\theta) = 0$ for $x \in (1,x_1]$. Note that $x = x(\theta_x)$ by definition of $x(\theta)$.

As an intermediate step, we establish

$$\left.\frac{\partial \Gamma(x,\theta)}{\partial \theta}\right|_{\theta=\theta_x} \geq 0. \tag{A.6}$$

We confirm this inequality in two steps, first showing that it holds over the open interval $x \in (1, x_1)$ and then showing that it also holds at the boundary $x = x_1$. For the first step, note that $\lim_{\theta \to 1} \Gamma(x, \theta) = 0$. If $x \in (1, x_1)$, we know from results (ii) and (iii) of Lemma A.2 that

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x, \theta)}{\partial \theta} = H(x, 1) < 0.$$

Therefore, $\Gamma(x, \theta) < 0$ for θ close to but larger than 1. Consequently, the graph of $\Gamma(x, \theta)$ as a function of θ must cross 0 at θ_x from below. Therefore, (A.6) must hold for $x \in (1, x_1)$.

Now we move to the second step, showing that (A.6) also holds at $x = x_1$. For $x = x_1$ we use Lemma A.2 (ii), which implies

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x_1, \theta)}{\partial \theta} = H(x_1, 1) = 0.$$

To evaluate $\Gamma(x_1, \theta)$ in the neighborhood of $\theta = 1$ we use a second order approximation of the function. Using the definition of $\Gamma(x, \theta)$, we have

$$\begin{aligned} \frac{x^\theta - 1}{(x-1)^\theta} \frac{\partial^2 \Gamma(x, \theta)}{\partial \theta^2} = &\left(\ln(x-1) - \frac{x^\theta}{x^\theta - 1} \ln(x) \right) \left(\frac{x^\theta - 1}{(x-1)^\theta} \frac{\partial \Gamma(x, \theta)}{\partial \theta} + 1 \right) \\ &+ \left(\frac{\ln(x)}{x^\theta - 1} \right)^2 \theta x^\theta. \end{aligned}$$

Evaluating this expression at $x = x_1$ and taking the limit of $\theta \to 1$, we obtain

$$\begin{aligned} \lim_{\theta \to 1} \frac{\partial^2 \Gamma(x_1, \theta)}{\partial \theta^2} &= \left(\ln(x_1 - 1) - \frac{x_1}{x_1 - 1} \ln(x_1) \right) \left(\lim_{\theta \to 1} \frac{\partial \Gamma(x_1, \theta)}{\partial \theta} + 1 \right) + \left(\frac{\ln(x_1)}{x_1 - 1} \right)^2 x_1 \\ &= (H(x_1, 1) - 1)(H(x_1, 1) + 1) + \frac{(1 + \ln(x_1 - 1) - H(x_1, 1))^2}{x_1} \\ &= -1 + \frac{(1 + \ln(x_1 - 1))^2}{x_1} < 0, \end{aligned}$$

where the second line uses the definition of $H(x, 1)$ and Lemma A.2 (ii), the third line uses Lemma A.2 (iii), and the inequality is due to the fact that $x_1 \in (2, 3)$.

Therefore, $\Gamma(x_1, \theta)$ is a concave function of θ in the neighborhood of $\theta = 1$. Because this function and its partial derivative both equal 0 at $\theta = 1$, $\Gamma(x_1, \theta) < 0$ for θ close to but larger than 1. This fact means that $\Gamma(x_1, \theta)$ is increasing in the neighborhood of θ_x; thus, (A.6) holds for $x = x_1$.

We now falsify the hypothesis. By definitions of $x(\theta)$ and θ_x, $x = x(\theta_x)$.

Lemma A.2 (i) shows that (A.6) implies

$$H(x, \theta_x) \geq 0,$$

which contradicts Lemma A.2 (iv). Therefore, we conclude that there is no $\theta \in (1, \infty)$ such that $\Gamma(x, \theta) = 0$ for $x \in (1, x_1]$. □

The next lemma confirms that for $x > x_1$ there exists a unique $\theta > 1$ that satisfies $\Gamma(x, \theta) = 0$.

Lemma A.4. *For each $x \in (x_1, \infty)$, there exists a unique $\theta \in (1, \infty)$ such that $\Gamma(x, \theta) = 0$.*

Proof. Fix $x \in (x_1, \infty)$. Observe that

$$\lim_{\theta \to 1} \Gamma(x, \theta) = 0 > -1 = \lim_{\theta \to \infty} \Gamma(x, \theta)$$

and

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x, \theta)}{\partial \theta} = H(x, 1) > 0,$$

where the last equality follows from (A.3) and the next inequality follows from Lemma A.2 (iv). Hence, there exits at least one $\theta \in (1, \infty)$ such that $\Gamma(x, \theta) = 0$.

To prove the uniqueness of such θ, suppose to the contrary that $\Gamma(x, \theta) = 0$ has multiple roots with respect to θ. Let θ_x be the smallest root and $\theta'_x > \theta_x$ be the second smallest. The definition of $x(\theta)$ implies that $x(\theta_x) = x(\theta'_x) = x$.

By Lemma A.2 (ii), we know that

$$\lim_{\theta \to 1} \frac{\partial \Gamma(x, \theta)}{\partial \theta} = H(x, 1) > 0.$$

Therefore, the graph of $\Gamma(x, \theta)$ is positive for $\theta > 1$ in the neighborhood of $\theta = 1$. Consequently the graph of $\Gamma(x, \theta)$ as a function of θ either crosses 0 from above at θ_x, or the graph is tangent to 0 at that point. This observation implies the weak inequality

$$\left. \frac{\partial \Gamma(x, \theta)}{\partial \theta} \right|_{\theta = \theta_x} \leq 0,$$

which, by result (i) of Lemma A.2, is equivalent to

$$H(x, \theta_x) \leq 0. \tag{A.7}$$

We show that (A.7) and the hypothesis that $\Gamma(x, \theta) = 0$ has multiple roots imply

a contradiction. We need to consider two cases, where (A.7) holds as a strict inequality and where it holds as an equality.

CASE 1: Consider the case where (A.7) holds with strict inequality. Here, the graph of $\Gamma(x,\theta)$ crosses 0 at $\theta = \theta_x$ from above. Consequently, at $\theta = \theta'_x$, the graph of $\Gamma(x,\theta)$ must cross or touch 0 from below, implying

$$\left.\frac{\partial \Gamma(x,\theta)}{\partial \theta}\right|_{\theta=\theta'_x} \geq 0.$$

By result (i) of Lemma A.2, this inequality implies

$$H(x,\theta'_x) \geq 0.$$

Because $\theta'_x > \theta_x$ and because $H(x,\theta)$ is strictly decreasing in θ by result (iv) of Lemma A.2, we then have

$$H(x,\theta_x) > H(x,\theta'_x) \geq 0,$$

which contradicts (A.7).

CASE 2: Consider the case where $H(x,\theta_x) = 0$. Here the graph of $\Gamma(x,\theta)$ is tangent to 0 at $\theta = \theta_x$. The function is convex at this point because

$$\left.\frac{\partial^2 \Gamma(x,\theta)}{\partial \theta^2}\right|_{\theta=\theta_x} = \frac{1}{\theta_x}\frac{(x-1)^{\theta_x}}{x^{\theta_x}-1}\left[x^{\theta_x}\left(\frac{\ln(x^{\theta_x})}{x^{\theta_x}-1}\right)^2 - 1\right] > 0.$$

We establish the inequality using the fact that $\Gamma(x,\theta_x) = 0$ and $\partial\Gamma(x,\theta_x)/\partial\theta = H(x,\theta_x) = 0$. Consequently, $\Gamma(x,\theta)$ is positive in the neighborhood of θ_x except at θ_x where it equals 0.

Now, observe that for any $\tilde{x} \in (x_1, x)$, we have

$$\Gamma(\tilde{x},\theta) < \Gamma(x,\theta) \quad \forall \theta \in (1,\infty),$$

because $\Gamma(x,\theta)$ is strictly increasing in x by Lemma A.1. By making $\tilde{x}$ sufficiently close to x, then we can find $\theta_{\tilde{x}}$ and $\theta'_{\tilde{x}}$ such that $\theta_{\tilde{x}} < \theta_x < \theta'_{\tilde{x}}$,

$$\Gamma(\tilde{x},\theta_{\tilde{x}}) = \Gamma(\tilde{x},\theta'_{\tilde{x}}) = 0 \quad \text{and} \quad \Gamma(\tilde{x},\theta) < 0 \quad \forall \theta \in (\theta_{\tilde{x}},\theta'_{\tilde{x}}),$$

which implies

$$H(\tilde{x},\theta_{\tilde{x}}) < 0 < H(\tilde{x},\theta'_{\tilde{x}}). \tag{A.8}$$

However, because $\theta'_{\tilde{x}} > \theta_{\tilde{x}}$ and since $H(\tilde{x}, \theta)$ is strictly decreasing in θ by result (iv) of Lemma A.2, we then have

$$H(\tilde{x}, \theta_{\tilde{x}}) > H(\tilde{x}, \theta'_{\tilde{x}}),$$

which contradicts (A.8). □

We can now characterize $x(\theta)$ as an increasing function, which in turn allows us to prove Remark 1.

Lemma A.5. *Function $x(\theta)$ is strictly increasing in $\theta \in (1, \infty)$ with $\lim_{\theta \to 1} x(\theta) = x_1$ and $\lim_{\theta \to \infty} x(\theta) = \infty$. In particular, $x(2) = 3$ and $2 < x(\theta) < 3$ for all $\theta \in (1, 2)$.*

Proof. Combining Lemmas A.1, A.3, and A.4, we conclude that $x(\theta)$ is a bijection from $(1, \infty)$ onto (x_1, ∞). Hence, $x(\theta)$ must be monotonic. To prove that $x(\theta)$ is strictly increasing, it suffices to show that $x(2) < x(3)$. Observe

$$\Gamma(x, 2) = 0 \iff 2\frac{x-1}{x+1} = 1 \iff x = 3,$$

which implies $x(2) = 3$. Also,

$$\Gamma(3, 3) = 3\frac{(3-1)^3}{3^3 - 1} - 1 = -\frac{1}{13} < 0 = \Gamma(x(3), 3),$$

which implies $x(3) > 3$ because $\Gamma(x, 3)$ is strictly increasing in x by Lemma A.1. The last part of the lemma follows from the fact that $x(2) = 3$ and $x(\theta)$ is strictly increasing with $x(\theta) > x_1 > 2$ for all θ. □

Proof. (Remark 1) Fix $\gamma \in (1, \infty)$ so that $\theta = \gamma/(\gamma - 1)$ is fixed. Use (1) and observe that a coalition M with $|M| \geq 2$ is internally stable if and only if

$$\begin{aligned} u_{in}^{|M|} \geq u_{out}^{|M|-1} &\iff \frac{1}{\theta}|M|^\theta \geq (|M| - 1)^\theta + \frac{1}{\theta} \\ &\iff 0 \geq \Gamma(|M|, \theta). \end{aligned}$$

On the other hand, a coalition M with $|M| \leq n - 1$ is externally stable if and only if

$$\begin{aligned} u_{out}^{|M|} > u_{in}^{|M|+1} &\iff |M|^\theta + \frac{1}{\theta} \geq \frac{1}{\theta}(|M| + 1)^\theta \\ &\iff \Gamma(|M| + 1, \theta) > 0. \end{aligned}$$

Therefore, by defining m_* by (A.1), we conclude that M is stable if and only if $|M| = m_*$. Since $x(\theta)$ is unique by Lemma A.1, so is m_*. Also, since $x(\theta)$ is independent of c, so is m_*. Moreover, Lemma A.5 shows that m_* is weakly increasing in $\theta \in (1, \infty)$ with $\lim_{\theta \to 1} m_* = 2$, $\lim_{\theta \to \infty} m_* = n$. This result means that m_* is weakly decreasing in $\gamma \in (1, \infty)$ with $\lim_{\gamma \to \infty} m_* = 2$ and $\lim_{\gamma \to 1} m_* = n$, as claimed in the remark. Finally, Lemma A.5 also shows that $x(2) = 3$ and $2 < x(\theta) < 3$ for all $\theta \in (1, 2)$, which means that $m_* = 3$ when $\theta = 2$ (or when $\gamma = 2$) whereas $m_* = 2$ when $\theta < 2$ (or when $\gamma > 2$). □

A.2 Proof of Remark 2

To prove the 'if' part, put $m_* := \lceil 1/c \rceil$ and fix M such that $|M| = m_*$.

