International environmental agreements without commitment*

Larry S. Karp† Hiroaki Sakamoto‡

February 14, 2018

Abstract

We analyze a dynamic model of international agreements where countries cannot make long-term commitments and have no sanctions or rewards to induce participation. Countries can communicate with each other to build endogenous beliefs about the random consequences of (re)opening negotiation. Provided that countries are patient, many different agreements might emerge, including an effective agreement with many participants. Along the way, however, negotiation might yield a succession of short-lived agreements with a small number of participants. Beliefs are important, and negotiations matter. Our theoretical results are consistent with the existing empirical observations and they explain the 'paradox of international agreements'.

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

---

*Acknowledgements.

†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

International environmental agreements (IEAs) attempt to correct trans-national externalities while respecting the principle of national sovereignty. Without supra-national enforcement mechanisms, IEAs must be largely self-enforcing. Many countries benefit from agreements that reduce global externalities, but only members incur the costs of environmental protection. The temptation to free-ride often leads to small or ineffectual agreements.

However, countries sometimes manage to cooperate; the number of IEAs has steadily grown. Mitchell's (2018) database 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 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).

We contribute to the theory of IEAs by explaining how a large and effective IEA can randomly emerge as the equilibrium to a non-cooperative game. Like most of the noncooperative IEA literature, the building block consists of the one-shot participation game in d'Aspremont et al. (1983) and Palfrey and Rosenthal (1984).[1] Much of this literature concludes that large and effective IEAs are unlikely to emerge in equilibrium, especially when the potential gains from cooperation are large (Hoel, 1992; Carraro and Siniscalco, 1993; Barrett, 1994). This pessimistic finding explains the difficulty of building a successful IEA when the stakes are high but it misses the considerable success documented above. Kolstad and Toman (2005) describe this discrepancy between theory and evidence as the 'paradox of international agreements'.

There are many explanations for the occasional IEA success. Costs of non-participation such as trade sanctions or social norms (Barrett, 1997; Hoel and Schneider, 1997), or money transfers among countries (Barrett, 2001; Carraro et al., 2006) can lead to large IEAs. Membership might be high if countries use mixed strategies at the participation stage, although even here the expected membership is small (Dixit and Olson, 2000; Hong and Karp, 2012, 2014). Increasing-return-to-scale technologies can result in large IEAs (Barrett, 2006). Even with convex technologies, relaxing the parametric assumptions in previous

[1] 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), and Barrett (2005) survey the literature.

papers can lead to large and effective IEAs (Karp and Simon, 2013).

Dynamic settings can also generate large IEAs. 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) show that when countries can commit to the number of periods during which an IEA is binding, a holdup problem can lead to a large and long-lived IEA. In a similar setting, Schmidt and Kovac (2015) demonstrate that with commitment but even without the holdup problem, large IEAs are possible when deviation triggers a costly delay of long-term agreements. (Section 5 discusses these issues more thoroughly.)

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

In our model as in the world, signatories can review and reject a previously signed international agreement (e.g. Canada and the Kyoto Protocol). Countries adhere to an agreement only when it serves their national self-interest. Abandonment of an existing agreement triggers a new round of negotiation.

We follow the literature in using 'stable equilibrium' as a synonym for 'Nash equilibrium'. We assume that every round of negotiation results in a stable equilibrium, but perhaps not the one that had just been abandoned or the one that ultimately prevails. The interim stable agreement immediately takes effect, eliminating any punishment phase. Some stable equilibria have low membership and low welfare, as in the standard one-shot models. Members have two types of incentives to deviate from such equilibria: their deviation might result in a large equilibrium and high welfare; but even if it leads to another small IEA, the erstwhile member might enjoy the free-riding benefit as a non-member of the new agreement. A 'sustainable equilibrium' is both stable in the conventional sense, and sufficiently attractive that no country wants to deviate. Members recognize that if they deviate they might become free-riders in a subsequent equilibrium. However, the likelihood that for a long period of time future IEAs will be small and ineffective discourages members of a sustainable IEA from defecting.

Our model has the flavor of real-world negotiation: it might be painstakingly long, with uncertain outcomes (Benedick, 1998; Oberthur and Ott, 1999).[2] The equilibrium in this game is an endogenous probability distribution over the size of the next-period IEA and the identify of their members. This probability distribution is not unique because it depends on countries' endogenous beliefs. The lack of uniqueness is consistent with the heterogeneity of actual agreements. The negotiation process affects agents' beliefs, and those beliefs help determine the outcome. Negotiations might not be successful, but *ex ante* they are not a waste of time. The model results have a practical implication as 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. By recognizing the importance of negotiations in formulating beliefs, and demonstrating the role of these beliefs in determining outcomes, the model can help reframe attitudes toward IEAs.

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 reduced-form period payoff function of player i where $\mathcal{N}$ is the set of all subsets of N. In every period, players collectively choose a coalition $M \in \mathcal{N}$ and obtain their individual payoff $u_i(M)$. Player i's discounted

[2]Benedick (1998) documents in great detail 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 even shaped by chance.

present-value payoff at period t is

$$\sum_{s=t}^{\infty} \delta^{s-t} u_i(M_s). \tag{1}$$

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. Consider the parametric payoff function of player i

$$-\frac{1}{\gamma}(\bar{g}_i - g_i)^{\gamma} - c\sum_{j\in N} g_j \tag{2}$$

for $\gamma > 1$, $c > 0$, 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

$$u_i(M) = \begin{cases} c^{\frac{\gamma}{\gamma-1}}\left(|M|^{\frac{\gamma}{\gamma-1}} - |M| + n - \frac{1}{\gamma}|M|^{\frac{\gamma}{\gamma-1}}\right) - c\sum_{j\in N}\bar{g}_j & \forall i \in M \\ c^{\frac{\gamma}{\gamma-1}}\left(|M|^{\frac{\gamma}{\gamma-1}} - |M| + n - \frac{1}{\gamma}\right) - c\sum_{j\in N}\bar{g}_j & \forall i \notin M \end{cases} \tag{3}$$

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. Another familiar model uses the payoff function

$$g_i - c\sum_{j\in N} g_j, \tag{4}$$

with $0