Since $1 < 1/c \leq m_* < 1/c + 1$, we have

$$u_{in}^{|M|} - u_{out}^{|M|-1} = cm_* - 1 \geq 0,$$

meaning that M is internally stable. If $m_* = n$, there is no outsiders of M and we do not have to check its external stability. If $m_* < n$, the external stability condition is satisfied because

$$u_{out}^{|M|} - u_{in}^{|M|+1} = 1 - c > 0.$$

Hence, we conclude that M is stable.

To prove the 'only if' part, let M be a stable coalition. By (3), M cannot be internally stable if $|M| \geq 1/c+1$ and M cannot be externally stable if $|M| < 1/c$. Hence, either $M = n$ with $|M| < 1/c + 1$ or $M \neq n$ with $1/c \leq |M| < 1/c + 1$. Since $1/c < n$, we must have $1/c \leq |M| < 1/c + 1$ for both cases and therefore $|M| = \lceil 1/c \rceil = m_*$. This completes the proof.

A.3 Proof of Proposition 3.1

We first prove the following lemma, which states that all stable coalitions have either m_* or l^* members.

Lemma A.6. *Given the strategy profile* (15)*, M satisfies* (8) *only if* $|M| \in \{m_*, l^*\}$.

Proof. Let M be a coalition that satisfies (8) and assume that agents use the strategies (15). First, suppose $|M| \leq l^* - 1$. Then, the participation decision of

a single member does not change the continuation value:

$$V_i(M \cup \{i\}) = V_i(M \setminus \{i\}) \quad \forall i \in M$$

because strategy profile (15) instructs members to abandon the coalition in the following period. Therefore, the internal stability required in (8) implies

$$u_{in}^{|M|} \geq u_{out}^{|M|-1},$$

which, by (14), implies $|M| \leq m_*$. But $|M| < m_*$ is impossible because the external stability in (8) implies

$$u_{in}^{|M|+1} < u_{out}^{|M|},$$

which, by (13), requires $|M| \geq m_*$. Therefore, $|M| = m_*$ is the only possibility if $|M| \leq l^* - 1$.

We next show that $|M|$ cannot be greater than l^*. To confirm this claim, suppose to the contrary that for M satisfying (8), $|M| \geq l^* + 1$. Then, under strategy profile (15), which instructs all agents to remain even if one agent defects from the coalition,

$$V_i(M \cup \{i\}) = \frac{1}{1-\delta} u_{in}^{|M|} \quad \text{and} \quad V_i(M \setminus \{i\}) = \frac{1}{1-\delta} u_{out}^{|M|-1} \quad \forall i \in M.$$

The hypothesis $|M| \geq l^* + 1$, the fact that $l^* + 1 > m_*$, and (13), imply that $V_i(M \setminus \{i\}) > V_i(M \cup \{i\})$. This inequality and the internal stability required in (8) imply that

$$u_{in}^{|M|} \geq u_{out}^{|M|-1},$$

which, by (14), is possible only if $|M| \leq m_*$. But this contradicts the hypothesis $|M| \geq l^* + 1$ and the fact that $l^* > m_*$. Therefore, $|M| \leq l^*$.

It follows that under strategy profile (15), necessary conditions for stability are that either $|M| = m_*$ or $|M| = l^*$. □

Proof. (Proposition 3.1) We first prove the 'if' part of the proposition. Suppose that the discount factor δ satisfies

$$\delta < \delta_{l^*} := \frac{u_{out}^{l^*-1} - u_{in}^{l^*}}{u_{out}^{l^*-1} - \bar{u}^{m_*}} \in (0, 1]. \tag{A.9}$$

Let $(\pi_M)_{M \in \mathcal{M}}$ and $(a_i)_{i \in N}$ be defined as in Proposition 3.1. Then, given $(\pi_M)_{M \in \mathcal{M}}$

and $(a_i)_{i\in N}$, the value functions defined by

$$V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta}u_i(M_{-1}) & \text{if } |M_{-1}| \geq l^* \\ \frac{1}{1-\delta}\bar{u}^{m_*} & \text{otherwise} \end{cases}$$

satisfy (12). Since the support of the common belief only includes coalitions with $m_* < l^*$ members,

$$\begin{aligned} \mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] &= \mathbb{E}_\pi\left[u_i(\tilde{M})\right] + \delta\mathbb{E}_\pi\left[V_i(\tilde{M})\right] \\ &= \bar{u}^{m_*} + \delta\frac{1}{1-\delta}\bar{u}^{m_*} \\ &= \frac{1}{1-\delta}\bar{u}^{m_*}. \end{aligned} \tag{A.10}$$

The last two equalities imply that for any M_{-1}

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \iff u_i(M_{-1}) \geq \bar{u}^{m_*} \tag{A.11}$$

because if $|M_{-1}| \geq l^*$,

$$\begin{aligned} & u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \\ \iff & u_i(M_{-1}) + \frac{\delta}{1-\delta}u_i(M_{-1}) \geq \frac{1}{1-\delta}\bar{u}^{m_*} \\ \iff & u_i(M_{-1}) \geq \bar{u}^{m_*} \end{aligned}$$

and if $|M_{-1}| < l^*$,

$$\begin{aligned} & u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \\ \iff & u_i(M_{-1}) + \frac{\delta}{1-\delta}\bar{u}_i^{m_*} \geq \frac{1}{1-\delta}\bar{u}^{m_*} \\ \iff & u_i(M_{-1}) \geq \bar{u}^{m_*}. \end{aligned}$$

Notice that for $i \in M_{-1}$, $u_i(M_{-1}) = u_{in}^{|M_{-1}|}$, so by the definition of l^*

$$u_i(M_{-1}) \geq \bar{u}^{m_*} \iff u_{in}^{|M_{-1}|} \geq \bar{u}^{m_*} \iff |M_{-1}| \geq l^*. \tag{A.12}$$

In addition, for $i \notin M_{-1}$, where $u_i(M_{-1}) = u_{out}^{|M_{-1}|}$,

$$u_i(M_{-1}) \geq \bar{u}^{m_*} \iff u_{out}^{|M_{-1}|} \geq \bar{u}^{m_*} \iff |M_{-1}| \geq m_*. \tag{A.13}$$

One can confirm the last equivalence in (A.13) by observing

$$\begin{aligned} |M_{-1}| \geq m_* &\implies u_{out}^{|M_{-1}|} \geq u_{out}^{m_*} > u_{in}^{m_*} \\ &\implies u_{out}^{|M_{-1}|} > \frac{m_*}{n} u_{in}^{m_*} + \left(1 - \frac{m_*}{n}\right) u_{out}^{m_*} = \bar{u}^{m_*}, \end{aligned}$$

where we use Assumption 1-(a), and

$$\begin{aligned} |M_{-1}| < m_* &\implies u_{out}^{|M_{-1}|} \leq u_{in}^{|M_{-1}|+1} \leq u_{in}^{m_*} < u_{out}^{m_*} \\ &\implies u_{out}^{|M_{-1}|} < \frac{m_*}{n} u_{in}^{m_*} + \left(1 - \frac{m_*}{n}\right) u_{out}^{m_*} = \bar{u}^{m_*}, \end{aligned}$$

where we use (13) and Assumption 1-(a) and (d). Hence, it follows from (A.11), (A.12), and (A.13) that given $(\pi_M)_{M \in \mathcal{M}}$ and $(V_i)_{i \in N}$, the policy functions $(a_i)_{i \in N}$ defined by (15) do indeed satisfy (11).

To complete the proof of the 'if' part, we next show that given $(V_i)_{i \in N}$, M satisfies (8) if and only if $|M| = m_*$. There are two cases to consider. Consider first the case where $l^* = m_* + 1$. Let M be a coalition with $|M| = m_*$. Then for each $i \in M$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{m_*} + \frac{\delta}{1-\delta} \bar{u}_i^{m_*} \\ &\geq u_{out}^{m_*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned}$$

where the inequality follows from the definition of m_*. Therefore, the coalition M is internally stable. We now establish that this coalition is also externally stable. For each $i \notin M$, because (by hypothesis) $m_* + 1 = l^*$, we have

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &< u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M) + \delta V_i(M), \end{aligned}$$

where the inequality is due to (A.9). Therefore, the coalition M is externally stable. We conclude that if $l^* = m_* + 1$, coalitions of size m_* satisfy (8).

We need to prove that none of the other coalitions (i.e., those with $|M| \neq m_*$) satisfy (8). Because Lemma A.6 states that a coalition is stable only if its size is m_* or l^*, we need only show that coalitions of size l^* do not satisfy (8). In

fact, coalitions of size l^* are not internally stable because for each $i \in M$ with $|M| = l^*$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &< u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned}$$

where the inequality is again implied by (A.9).

Consider the other case where $l^* > m_* + 1$, where the definition of m_* directly implies that coalitions of size m_* satisfy (8). Also, exactly the same argument as in the first case shows that coalitions of size l^* are not internally stable. Hence, together with Lemma A.6, we conclude that M satisfies (8) if and only if $|M| = m_*$. This completes the proof of the 'if' part.

To prove the 'only if' part, suppose that $\delta \geq \delta_{l^*}$. We shall show that the common belief $(\pi_M)_{M \in \mathcal{M}}$ and the policy functions $(a_i)_{i \in N}$ defined in Proposition 3.1 do not constitute an equilibrium. In particular, we claim that coalitions of size l^* satisfy (8) if $\delta \geq \delta_{l^*}$. First, coalitions of size l^* are internally stable because for each $i \in M$ with $|M| = l^*$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &\geq u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned}$$

where the inequality follows from $\delta \geq \delta_{l^*}$. Also, coalitions of size l^* are externally stable because for each $i \notin M$ with $|M| = l^*$,

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{l^*+1} + \frac{\delta}{1-\delta} u_{in}^{l^*+1} \\ &< u_{out}^{l^*} + \frac{\delta}{1-\delta} u_{out}^{l^*} \\ &= u_i(M) + \delta V_i(M), \end{aligned}$$

where the inequality is due to (13) and the fact that $l^* \geq m_*$. However, the stability of l^* is inconsistent with the common belief defined in Proposition 3.1, which presumes that only coalitions of size m_* satisfy (8). □

A.4 Proof of Proposition 3.2

We begin with a roadmap of the proof. We first show that part (a) implies part (b). To this end, we verify that strategy (16) in part (a) constitutes an equilibrium only if (19) holds. We then show that this equation holds only if inequality (17) holds. We then show that part (b) imnplies part (a). To this end, we take as given a probability in the interval defined by (19) and we assume that δ satisfies (17). We then show that the equilibrium strategy satisfies (16).

Proof. To prove that statement a) in the proposition implies statement b), suppose that for some $(\pi_M)_{M\in\mathcal{M}}$, the strategy (16) in part (a), $(a_i)_{i\in N}$, constitutes an equilibrium. With $(\pi_M)_{M\in\mathcal{M}}$ given, denote as π^{m^*} the probability that any coalition of size m^* is drawn from the distribution, namely, $\pi^{m^*} := \sum_{|M|=m^*} \pi_M$. Obviously, π^{m^*} must satisfy $\pi^{m^*} > 0$. Also π^{m^*} must satisfy $1 > \pi^{m^*}$ because otherwise coalitions of size m_* would not be in the support.

Under the strategy profile $(a_i)_{i\in N}$, every player sticks with the coalition they inherit whenever its size is at least as large as m^*. For smaller inherited coalitions, members defect, initiating a new round of negotiation that results in either a coalition of size m^* or of size m_*, with probability π^{m^*} and $1-\pi^{m^*}$, respectively. Hence, the value functions $(V_i)_{i\in N}$ satisfy the recursion

$$V_i(M_{-1}) = \begin{cases} u_i(M_{-1}) + \delta V_i(M_{-1}) & \text{if } |M_{-1}| \geq m^* \\ \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] & \text{otherwise.} \end{cases} \tag{A.14}$$

Solving the first line of this equation yields

$$V_i(M_{-1}) = \frac{1}{1-\delta} u_i(M_{-1}) \tag{A.15}$$

for any M_{-1} with $|M_{-1}| \geq m^*$. Therefore

$$\mathbb{E}_\pi \left[V_i(\tilde{M}) \middle| |\tilde{M}| = m^* \right] = \frac{1}{1-\delta} \mathbb{E}_\pi \left[u_i(\tilde{M}) \middle| |\tilde{M}| = m^* \right] = \frac{1}{1-\delta} \bar{u}^{m^*}. \tag{A.16}$$

Note that the second line on the right side of (A.14) is independent of M_{-1} for $|M_{-1}| < m^*$. Because $m_* < m^*$, it follows that

$$\begin{aligned} \mathbb{E}_\pi \left[V_i(\tilde{M}) \middle| |\tilde{M}| = m_* \right] &= \mathbb{E}_\pi \left[\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \middle| |\tilde{M}| = m_* \right] \\ &= \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]. \end{aligned} \tag{A.17}$$

Combining (A.16) and (A.17), we obtain

$$\begin{aligned}
\mathbb{E}_\pi\left[u_i(\tilde{M})+\delta V_i(\tilde{M})\right] &= \mathbb{E}_\pi\left[u_i(\tilde{M})+\delta V_i(\tilde{M})\middle||\tilde{M}|=m^*\right]\pi^{m^*} \\
&\quad + \mathbb{E}_\pi\left[u_i(\tilde{M})+\delta V_i(\tilde{M})\middle||\tilde{M}|=m_*\right]\left(1-\pi^{m^*}\right) \\
&= \left(\bar{u}^{m^*}+\frac{\delta}{1-\delta}\bar{u}^{m^*}\right)\pi^{m^*} \\
&\quad + \left(\bar{u}^{m_*}+\delta\mathbb{E}_\pi\left[V_i(\tilde{M})\middle||\tilde{M}|=m_*\right]\right)\left(1-\pi^{m^*}\right) \\
&= \frac{1}{1-\delta}\bar{u}^{m^*}\pi^{m^*}+\bar{u}^{m_*}\left(1-\pi^{m^*}\right) \\
&\quad + \delta\mathbb{E}_\pi\left[u_i(\tilde{M})+\delta V_i(\tilde{M})\right]\left(1-\pi^{m^*}\right).
\end{aligned}$$

We set the first and last expressions in this string of equalities equal to each other and solve for $\mathbb{E}_\pi[V_i(\tilde{M})]$. Using this expression, we have

$$\mathbb{E}_\pi\left[u_i(\tilde{M})+\delta V_i(\tilde{M})\right]=\frac{1}{1-\delta}\bar{u}^\pi, \tag{A.18}$$

where we define

$$\bar{u}^\pi := \bar{u}^{m^*}\frac{\pi^{m^*}}{1-\delta(1-\pi^{m^*})}+\bar{u}^{m_*}\left(1-\frac{\pi^{m^*}}{1-\delta(1-\pi^{m^*})}\right). \tag{A.19}$$

This $\bar{u}^\pi$ represents players' expected per-period payoff if they reopen the negotiation process.

Internal stability for a coalition of size m^* requires

$$u_i(M)+\delta V_i(M)\geq u_i(M\setminus\{i\})+\delta V_i(M\setminus\{i\})\quad\forall i\in M,$$

for M with $|M|=m^*$. Using (A.14), (A.15), and (A.18), this inequality can be written as

$$\frac{1}{1-\delta}u_{in}^{m^*}\geq u_{out}^{m^*-1}+\delta\frac{1}{1-\delta}\bar{u}^\pi. \tag{A.20}$$

Rearranging terms yields the upper bound of $\Pi_\delta^{m^*}$ in (19):

$$\pi^{m^*}\leq\frac{\delta-\frac{u_{out}^{m^*-1}-u_{in}^{m^*}}{u_{out}^{m^*-1}-\bar{u}^{m_*}}}{\delta+\frac{\delta}{1-\delta}\frac{\bar{u}^{m^*}-u_{in}^{m^*}}{u_{out}^{m^*-1}-\bar{u}^{m_*}}}. \tag{A.21}$$

We note that the right-hand side of this inequality is smaller than 1 because

$$1 > \frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}} \iff \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} > 0$$

$$\iff \delta\bar{u}^{m^*} + (1-\delta)u_{out}^{m^*-1} - u_{in}^{m^*} > 0, \tag{A.22}$$

where we use the fact that $m^* \geq l^* > m_*$ and therefore $\bar{u}^{m^*} \geq u_{in}^{m^*}$ and $u_{out}^{m^*-1} > u_{in}^{m^*} \geq u_{in}^{l^*} \geq \bar{u}^{m_*}$. The inequality in the second line of (A.22) always holds because $\bar{u}^{m^*} \geq u_{in}^{m^*}$ and $u_{out}^{m^*-1} > u_{in}^{m^*}$.

Moreover, for the proposed strategy profile to constitute an equilibrium, members of an inherited coalition must prefer reopening the negotiation if the size of the inherited coalition is smaller than m^*. Hence, it must be the case that

$$\mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] > u_i(M_{-1}) + \delta V_i(M_{-1}) \quad \forall i \in M_{-1} \tag{A.23}$$

whenever $|M_{-1}| < m^*$. Using (A.14) and (A.18), inequality (A.23) can be written as

$$\frac{1}{1-\delta}\bar{u}^\pi > u_{in}^{|M_{-1}|} + \delta\frac{1}{1-\delta}\bar{u}^\pi.$$

This inequality must hold when $|M_{-1}| = m^* - 1$ in particular, implying

$$\bar{u}^\pi > u_{in}^{m^*-1}, \tag{A.24}$$

which by (A.19) is equivalent to the lower bound of $\Pi_\delta^{m^*}$ in (19) (when this exceeds 0):

$$\pi^{m^*} > \frac{(1-\delta)\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}{\bar{u}^{m^*} - \bar{u}^{m_*} - \delta\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}. \tag{A.25}$$

We note that because $\bar{u}^{m^*} > u_{in}^{m^*-1}$, the right-hand side of this inequality is negative if and only if $u_{in}^{m^*-1} < \bar{u}^{m_*}$, which is the case for $m^* = l^*$. For any $m^* > l^*$, the right-hand side is non-negative.

Combining (A.21) and (A.25) yields

$$\frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}} \geq \pi^{m^*} > \frac{(1-\delta)\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}{\bar{u}^{m^*} - \bar{u}^{m_*} - \delta\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}. \tag{A.26}$$

Because π^{m^*} satisfies $1 > \pi^{m^*} > 0$, (A.26) requires

$$\frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}} > \max\left\{0, \frac{(1-\delta)\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}{\bar{u}^{m^*} - \bar{u}^{m_*} - \delta\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}\right\}. \tag{A.27}$$

We now show that (A.27) is equivalent to

$$\delta > \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \max\{\bar{u}^{m_*}, u_{in}^{m^*-1}\}} = \delta_{m^*}, \tag{A.28}$$

thus proving that statement a) implies statement b). We need to consider two cases.

CASE 1: Consider first the case where $m^* = l^*$. By definition of l^*, we have $u_{in}^{l^*-1} < \bar{u}^{m_*}$ and therefore the right-hand side of (A.27) is 0. Hence, (A.27) is equivalent to

$$\delta > \frac{u_{out}^{l^*-1} - u_{in}^{l^*}}{u_{out}^{l^*-1} - \bar{u}^{m_*}},$$

which coincides with (A.28) because $\max\{\bar{u}^{m_*}, u_{in}^{l^*-1}\} = \bar{u}^{m_*}$.

CASE 2: Next consider the case where $l^* < m^* \leq n$; here, the left-hand side of (A.27) is non-negative. Define functions $\overline{\pi}^{m^*}(\delta)$ and $\underline{\pi}^{m^*}(\delta)$ as

$$\overline{\pi}^{m^*}(\delta) := \frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}$$

and

$$\underline{\pi}^{m^*}(\delta) := \frac{(1-\delta)\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}{\bar{u}^{m^*} - \bar{u}^{m_*} - \delta\left(u_{in}^{m^*-1} - \bar{u}^{m_*}\right)}.$$

We claim that

$$\overline{\pi}^{m^*}(\delta) > \underline{\pi}^{m^*}(\delta) \iff \delta > \delta_{m^*}. \tag{A.29}$$

To prove this claim, define

$$\alpha_{m^*} := \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} \in (0,1), \quad \beta_{m^*} := \frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} \geq 0,$$

and

$$\eta_{m^*} := \frac{u_{in}^{m^*-1} - \bar{u}^{m_*}}{\bar{u}^{m^*} - \bar{u}^{m_*}} \in [0,1),$$

so that we can write

$$\overline{\pi}^{m^*}(\delta) = \frac{\delta - \alpha_{m^*}}{\delta + \frac{\delta}{1-\delta}\beta_{m^*}} \quad \text{and} \quad \underline{\pi}^{m^*}(\delta) = \frac{(1-\delta)\eta_{m^*}}{1 - \delta\eta_{m^*}}.$$

We note that $\beta_{m^*} = 0$ for $m^* = n$ because $\bar{u}^n = u_{in}^n$. Otherwise, β_{m^*} is strictly positive. Observe that

$$\begin{aligned}
\overline{\pi}^{m^*}(\delta) - \underline{\pi}^{m^*}(\delta) = 0 &\iff \frac{(1-\delta)\eta_{m^*}}{1 - \delta\eta_{m^*}} = \frac{\delta - \alpha_{m^*}}{\delta + \frac{\delta}{1-\delta}\beta_{m^*}} \\
&\iff \delta = \frac{\alpha_{m^*}}{1 - \eta_{m^*}\left(\beta_{m^*} + 1 - \alpha_{m^*}\right)} \\
&\iff \delta = \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - u_{in}^{m^*-1}} \\
&\iff \delta = \delta_{m^*},
\end{aligned}$$

where the last line uses the fact that $m^* > l^* > m_*$ and therefore $\max\{\bar{u}^{m_*}, u_{in}^{m^*-1}\} = u_{in}^{m^*-1}$. Hence, $\delta = \delta_{m^*} \in (0,1)$ is the unique root of the equation $\overline{\pi}^{m^*}(\delta) - \underline{\pi}^{m^*}(\delta) = 0$. Also observe

$$\lim_{\delta \to 0} \overline{\pi}^{m^*}(\delta) = -\infty < 0 \leq \eta_{m^*} = \lim_{\delta \to 0} \underline{\pi}^{m^*}(\delta),$$

which means that $\overline{\pi}^{m^*}(\delta) - \underline{\pi}^{m^*}(\delta) < 0$ for small δ. This fact and the fact that $\delta = \delta_{m^*}$ is the unique root of $\overline{\pi}^{m^*}(\delta) - \underline{\pi}^{m^*}(\delta) = 0$, imply (A.29).

Statement a) of the proposition requires that (A.27) be true, and we have shown that (A.27) is equivalent to $\delta > \delta_{m^*}$, which is statement b). Therefore, we conclude that statement a) implies statement b).

We now prove the converse: statement b) in the proposition implies statement a). Suppose that $\delta > \delta_{m^*}$ and construct an equilibrium combination of belief and strategy as follows. First, let $\Pi_\delta^{m^*} \subset (0,1)$ be the interval defined as (19):

$$\Pi_\delta^{m^*} = \Big(\max\{0, \underline{\pi}^{m^*}(\delta)\}, \overline{\pi}^{m^*}(\delta) \Big].$$

We have already shown that $\Pi_\delta^{m^*}$ is nonempty if and only if $\delta > \delta_{m^*}$. Therefore, we can choose $\pi^{m^*} \in \Pi_\delta^{m^*}$ and let $(\pi_M)_{M \in \mathcal{M}}$ be the belief defined as (18). Let $(a_i)_{i \in N}$ be the strategy profile defined as (16) where we choose $k^* \in \{m_*, \ldots, n-1\}$ such that

$$u_{out}^{k^*} \geq \bar{u}^\pi > u_{out}^{k^*-1}, \tag{A.30}$$

where $\bar{u}^\pi$ is defined in (A.19). Outsiders want to stick with an inherited coali-

tion having k^* members but they prefer to reopen negotiations if the inherited coalition has k^*-1 members. Under Assumption 1, such a k^* always exists and is unique because

$$u_{out}^{n-1} > u_{in}^{n} = \bar{u}^{n} > \bar{u}^{\pi} > \bar{u}^{m_*} > u_{in}^{m_*} \geq u_{out}^{m_*-1} \tag{A.31}$$

and u_{out}^{m} is strictly increasing in $m \geq m_* - 1$. By (A.31), the first inequality in (A.30) is satisfied at $k^* = n-1$ and the second inequality is satisfied at $k^* = m_*$. Choose k^* as the smallest integer (which of course is unique) that satisfies the first inequality in (A.30); then k^*-1 also satisfies the second inequality. We can now show that the belief $(\pi_M)_{M\in\mathcal{M}}$ and the strategy $(a_i)_{i\in N}$ in (16) constitute an equilibrium.

We know from the discussion above (in particular, (A.14), (A.15), and (A.18)) that with this combination of belief and strategy, the associated value functions $(V_i)_{i\in N}$ are given by

$$V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_i(M_{-1}) & \text{if } |M_{-1}| \geq m^* \\ \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] = \frac{1}{1-\delta}\bar{u}^{\pi} & \text{otherwise.} \end{cases} \tag{A.32}$$

In what follows, we show that with $(V_i)_{i\in N}$ and $(\pi_M)_{M\in\mathcal{M}}$ given, M satisfies (8) (i.e., M is stable) if and only if $|M| \in \{m_*, m^*\}$ and $(a_i)_{i\in N}$ solves (11). These two requirements are necessary and sufficient for $(\pi_M)_{M\in\mathcal{M}}$ and $(a_i)_{i\in N}$ to constitute an equilibrium.

We first show that M satisfies (8) if and only if $|M| \in \{m_*, m^*\}$. The 'only if' part follows from exactly the same argument as in the proof of Lemma A.6. To prove the 'if' part, take M such that $|M| = m_*$. Because $m_* < m^* - 1$, (A.32) implies that

$$V_i(M \cup \{i\}) = V_i(M \setminus \{i\}) = \frac{1}{1-\delta}\bar{u}^{\pi}$$

for all $i \in N$.[19] Hence, in this case, it follows from the definition of m_* that M satisfies (8). Now take M such that $|M| = m^*$. Then M is internally stable

[19]This part requires $m^* \neq m_* + 1$. If $m^* = m_* + 1$, coalitions with m_* members will not be externally stable.

because for each $i \in M$,

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{m^*} + \frac{\delta}{1-\delta} u_{in}^{m^*} \\ &\geq u_{out}^{m^*-1} + \frac{\delta}{1-\delta} \bar{u}^{\pi} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned}$$

where the inequality follows from the fact that $\pi^{m^*} \leq \overline{\pi}^{m^*}(\delta)$. (See also the equivalence between (A.20) and (A.21).) Also, M is externally stable because for each $i \notin M$,

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{m^*+1} + \frac{\delta}{1-\delta} u_{in}^{m^*+1} \\ &< u_{out}^{m^*} + \frac{\delta}{1-\delta} u_{out}^{m^*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned}$$

where the inequality follows from the fact that $m^* > m_*$. Hence, M satisfies (8). We conclude that M satisfies (8) if and only if $|M| \in \{m_*, m^*\}$.

To complete the proof, we need to show that $(a_i)_{i \in N}$ solves (11) given $(V_i)_{\in N}$ and $(\pi_M)_{M \in \mathcal{M}}$. Fix M_{-1} and first consider an arbitrary member $i \in M_{-1}$. If this player sticks with M_{-1}, she obtains the payoff

$$u_i(M_{-1}) + \delta V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_{in}^{|M_{-1}|} & \text{if } |M_{-1}| \geq m^* \\ u_{in}^{|M_{-1}|} + \frac{\delta}{1-\delta} \bar{u}^{\pi} & \text{if } |M_{-1}| < m^*. \end{cases} \tag{A.33}$$

If she abandons M_{-1}, she obtains the payoff

$$\mathbb{E}_{\pi}\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] = \frac{1}{1-\delta} \bar{u}^{\pi}. \tag{A.34}$$

Combining (A.33) and (A.34) implies the equivalence

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_{\pi}\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \iff u_{in}^{|M_{-1}|} \geq \bar{u}^{\pi}$$

for $i \in M_{-1}$. This equivalence is true regardless of the choice of M_{-1}. Therefore, the strategy profile defined by (16) is optimal for members of any existing coalition if

$$u_{in}^{m^*} \geq \bar{u}^{\pi} > u_{in}^{m^*-1}. \tag{A.35}$$

The first inequality states that members of an existing coalition prefer sticking with the coalition whenever it has at least m^* members. The second inequality states that they would rather reopen the negotiation if the existing coalition is smaller than m^*. We need to show that (A.35) in fact holds. Because $\pi^{m^*} \leq \overline{\pi}^{m^*}(\delta)$, it follows from the equivalence between (A.20) and (A.21) that

$$u_{in}^{m^*} \geq (1-\delta)u_{out}^{m^*-1} + \delta\bar{u}^{\pi}. \tag{A.36}$$

Because $u_{out}^{m^*-1} > u_{in}^{m^*}$, we then have

$$u_{out}^{m^*-1} > (1-\delta)u_{out}^{m^*-1} + \delta\bar{u}^{\pi}$$

and therefore

$$u_{out}^{m^*-1} > \bar{u}^{\pi}. \tag{A.37}$$

Combining (A.36) and (A.37) yields

$$u_{in}^{m^*} \geq (1-\delta)u_{out}^{m^*-1} + \delta\bar{u}^{\pi} > \bar{u}^{\pi},$$

which proves the first inequality in (A.35). The second inequality in (A.35) directly follows from the fact that $\pi^{m^*} > \underline{\pi}^{m^*}(\delta)$ and the equivalence between (A.24) and (A.25). We have therefore proved that the strategy profile defined by (16) is optimal for members of M_{-1}.

Next consider an arbitrary nonmember $i \notin M_{-1}$. If this player sticks with M_{-1}, she obtains the payoff

$$u_i(M_{-1}) + \delta V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta}u_{out}^{|M_{-1}|} & \text{if } |M_{-1}| \geq m^* \\ u_{out}^{|M_{-1}|} + \frac{\delta}{1-\delta}\bar{u}^{\pi} & \text{if } |M_{-1}| < m^*. \end{cases} \tag{A.38}$$

If instead she defects, triggering a new round of negotiation, her payoff is

$$\mathbb{E}_{\pi}\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] = \frac{1}{1-\delta}\bar{u}^{\pi}. \tag{A.39}$$

Combining (A.38) and (A.39) implies the equivalence

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_{\pi}\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \iff u_{out}^{|M_{-1}|} \geq \bar{u}^{\pi}$$

for $i \notin M_{-1}$. This equivalence is true regardless of the choice of M_{-1}. Therefore, the strategy profile defined by (16) is optimal for nonmembers of any existing

coalition if

$$u_{out}^{k^*} \geq \bar{u}^{\pi} > u_{out}^{k^*-1}. \tag{A.40}$$

The interpretation of these inequalities is analogous to that of (A.35). By construction of k^*, (A.40) in fact holds. Therefore, the strategy profile defined by (16) is also optimal for nonmembers of M_{-1}. □

A.5 Proof of Proposition 3.3

As above, π denotes a symmetric equilibrium belief and $\mathcal{M}$ its support; $(a_i)_{i\in N}$ and $(V_i)_{i\in N}$ are the equilibrium policy functions and the value functions, respectively. We begin with a roadmap of the proof. We first use the assumptions that beliefs and reduced form payoffs are symmetric (Definition 2.2 and Assumption 1) to show that the expected payoff from reopening the negotiation process must be the same for all countries (Lemmas A.7 and A.8). Then we show that coalitions with fewer than m_* members cannot be included in $\mathcal{M}$ (Lemmas A.9 and A.10). We also show that any coalition in $\mathcal{M}$ with more than m_* members must be sustainable and any defection from such a coalition must make it unsustainable (Lemma A.11). It follows that if $\mathcal{M}$ contains coalitions of three or more distinct sizes, we can find $M, M' \in \mathcal{M}$ such that $|M| > |M'| > m_*$ and M' is sustainable but $M \setminus \{i\}$ is not, for any $i \in M$, in spite of the fact that $M \setminus \{i\}$ is not smaller than M'. This observation, together with the inequalities derived in Lemma A.12, causes a contradiction.

Lemma A.7. *For any $M, M' \in \mathcal{N}$ with $|M| = |M'|$, if M satisfies $a_i(M) = 1$ for all $i \in N$, so does M'.*

Proof. Fix $M, M' \in \mathcal{N}$ such that $|M| = |M'|$ and suppose that M satisfies $a_i(M) = 1$ for all $i \in N$. Once M is formed, players keep using it so we have

$$V_i(M) = u_i(M) + \delta V_i(M) \quad \forall i \in N,$$

which implies

$$V_i(M) = \frac{1}{1-\delta} u_i(M) \quad \forall i \in N. \tag{A.41}$$

Because $a_i(M) = 1$ for each $i \in N$, it must be the case that

$$u_i(M) + \delta V_i(M) \geq \mathbb{E}_{\pi}\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \quad \forall i \in N. \tag{A.42}$$

Combining (A.41) and (A.42), we have

$$\frac{1}{1-\delta}u_i(M) \geq \mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] \quad \forall i \in N,$$

which under the symmetry of the reduced-form payoff functions implies

$$\frac{1}{1-\delta}\min\left\{u_{in}^{|M|}, u_{out}^{|M|}\right\} \geq \max_{i\in N}\left\{\mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right]\right\}. \tag{A.43}$$

Now suppose that $a_{i'}(M') = 0$ for some $i' \in N$. We establish the Lemma by falsifying this hypothesis. Under the hypothesis, when M' is inherited from the preceding period, player i' strictly prefers reopening the negotiation process. So we must have

$$u_{i'}(M') + \delta V_{i'}(M') < \mathbb{E}_\pi\left[u_{i'}(\tilde{M}) + \delta V_{i'}(\tilde{M})\right],$$

where (because $\prod_{j\in N} a_j(M') = 0$)

$$V_{i'}(M') = \mathbb{E}_\pi\left[u_{i'}(\tilde{M}) + \delta V_{i'}(\tilde{M})\right],$$

implying

$$\frac{1}{1-\delta}u_{i'}(M') < \mathbb{E}_\pi\left[u_{i'}(\tilde{M}) + \delta V_{i'}(\tilde{M})\right].$$

Under the symmetry of the reduced-form payoff functions, this inequality implies

$$\frac{1}{1-\delta}\min\left\{u_{in}^{|M'|}, u_{out}^{|M'|}\right\} < \mathbb{E}_\pi\left[u_{i'}(\tilde{M}) + \delta V_{i'}(\tilde{M})\right]. \tag{A.44}$$

Because $|M| = |M'|$, combining (A.43) and (A.44) yields

$$\max_{i\in N}\left\{\mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right]\right\} < \mathbb{E}_\pi\left[u_{i'}(\tilde{M}) + \delta V_{i'}(\tilde{M})\right],$$

a contradiction. □

Lemma A.8. *The expected payoff from reopening the negotiation process is identical for all players, namely,*

$$\mathbb{E}_\pi\left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] = \mathbb{E}_\pi\left[u_j(\tilde{M}) + \delta V_j(\tilde{M})\right] \quad \forall i, j \in N. \tag{A.45}$$

Proof. Let $\mathcal{L}$ be the set of all sustainable coalitions, namely,

$$\mathcal{L} := \{M \in \mathcal{N} \,|\, a_i(M) = 1 \forall i \in N\}. \tag{A.46}$$

Then we may write

$$M \in \mathcal{L} \implies V_i(M) = u_i(M) + \delta V_i(M) = \frac{1}{1-\delta} u_i(M) \quad \forall i \in N \tag{A.47}$$

and

$$\begin{aligned} M \notin \mathcal{L} \implies V_i(M) &= \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \\ &= \mathbb{E}_\pi \left[u_i(\tilde{M}) \right] + \delta \mathbb{E}_\pi \left[V_i(\tilde{M}) \right] \quad \forall i \in N. \end{aligned} \tag{A.48}$$

Combining (A.47) and (A.48), we obtain

$$\begin{aligned} \mathbb{E}_\pi \left[V_i(\tilde{M}) \right] &= \mathbb{E}_\pi \left[V_i(\tilde{M}) \middle| \tilde{M} \in \mathcal{L} \right] \pi^{\mathcal{L}} + \mathbb{E}_\pi \left[V_i(\tilde{M}) \middle| \tilde{M} \notin \mathcal{L} \right] \left(1 - \pi^{\mathcal{L}} \right) \\ &= \frac{1}{1-\delta} \mathbb{E}_\pi \left[u_i(\tilde{M}) \middle| \tilde{M} \in \mathcal{L} \right] \pi^{\mathcal{L}} \\ &\quad + \left(\mathbb{E}_\pi \left[u_i(\tilde{M}) \right] + \delta \mathbb{E}_\pi \left[V_i(\tilde{M}) \right] \right) \left(1 - \pi^{\mathcal{L}} \right) \quad \forall i \in N, \end{aligned} \tag{A.49}$$

where $\pi^{\mathcal{L}} \in [0,1]$ denotes the probability of drawing a sustainable coalition under the equilibrium belief,

$$\pi^{\mathcal{L}} := \sum_{M \in \mathcal{L}} \pi_M. \tag{A.50}$$

Solving (A.49) for $\mathbb{E}_\pi \left[V_i(\tilde{M}) \right]$ yields

$$\mathbb{E}_\pi \left[V_i(\tilde{M}) \right] = \frac{1}{1-\delta} \left(\frac{\pi^{\mathcal{L}}}{1 - \delta(1-\pi^{\mathcal{L}})} \mathbb{E}_\pi \left[u_i(\tilde{M}) \middle| \tilde{M} \in \mathcal{L} \right] + \frac{(1-\delta)(1-\pi^{\mathcal{L}})}{1 - \delta(1-\pi^{\mathcal{L}})} \mathbb{E}_\pi \left[u_i(\tilde{M}) \right] \right),$$

which implies

$$\begin{aligned} \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] &= \mathbb{E}_\pi \left[u_i(\tilde{M}) \right] + \delta \mathbb{E}_\pi \left[V_i(\tilde{M}) \right] \\ &= \frac{1}{1 - \delta(1-\pi^{\mathcal{L}})} \mathbb{E}_\pi \left[u_i(\tilde{M}) \right] \\ &\quad + \frac{\pi^{\mathcal{L}}}{1 - \delta(1-\pi^{\mathcal{L}})} \frac{\delta}{1-\delta} \mathbb{E}_\pi \left[u_i(\tilde{M}) \middle| \tilde{M} \in \mathcal{L} \right] \end{aligned} \tag{A.51}$$

for all $i \in N$. Note that by assumption the reduced-form payoff functions $(u_i)_{i \in N}$ are symmetric across players and so is the equilibrium belief π, also by assumption. Moreover, by Lemma A.7, the set $\mathcal{L}$ treats players symmetrically. Therefore, we conclude that the right-hand side of (A.51) is independent of i, which completes the proof. □

The following lemma states that if a stable coalition has fewer than m_* members, then that coalition is sustainable, but the addition of an additional member renders it not sustainable. We use this intermediate result in the subsequent lemma to show that there are no stable coalitions with fewer than m_* members.

Lemma A.9. *If $M \in \mathcal{M}$ and $|M| < m_*$, then M is sustainable but $M \cup \{i\}$ is not sustainable for any $i \in N \setminus M$.*

Proof. Fix $M \in \mathcal{M}$ such that $|M| < m_*$. Because M is externally stable,

$$u_i(M) + \delta V_i(M) > u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \quad \forall i \in N \setminus M. \tag{A.52}$$

By (13), $|M| < m_*$ implies that there exists $i' \in N \setminus M$ such that

$$u_{i'}(M) \leq u_{i'}(M \cup \{i'\}),$$

which by the assumed symmetry of the reduced-form payoff functions implies

$$u_i(M) < u_i(M \cup \{i\}) \quad \forall i \in N \setminus M. \tag{A.53}$$

Combining (A.53) and (A.52) yields

$$V_i(M) > V_i(M \cup \{i\}) \quad \forall i \in N \setminus M. \tag{A.54}$$

Now choose arbitrary $i \in N \setminus M$ arbitrarily. It follows from (A.54) that at either M or $M \cup \{i\}$ is sustainable; if this were not the case then $V_i(M) = V_i(M \cup \{i\}) = \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]$, contradicting (A.54). Also, M and $M \cup \{i\}$ cannot both be sustainable because otherwise (A.53) implies

$$V_i(M) = \frac{1}{1-\delta} u_i(M) \leq \frac{1}{1-\delta} u_i(M \cup \{i\}) = V_i(M \cup \{i\}),$$

which contradicts (A.54). Thus, to complete the proof we need only show that $M \cup \{i\}$ is not sustainable. Suppose to the contrary that $M \cup \{i\}$ is sustainable (which implies M is not sustainable). Then

$$V_i(M \cup \{i\}) = u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \tag{A.55}$$

and

$$V_i(M) = \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]. \tag{A.56}$$

Because $M \cup \{i\}$ is sustainable, we must have $a_i(M \cup \{i\}) = 1$, implying

$$u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \tag{A.57}$$

Combining (A.55)–(A.57) yields

$$\begin{aligned} V_i(M \cup \{i\}) &= u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \\ &\geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \\ &= V_i(M), \end{aligned}$$

which again contradicts (A.54). This completes the proof. □

Lemma A.10. *If $M \in \mathcal{M}$, it must be the case that $|M| \geq m_*$.*

Proof. Suppose to the contrary that there exists $M \in \mathcal{M}$ such that $|M| < m_*$. We know from Lemma A.9 that M is sustainable. Hence, $a_i(M) = 1$ for all $i \in N$, which implies

$$u_i(M) + \delta V_i(M) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]$$

with

$$V_i(M) = \frac{1}{1-\delta} u_i(M)$$

for all $i \in N$. Combining these expressions for $i \in M$ implies

$$\frac{1}{1-\delta} u_{in}^{|M|} \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]. \tag{A.58}$$

Fix $i' \in N \setminus M$ so that by Lemma A.9 $M \cup \{i'\}$ is not sustainable. Then there must exist $j \in N$ such that $a_j(M \cup \{i'\}) = 0$, which implies

$$\mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right] > u_j(M \cup \{i'\}) + \delta V_j(M \cup \{i'\})$$

with

$$V_j(M \cup \{i'\}) = \mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right].$$

Combining these expressions implies

$$\frac{1}{1-\delta} u_j(M \cup \{i'\}) < \mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right]. \tag{A.59}$$

By Lemma A.8, we know that the right-hand sides of (A.58) and (A.59) are

identical. Hence, under Assumption 1-a) and -d), combining (A.58) and (A.59) yields

$$u_{in}^{|M|+1} \leq u_j(M \cup \{i'\}) < u_{in}^{|M|} \leq u_{in}^{|M|+1},$$

a contradiction. Therefore, we conclude that any coalition in $\mathcal{M}$ must have at least m_* members in it. □

Lemma A.11. *If $M \in \mathcal{M}$ and $|M| > m_*$, then M is sustainable but $M \setminus \{i\}$ is not sustainable for any $i \in M$.*

Proof. The proof is analogous to the proof of Lemma A.9. Fix $M \in \mathcal{M}$ such that $|M| > m_*$. Because M is internally stable,

$$u_i(M) + \delta V_i(M) \geq u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}) \quad \forall i \in M. \tag{A.60}$$

By (14), $|M| > m_*$ implies that there exists $i' \in M$ such that

$$u_{i'}(M) < u_{i'}(M \setminus \{i'\}),$$

which by the assumed symmetry of the reduced-form payoff functions implies

$$u_i(M) < u_i(M \setminus \{i\}) \quad \forall i \in M. \tag{A.61}$$

Combining (A.61) and (A.60) yields

$$V_i(M) > V_i(M \setminus \{i\}) \quad \forall i \in M. \tag{A.62}$$

Now choose arbitrary $i \in M$. It follows from (A.62) that either M or $M \setminus \{i\}$ is sustainable; if this were not true, then $V_i(M) = V_i(M \setminus \{i\}) = \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]$, contradicting (A.62). Also, M and $M \setminus \{i\}$ cannot both be sustainable because otherwise (A.61) implies

$$V_i(M) = \frac{1}{1-\delta} u_i(M) < \frac{1}{1-\delta} u_i(M \setminus \{i\}) = V_i(M \setminus \{i\}),$$

which contradicts (A.62). Thus, to complete the argument we need only show that $M \setminus \{i\}$ is not sustainable. Suppose to the contrary that $M \setminus \{i\}$ is sustainable (which implies that M is not sustainable). Then

$$V_i(M \setminus \{i\}) = u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}) \tag{A.63}$$

and

$$V_i(M) = \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]. \tag{A.64}$$

Because $M \setminus \{i\}$ is sustainable, we must have $a_i(M \setminus \{i\}) = 1$, implying

$$u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right]. \tag{A.65}$$

Combining (A.63)–(A.65) yields

$$\begin{aligned} V_i(M \setminus \{i\}) &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}) \\ &\geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \\ &= V_i(M), \end{aligned}$$

which again contradicts (A.62). This completes the proof. □

Lemma A.12. *If $M \in \mathcal{M}$ and $|M| \neq m_*$, it must be the case that*

$$u_{in}^{|M|} \geq (1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] > u_{in}^{|M|-1} \quad \forall i \in N. \tag{A.66}$$

Proof. The proof is analogous to the proof of Lemma A.10. Fix $M \in \mathcal{M}$ such that $|M| \neq m_*$. By Lemma A.10, we know that $|M| > m_*$. Then it follows from Lemma A.11 that M is sustainable. Hence, $a_i(M) = 1$ for all $i \in N$, which implies

$$u_i(M) + \delta V_i(M) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \tag{A.67}$$

with

$$V_i(M) = \frac{1}{1-\delta} u_i(M) \tag{A.68}$$

for all $i \in N$. Combining these expressions for $i \in M$ implies

$$u_{in}^{|M|} \geq (1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \quad \forall i \in M. \tag{A.69}$$

Noticing that by Lemma A.8 the right-hand side of this inequality is identical for all players establishes the first inequality in (A.66).

To derive the second inequality in (A.66), fix $i' \in M$. Lemma A.11 shows that $M \setminus \{i'\}$ is not sustainable. Hence there must exist $j \in N$ such that $a_j(M \setminus \{i'\}) = 0$, which implies

$$\mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right] > u_j(M \setminus \{i'\}) + \delta V_j(M \setminus \{i'\}) \tag{A.70}$$

with

$$V_j(M \setminus \{i'\}) = \mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right]. \tag{A.71}$$

Combining these expressions implies

$$\frac{1}{1-\delta} u_j(M \setminus \{i'\}) < \mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right]. \tag{A.72}$$

Under Assumption 1-a) and -d), (A.72) implies

$$u_{in}^{|M|-1} \leq u_j(M \setminus \{i'\}) < \mathbb{E}_\pi \left[u_j(\tilde{M}) + \delta V_j(\tilde{M}) \right], \tag{A.73}$$

Again, by Lemma A.8, the right-hand side of the last inequality is identical for all players, which establishes the second inequality in (A.66). □

Proof. (Proposition 3.3) Suppose that the support $\mathcal{M}$ of the symmetric equilibrium belief contains coalitions of three or more distinct sizes. Then we can choose $M, M' \in \mathcal{M}$ such that $|M| \neq |M'|$ and neither of them is of size m_*. Assume that $|M| > |M'|$ without loss of generality.

Fix $i \in N$ arbitrarily. By Lemma A.12, we have

$$u_{in}^{|M|} \geq (1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] > u_{in}^{|M|-1} \tag{A.74}$$

and

$$u_{in}^{|M'|} \geq (1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] > u_{in}^{|M'|-1}. \tag{A.75}$$

Because $|M| - 1 \geq |M'|$, Assumption 1-a) implies

$$u_{in}^{|M|-1} \geq u_{in}^{|M'|}. \tag{A.76}$$

Combining (A.74)–(A.76) yields

$$(1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] > (1-\delta)\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right], \tag{A.77}$$

a contradiction. Therefore we conclude that the support of any symmetric equilibrium belief cannot contain coalitions of three or more distinct sizes. □

A.6 Proof of Proposition 3.4

Proof. As in the proof of Remark 1, define $\theta = \gamma/(\gamma-1)$. Using (1), it is easy to see that

$$\frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{c^\theta} = (m^*-1)^\theta - \frac{1}{\theta}(m^*)^\theta + \frac{1}{\theta},$$

and

$$\frac{u_{out}^{m^*-1} - u_{in}^{m^*-1}}{c^\theta} = \frac{\theta-1}{\theta}\left((m^*-1)^\theta - 1\right).$$

Because $m^* > l^*$, we have $\max\{\bar{u}^{m_*}, u_{in}^{m^*-1}\} = u_{in}^{m^*-1}$. Therefore, using (17)

$$\begin{aligned}\delta_{m^*} &= \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - u_{in}^{m^*-1}} \\ &= \frac{\frac{\theta}{\theta-1}(m^*-1)^\theta - \frac{1}{\theta-1}\left((m^*)^\theta - 1\right)}{(m^*-1)^\theta - 1}\end{aligned}$$

as claimed in the proposition. To prove that δ_{m^*} is increasing in m^*, note that

$$\frac{\partial \delta_{m^*}}{\partial m^*}\frac{1}{\delta_{m^*}} = \frac{\theta^2(m^*-1)^{\theta-1} - \theta(m^*)^{\theta-1}}{\theta(m^*-1)^\theta - ((m^*)^\theta - 1)} - \frac{\theta(m^*-1)^{\theta-1}}{(m^*-1)^\theta - 1},$$

which is positive if and only if

$$(m^*)^{\theta-1} - \theta > 1 - \left(\frac{m^*}{m^*-1}\right)^{\theta-1}. \tag{A.78}$$

Because $\theta > 1$, the right-hand side of (A.78) is negative for any $m^* \geq 2$. The left-hand side of (A.78) is an increasing function of m^* and it is easy to verify that it is positive at $m^* = e$. Therefore, (A.78) holds if $m^* \geq e$. By Remark 1, we know that $m_* \geq 2$ for all $\theta > 1$. The fact that $m^* > m_* + 1$ implies $m^* > e$, so we conclude that for all $\theta > 1$ δ_{m^*} is increasing in m^*. □

A.7 Proof of Proposition 3.5

Proof. Because $m^* > m_*$, using (3) yields

$$u_{out}^{m^*-1} - u_{in}^{m^*} = 1 - c \quad \text{and} \quad u_{out}^{m^*-1} - u_{in}^{m^*-1} = 1,$$

from which we obtain

$$\delta_{m^*} = \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - u_{in}^{m^*-1}} = 1 - c.$$

□

A.8 Proof of Proposition 3.6

To make this proof self-contained, we repeat some definitions used in the proof of Proposition 3.2:

$$\overline{\pi}^{m^*}(\delta) = \frac{\delta - \alpha_{m^*}}{\delta + \frac{\delta}{1-\delta}\beta_{m^*}} \quad \text{and} \quad \underline{\pi}^{m^*}(\delta) = \frac{(1-\delta)\eta_{m^*}}{1 - \delta\eta_{m^*}}. \tag{A.79}$$

$$\alpha_{m^*} := \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} \in (0,1), \quad \beta_{m^*} := \frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} \geq 0, \tag{A.80}$$

and

$$\eta_{m^*} := \frac{u_{in}^{m^*-1} - \bar{u}^{m_*}}{\bar{u}^{m^*} - \bar{u}^{m_*}} \in [0,1). \tag{A.81}$$

With these definitions, we write (19) as

$$\Pi_\delta^{m^*} = \Big(\max\{0, \underline{\pi}^{m^*}(\delta)\}, \overline{\pi}^{m^*}(\delta) \Big].$$

Proof. (Proposition 3.6-(a)) We note that $\beta_{m^*} = 0$ for $m^* = n$ because $\bar{u}^n = u_{in}^n$. Otherwise, β_{m^*} is strictly positive. Hence,

$$\lim_{\delta \to 1} \overline{\pi}^{m^*}(\delta) = \begin{cases} 0 & \text{if } m^* < n \\ 1 - \alpha_n = \frac{u_{in}^n - \bar{u}^{m_*}}{u_{out}^{n-1} - \bar{u}^{m_*}} > 0 & \text{if } m^* = n, \end{cases}$$

and

$$\lim_{\delta \to 1} \underline{\pi}^{m^*}(\delta) = 0.$$

Therefore,

$$\lim_{\delta \to 1} \left(\max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \right) = 0$$

for any $m^* \neq n$, whereas

$$\lim_{\delta \to 1} \left(\max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \right) = \frac{u_{in}^n - \bar{u}^{m_*}}{u_{out}^{n-1} - \bar{u}^{m_*}} > 0.$$

for $m^* = n$. It follows that there exists $\delta^* \in (0,1)$ such that

$$\delta > \delta^* \implies \max \Pi_\delta^n - \inf \Pi_\delta^n > \max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \quad \forall m^* \neq n,$$

which proves statement (a) of the proposition. □

Proof. (Proposition 3.6-(b)) First observe in (A.79) that $\overline{\pi}^{m^*}(\delta)$ is decreasing in α_{m^*} and β_{m^*} and $\overline{\pi}^{m^*}(\delta)$ is increasing in η_{m^*}. Hence for the statement (b) of the proposition to be true, it suffices to show that α_{m^*} and β_{m^*} are both decreasing in m^* and η_{m^*} is increasing in m^*.

For Example 2, where $m_* = \lceil 1/c \rceil \geq 1/c$, we know from (3) that

$$u_{in}^{m^*} = -c(n - m^*), \quad u_{out}^{m^*} = 1 - c(n - m^*),$$

and

$$\bar{u}^{m^*} = \frac{m^*}{n} u_{in}^{m^*} + \frac{n - m^*}{n} u_{out}^{m^*} = \frac{n - m^*}{n} - c(n - m^*) \tag{A.82}$$

for any $m^* \geq m_*$ and therefore

$$\alpha_{m^*} = \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} = \frac{n - cn}{cn(m^* - 1 - m_*) + m_*},$$

$$\beta_{m^*} = \frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}} = \frac{n - m^*}{cn(m^* - 1 - m_*) + m_*},$$

$$\eta_{m^*} = \frac{u_{in}^{m^*-1} - \bar{u}^{m_*}}{\bar{u}^{m^*} - \bar{u}^{m_*}} = \frac{cn - \frac{n - m_* + cn}{m^* - m_*}}{cn - 1}$$

for any $m^* \geq m_* + 1$. A brief inspection of these expressions should reveal that α_{m^*} and β_{m^*} are both decreasing in m^* and η_{m^*} is increasing in m^*, as desired. □

A.9 Proof of Proposition 4.1

Let $(\pi, (a_i)_{i \in N})$ be an equilibrium of reduced-form model $\langle \delta, N, (u_i^\infty)_{i \in N} \rangle$, where

$$u_i^\infty(M) = \phi_i(\hat{\boldsymbol{g}}^\infty(M)) - \frac{c}{1 - \delta\sigma} f(\hat{\boldsymbol{g}}^\infty(M)). \tag{A.83}$$

Then, by definition, there exist value functions $(V_i)_{i \in N}$ such that $\mathcal{M}$ is the collection of all $M \in \mathcal{N}$ satisfying

$$i \in M \iff u_i^\infty(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \geq u_i^\infty(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \tag{A.84}$$

the policy functions $(a_i)_{i\in N}$ satisfy

$$a_i(M_{-1}) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ \left[u_i^\infty(M_{-1}) + \delta V_i(M_{-1})\right] a_i + \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta V_i(\tilde{M})\right](1-a_i) \Bigg\}, \tag{A.85}$$

and the value functions $(V_i)_{i\in N}$ solve

$$V_i(M_{-1}) = \begin{cases} u_i^\infty(M_{-1}) + \delta V_i(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta V_i(\tilde{M})\right] & \text{otherwise.} \end{cases} \tag{A.86}$$

Now define functions $(V_i^\infty)_{i\in N}$ by

$$V_i^\infty(M_{-1}, G_{-1}) := V_i(M_{-1}) - \frac{c}{1-\delta\sigma}\sigma G_{-1}.$$

Given (A.83), (A.84), (A.85), (A.86), (26), and (27), it is straightforward to see that $(\pi, (a_i)_{i\in N}, (\hat{g}_i^\infty)_{i\in N})$ satisfies Definition 4.1 as an equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$ with $(V_i^\infty)_{i\in N}$ being the value functions associated with the structural model.

A.10 Proof of Proposition 4.2

We first prove the following lemma.

Lemma A.13. *Under Assumptions 2 and 3, if $(\pi^1, (a_i^1)_{i\in N}, (g_i^1)_{i\in N})$ is an equilibrium of the structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, 1\rangle$, then the support $\mathcal{M}^1$ of the belief and a_i^1 are both independent of G_{-1} and*

$$g_i^1(M, G_{-1}, 1) = \hat{g}_i^1(M), \tag{A.87}$$

where $\hat{g}_i^1$ is defined in Assumption 3. The value function associated with this model is given by

$$V_i^1(M_{-1}, G_{-1}, 1) = v_i^1(M_{-1}) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma}\sigma G_{-1}$$

for some function v_i^1.

Proof. Since $(g_i^1(M, G_{-1}, 1))_{i\in N}$ is the equilibrium profile of emission levels cho-

sen by players, it must simultaneously satisfy

$$(g_i^1(M, G_{-1}, 1))_{i\in M} \in \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \textstyle\sum_{i\in M} \Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) \quad \text{s.t.} \quad g_j = g_j^1(M, G_{-1}, 1) \quad \forall j \notin M,$$

and

$$g_i^1(M, G_{-1}, 1) \in \underset{g_i}{\operatorname{argmax}} \Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) \quad \text{s.t.} \quad g_j = g_j^1(M, G_{-1}, 1) \quad \forall j \in N \setminus \{i\} \qquad \forall i \notin M$$

for each $M \in \mathcal{N}$. By Assumption 2, we may write

$$\begin{aligned}\underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) &= \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \{\phi_i(\boldsymbol{g}) - c\,[\sigma G_{-1} + f(\boldsymbol{g})]\} \\ &= \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{ \phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma} f(\boldsymbol{g}) \right\}\end{aligned}$$

and

$$\begin{aligned}\underset{g_i}{\operatorname{argmax}}\, \Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) &= \underset{g_i}{\operatorname{argmax}}\, \{\phi_i(\boldsymbol{g}) - c\,[\sigma G_{-1} + f(\boldsymbol{g})]\} \\ &= \underset{g_i}{\operatorname{argmax}} \left\{ \phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma} f(\boldsymbol{g}) \right\}.\end{aligned}$$

Hence, by Assumption 3, we have $(g_i^1(M, G_{-1}, 1))_{i\in N} = (\hat{g}_i^1(M))_{i\in N}$. Notice in particular that $g_i^1(M, G_{-1}, 1)$ is independent of G_{-1}. With this result, we may characterize $\mathcal{M}^1$ as the collection of all M such that

$$\begin{aligned} i \in M &\iff \phi_i(\boldsymbol{g}^1(M \cup \{i\}, G_{-1}, 1)) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(M \cup \{i\}, G_{-1}, 1))\right] \\ &\qquad \geq \phi_i(\boldsymbol{g}^1(M \setminus \{i\}, G_{-1}, 1)) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(M \setminus \{i\}, G_{-1}, 1))\right] \\ &\iff \phi_i(\hat{\boldsymbol{g}}^1(M \cup \{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^1(M \cup \{i\}))\right] \\ &\qquad \geq \phi_i(\hat{\boldsymbol{g}}^1(M \setminus \{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^1(M \setminus \{i\}))\right] \\ &\iff u_i^1(M \cup \{i\}) \geq u_i^1(M \setminus \{i\}), \end{aligned} \tag{A.88}$$

where we define

$$u_i^1(M) := \phi_i(\hat{\boldsymbol{g}}^1(M)) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma} f(\hat{\boldsymbol{g}}^1(M)).$$

Since the last line of (A.88) is independent of G_{-1}, we conclude that $\mathcal{M}^1$ is independent of G_{-1}. The policy functions $(a_i^1)_{i\in N}$ are also independent of G_{-1} because they must solve

$$\begin{aligned}
a_i^1(M_{-1}, G_{-1}, 1) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left[\Phi_i(\boldsymbol{g}^1(M_{-1}, G_{-1}, 1), F(\boldsymbol{g}^1(M_{-1}, G_{-1}, 1), G_{-1}))\right] a_i \\
& + \mathbb{E}_\pi \left[\Phi_i(\boldsymbol{g}^1(\tilde{M}, G_{-1}, 1), F(\boldsymbol{g}^1(\tilde{M}, G_{-1}, 1), G_{-1}))\right] (1 - a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left[\Phi_i(\hat{\boldsymbol{g}}^1(M_{-1}), F(\hat{\boldsymbol{g}}^1(M_{-1}), G_{-1}))\right] a_i \\
& + \mathbb{E}_\pi \left[\Phi_i(\hat{\boldsymbol{g}}^1(\tilde{M}), F(\hat{\boldsymbol{g}}^1(\tilde{M}), G_{-1}))\right] (1 - a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left[\phi_i(\hat{\boldsymbol{g}}^1(M_{-1})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^1(M_{-1}))\right]\right] a_i \\
& + \mathbb{E}_\pi \left[\phi_i(\hat{\boldsymbol{g}}^1(\tilde{M})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^1(\tilde{M}))\right]\right] (1 - a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & u_i^1(M_{-1}) a_i + \mathbb{E}_\pi \left[u_i^1(\tilde{M})\right] (1 - a_i) \Bigg\}.
\end{aligned}$$

Finally, it is easy to see that the associated value functions $(V_i^1)_{i\in N}$ are given by

$$V_i^1(M_{-1}, G_{-1}) = v_i^1(M_{-1}) - c\frac{1 - (\delta\sigma)^1}{1 - \delta\sigma}\sigma G_{-1},$$

where

$$v_i^1(M_{-1}) := \begin{cases} u_i^1(M_{-1}) & \text{if } \prod_{j\in N} a_j^1(M_{-1}, G_{-1}, 1) = 1 \\ \mathbb{E}_\pi \left[u_i^1(\tilde{M})\right] & \text{otherwise.} \end{cases}$$

This completes the proof. □

The next lemma generalizes Lemma A.13.

Lemma A.14. *Under Assumptions 2 and 3, for each $T < \infty$, if $(\pi^T, (a_i^T)_{i\in N}, (g_i^T)_{i\in N})$ is an equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, T\rangle$, then the support $\mathcal{M}^T$ of the belief and a_i^T are both independent of G_{-1} and*

$$g_i^T(M, G_{-1}, \tau) = \hat{g}_i^\tau(M) \tag{A.89}$$

for each $\tau \leq T$, where $\hat{g}_i^\tau$ is defined in Assumption 3. The value function

associated with this model is given by

$$V_i^T(M_{-1}, G_{-1}, \tau) = v_i^\tau(M_{-1}) - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma}\sigma G_{-1}$$

for some function v_i^τ for each $\tau \leq T$.

Proof. Suppose, as an induction hypothesis, that the statement is true for some $T < \infty$. Let $(\pi_M^{T+1}, (a_i^{T+1})_{i\in N}, (g_i^{T+1})_{i\in N})$ be an equilibrium of the $T+1$-period structural model. We shall show that the support $\mathcal{M}^{T+1}$ of the belief and a_i^{T+1} are both independent of G_{-1}, the policy function g_i^{T+1} satisfies

$$g_i^{T+1}(M, G_{-1}, \tau) = \hat{g}_i^\tau(M) \tag{A.90}$$

for each $\tau \leq T+1$, and the value function satisfies

$$V_i^{T+1}(M_{-1}, G_{-1}, \tau) = v_i^\tau(M_{-1}) - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma}\sigma G_{-1}. \tag{A.91}$$

for some function v_i^τ for each $\tau \leq T+1$. Note that by the induction hypothesis, (A.90) and (A.91) must be true for $\tau = 1, 2, \ldots, T$.

Since $(g_i^{T+1}(M, G_{-1}, T+1))_{i\in N}$ is the equilibrium profile of emission levels chosen by players, it must simultaneously satisfy

$$\begin{aligned}(g_i^{T+1}(M, G_{-1}, T+1))_{i\in M} \in \quad & \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \textstyle\sum_{i\in M} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^{T+1}(M, F(\boldsymbol{g}, G_{-1}), T)\right\} \\ & \text{s.t.} \quad g_j = g_j^{T+1}(M, G_{-1}, T+1) \quad \forall j \notin M,\end{aligned}$$

and

$$\begin{aligned}g_i^{T+1}(M, G_{-1}, T+1) \in \quad & \underset{g_i}{\operatorname{argmax}} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^{T+1}(M, F(\boldsymbol{g}, G_{-1}), T)\right\} \\ & \text{s.t.} \quad g_j = g_j^{T+1}(M, G_{-1}, T+1) \quad \forall j \in N \setminus \{i\}\end{aligned} \qquad \forall i \notin M$$

for each $M \in \mathcal{N}$. By Assumption 2 and the induction hypothesis, we may write

$$\begin{aligned}
&\underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^{T+1}(M, F(\boldsymbol{g}, G_{-1}), T)\right\} \\
&\quad = \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - cF(\boldsymbol{g}, G_{-1}) + \delta v_i^T(M_{-1}) - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\delta\sigma F(\boldsymbol{g}, G_{-1})\right\} \\
&\quad = \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}F(\boldsymbol{g}, G_{-1})\right\} \\
&\quad = \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\left[f(\boldsymbol{g}) + \sigma G_{-1}\right]\right\} \\
&\quad = \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}f(\boldsymbol{g})\right\}
\end{aligned}$$

and similarly

$$\begin{aligned}
&\underset{g_i}{\operatorname{argmax}} \left\{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i^{T+1}(M, F(\boldsymbol{g}, G_{-1}), T)\right\} \\
&\quad = \underset{g_i}{\operatorname{argmax}} \left\{\phi_i(\boldsymbol{g}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}f(\boldsymbol{g})\right\}.
\end{aligned}$$

Hence, by Assumption 3, we have $(g_i^{T+1}(M, G_{-1}), T+1)_{i\in N} = (g_i^{T+1}(M))_{i\in N}$. Notice in particular that $g_i^{T+1}(M, G_{-1}, T+1)$ is independent of G_{-1}.

With this result, we may characterize $\mathcal{M}^{T+1}$ as the collection of all M such that

$$\begin{aligned}
i \in M \iff\ & \phi_i(\boldsymbol{g}^{T+1}(M\cup\{i\}, G_{-1}, T+1)) - cF(\boldsymbol{g}^{T+1}(M\cup\{i\}, G_{-1}, T+1), G_{-1}) \\
&\quad + \delta V_i^{T+1}(M\cup\{i\}, F(\boldsymbol{g}^{T+1}(M\cup\{i\}, G_{-1}, T+1), G_{-1})) \\
&\geq \phi_i(\boldsymbol{g}^{T+1}(M\setminus\{i\}, G_{-1}, T+1)) - cF(\boldsymbol{g}^{T+1}(M\setminus\{i\}, G_{-1}, T+1), G_{-1}) \\
&\quad + \delta V_i^{T+1}(M\setminus\{i\}, F(\boldsymbol{g}^{T+1}(M\setminus\{i\}, G_{-1}, T+1), G_{-1})) \\
\iff\ & \phi_i(\hat{\boldsymbol{g}}^{T+1}(M\cup\{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^{T+1}(M\cup\{i\}))\right] \\
&\quad + \delta V_i^T(M\cup\{i\}, \sigma G_{-1} + f(\hat{\boldsymbol{g}}^{T+1}(M\cup\{i\}))) \\
&\geq \phi_i(\hat{\boldsymbol{g}}^{T+1}(M\setminus\{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^{T+1}(M\setminus\{i\}))\right] \\
&\quad + \delta V_i^T(M\setminus\{i\}, \sigma G_{-1} + f(\hat{\boldsymbol{g}}^{T+1}(M\setminus\{i\}))) \\
\iff\ & u_i^{T+1}(M\cup\{i\}) + \delta v_i^T(M\cup\{i\}) \\
&\geq u_i^{T+1}(M\setminus\{i\}) + \delta v_i^T(M\setminus\{i\}),
\end{aligned} \tag{A.92}$$

where

$$u_i^{T+1}(M) := \phi_i(\hat{\boldsymbol{g}}^{T+1}(M)) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma} f(\hat{\boldsymbol{g}}^{T+1}(M)).$$

Since the last line of (A.92) is independent of G_{-1}, we conclude that $\mathcal{M}^{T+1}$ is independent of G_{-1}. The policy functions $(a_i^{T+1})_{i\in N}$ are also independent of G_{-1}. First, by the induction hypothesis, $a_i^{T+1}(M_{-1}, G_{-1}, \tau)$ is independent of G_{-1} for all $\tau = 1, 2, \ldots, T$. Also, $a_i^{T+1}(M_{-1}, G_{-1}, T+1)$ must solve

$$\begin{aligned}
a_i^{T+1}(M_{-1}, G_{-1}, T+1) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left(\hat{\Phi}_i^{T+1}(M_{-1}, G_{-1}) + \delta \hat{V}_i^T(M_{-1}, G_{-1})\right) a_i \\
& + \mathbb{E}_\pi\left[\hat{\Phi}_i^{T+1}(\tilde{M}, G_{-1}) + \delta \hat{V}_i^T(\tilde{M}, G_{-1})\right](1-a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left(u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1}\right) a_i \\
& + \left(\mathbb{E}_\pi\left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M})\right] - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1}\right)(1-a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left[u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1})\right] a_i \\
& + \mathbb{E}_\pi\left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M})\right](1-a_i) \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Phi}_i^{T+1}(M, G_{-1}) &:= \Phi_i(\hat{\boldsymbol{g}}^{T+1}(M), F(\hat{\boldsymbol{g}}^{T+1}(M), G_{-1})) \\
&= \phi_i(\hat{\boldsymbol{g}}^{T+1}(M)) - cf(\hat{\boldsymbol{g}}^{T+1}(M)) - c\sigma G_{-1}
\end{aligned}$$

and

$$\begin{aligned}
\hat{V}_i^T(M, G_{-1}) &:= V_i^{T+1}(M, F(\hat{\boldsymbol{g}}^{T+1}(M), G_{-1}), T) \\
&= v_i^T(M) - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\sigma\left[f(\hat{\boldsymbol{g}}^{T+1}(M)) + \sigma G_{-1}\right].
\end{aligned}$$

Finally, we can compute the associated value functions $(V_i^{T+1})_{i\in N}$ as

$$V_i^{T+1}(M_{-1}, G_{-1}, \tau) = v_i^\tau(M_{-1}) - c\frac{1-(\delta\sigma)^\tau}{1-\delta\sigma}\sigma G_{-1}$$

for each $\tau \leq T+1$, where

$$v_i^{T+1}(M_{-1}) := \begin{cases} u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1}) & \text{if } \prod_{j\in N} a_j^{T+1}(M_{-1}, G_{-1}, T+1) = 1 \\ \mathbb{E}_{\pi^T}\left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M})\right] & \text{otherwise.} \end{cases}$$

Therefore, the statement of the lemma is true for $T+1$ as well. Together with Lemma A.13, the induction argument then completes the proof of the lemma. □

Proof. (Proposition 4.2) Let $(\pi^\infty, (a_i^\infty)_{i\in N}, (g_i^\infty)_{i\in N})$ be a limit equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$. Then, by Lemma A.14, the corresponding value functions $(V_i^\infty)_{i\in N}$ are given by

$$V_i^\infty(M_{-1}, G_{-1}) = \lim_{T\to\infty} V_i^T(M_{-1}, G_{-1}, T) = v_i^\infty(M_1) - \frac{c}{1-\delta\sigma}\sigma G_{-1} \qquad \text{(A.93)}$$

for some functions $(v_i^\infty)_{i\in N}$. Also, the support $\mathcal{M}^\infty$ of the belief and $(a_i^\infty)_{i\in N}$ are both independent of G_{-1} and the policy functions $(g_i^\infty)_{i\in N}$ coincide with $(\hat{g}_i^\infty)_{i\in N}$.

Since $(\pi_M^\infty, (a_i^\infty)_{i\in N}, (g_i^\infty)_{i\in N})$ is an equilibrium of $\langle \delta, N, (\Phi_i)_{i\in N}, F, \infty\rangle$, it satisfies (23), (24), and (25). It follows that $\mathcal{M}^\infty$ is the collection of all $M \in \mathcal{N}$ that satisfies

$$\begin{aligned} i \in M \iff & \phi_i(\hat{\boldsymbol{g}}^\infty(M\cup\{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^\infty(M\cup\{i\}))\right] \\ & \quad + \delta V_i^\infty(M\cup\{i\}, \sigma G_{-1} + f(\hat{\boldsymbol{g}}^\infty(M\cup\{i\}))) \\ & \geq \phi_i(\hat{\boldsymbol{g}}^\infty(M\setminus\{i\})) - c\left[\sigma G_{-1} + f(\hat{\boldsymbol{g}}^\infty(M\setminus\{i\}))\right] \\ & \quad + \delta V_i^\infty(M\setminus\{i\}, \sigma G_{-1} + f(\hat{\boldsymbol{g}}^\infty(M\setminus\{i\}))) \\ \iff & u_i^\infty(M\cup\{i\}) + \delta v_i^\infty(M\cup\{i\}) \\ & \geq u_i^\infty(M\setminus\{i\}) + \delta v_i^\infty(M\setminus\{i\}), \end{aligned}$$

where

$$u_i^\infty(M) = \phi_i(\hat{\boldsymbol{g}}^\infty(M)) - \frac{c}{1-\delta\sigma} f(\hat{\boldsymbol{g}}^\infty(M)).$$

Also, the policy functions $(a_i^\infty)_{i\in N}$ satisfy

$$\begin{aligned}
a_i^\infty(M_{-1}) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left(\hat{\Phi}_i^\infty(M_{-1},G_{-1}) + \delta\hat{V}^\infty(M_{-1},G_{-1})\right) a_i \\
& + \mathbb{E}_\pi\left[\hat{\Phi}_i^\infty(\tilde{M},G_{-1}) + \delta\hat{V}^\infty(\tilde{M},G_{-1})\right](1-a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left(u_i^\infty(M_{-1}) + \delta v_i^\infty(M_{-1}) - c\frac{1-(\delta\sigma)^\infty}{1-\delta\sigma}\sigma G_{-1}\right) a_i \\
& + \left(\mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta v_i^\infty(\tilde{M})\right] - c\frac{1-(\delta\sigma)^\infty}{1-\delta\sigma}\sigma G_{-1}\right)(1-a_i) \Bigg\} \\
= \underset{a_i\in\{0,1\}}{\operatorname{argmax}} \Bigg\{ & \left[u_i^\infty(M_{-1}) + \delta v_i^\infty(M_{-1})\right] a_i \\
& + \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta v_i^\infty(\tilde{M})\right](1-a_i) \Bigg\},
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Phi}_i^\infty(M,G_{-1}) &:= \Phi_i(\hat{\boldsymbol{g}}^\infty(M), F(\hat{\boldsymbol{g}}^\infty(M),G_{-1})) \\
&= \phi_i(\hat{\boldsymbol{g}}^\infty(M)) - cf(\hat{\boldsymbol{g}}^\infty(M)) - c\sigma G_{-1} \qquad \text{(A.94)}
\end{aligned}$$

and

$$\begin{aligned}
\hat{V}^\infty(M,G_{-1}) &:= V^\infty(M, F(\hat{\boldsymbol{g}}^\infty(M),G_{-1})) \\
&= v^\infty(M) - c\frac{1-(\delta\sigma)^\infty}{1-\delta\sigma}\sigma\left[f(\hat{\boldsymbol{g}}^\infty(M)) + \sigma G_{-1}\right]. \qquad \text{(A.95)}
\end{aligned}$$

Finally,

$$V_i^\infty(M_{-1},G_{-1}) = \begin{cases} \hat{\Phi}_i^\infty(M_{-1},G_{-1}) + \delta\hat{V}_i^\infty(M_{-1},G_{-1}) & \text{if } \prod_{j\in N} a_j^\infty(M_{-1},G_{-1}) = 1 \\ \mathbb{E}_\pi\left[\hat{\Phi}_i^\infty(\tilde{M},G_{-1}) + \delta\hat{V}^\infty(\tilde{M},G_{-1})\right] & \text{otherwise,} \end{cases}$$

which, together with (A.93), (A.94), and (A.95), implies

$$v_i^\infty(M_{-1}) = \begin{cases} u_i^\infty(M_{-1}) + \delta v_i^\infty(M_{-1}) & \text{if } \prod_{j\in N} a_j^\infty(M_{-1},G_{-1}) = 1 \\ \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta v_i^\infty(\tilde{M})\right] & \text{otherwise.} \end{cases}$$

Hence $(\pi^\infty, (a_i^\infty)_{i\in N})$, with the value functions $(v_i^\infty)_{i\in N}$, satisfies Definition 2.1

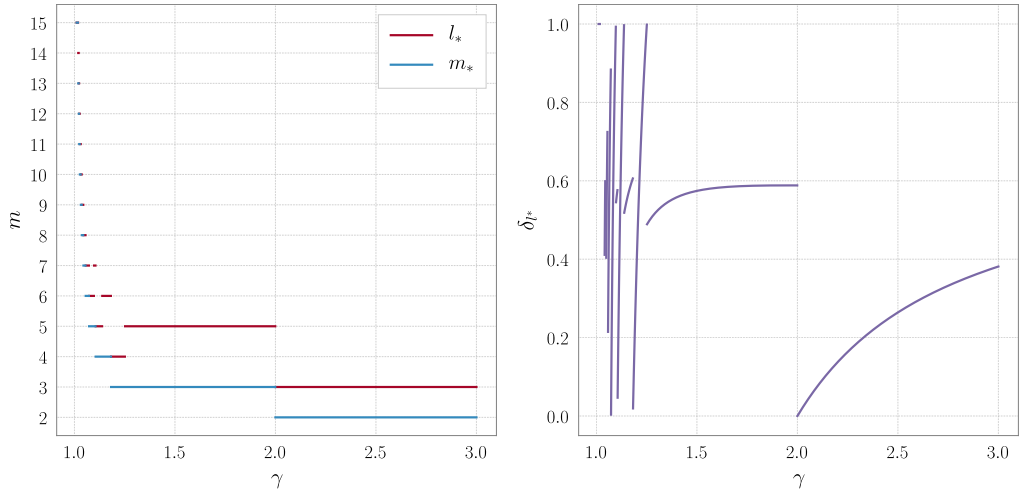

Figure 5: The values of m_* and l^* (left panel) and the threshold value δ_{l^*} of discount factor (right panel) in the model of Example 1. The number of players is set to $n = 15$.

as an equilibrium of reduced-form model $\langle \delta, N, (u_i^\infty)_{i \in N} \rangle$. □

B Numerical examples

Figure 5 illustrates Proposition 3.1 based on Example 1. As Remark 1 shows, the value of m_* quickly declines as γ increases, converging to $m_* = 2$ for all $\gamma > 2$. The equilibrium cut-off size, l^*, is equal to or slightly larger than m_* and for the most part follows the same pattern as m_*, although it is not monotonic in γ. Here, players are pessimistic about future negotiations and therefore willing to keep the coalition they inherit if it is slightly larger than m_*. But, provided that the initial ($t = 0$) coalition is smaller than l^* (and unless the discount factor is greater than the threshold value δ_{l^*}) players always inherit a coalition of size m_*. They repeatedly reopen the negotiation process.

The right panel of Figure 5 shows that δ_{l^*} as a function of γ changes discontinuously as m_* and l^* jump. With m_* and l^* being given, however, a larger value of γ always implies a larger value of δ_{l^*}, making it more likely that this type of pessimistic equilibrium emerges. Many papers use the quadratic model ($\gamma = 2$), where the stable coalition contains either two or three members, depending on the tie-breaking assumption. Figure 5 shows, for our tie-breaking assumption, that $m_* \in \{2, 3\}$ for $\gamma > 1.2$. Over this range, the pessimistic equilibrium, where all stable coalitions have m_* members, requires $\delta < 0.6$. Thus, although our dynamic model produces the pessimistic static result in some circumstances, a

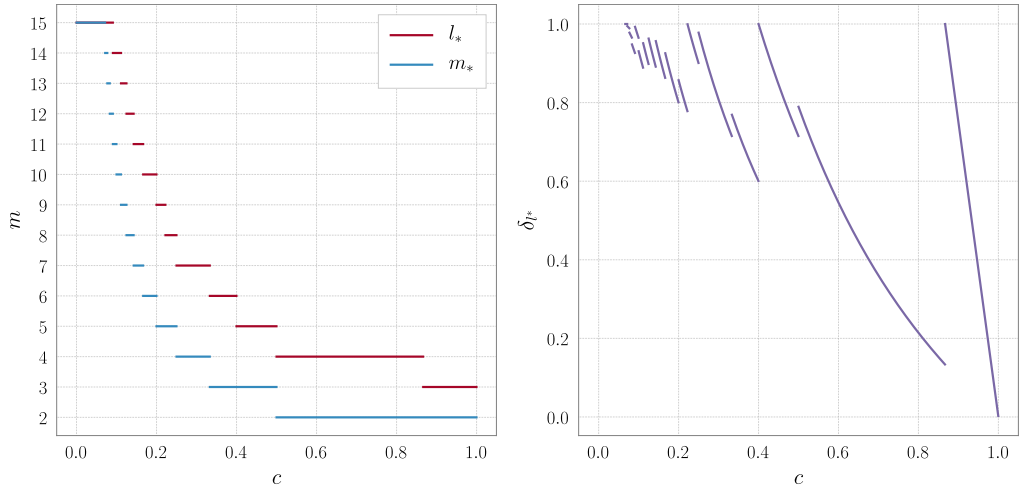

Figure 6: The values of m_* and l^* (left panel) and the threshold value δ_{l^*} of discount factor (right panel) in the model of Example 2. The number of players is set to $n = 15$.

moderate level of patience implies that, for the same γ, equilibrium beliefs always include larger coalitions.[20] The dynamic and static versions of the model therefore have quite different implications.

Example 2 suggests a slightly different relation, depicted in Figure 6. As Remark 2 shows, the value of m_* is small unless c, the marginal damage parameter, is also small. The cut-off size l^* closely follows the pattern of m_*, but the difference between the two is somewhat larger here than in Example 1. The right panel shows that the value of δ_{l^*} depends on c; the discontinuous points are due to discontinuity of m_* and l_*. Interestingly, here (unlike Example 1) with m_* and l_* given, a larger value of c always implies a smaller value of δ_{l^*}. Here, a larger marginal damage makes it less, not more, likely that this type of pessimistic equilibrium exists.

[20]For example, with an annual discount rate of 7% and a time step of five years, the per period discount factor is $\delta = 0.7$.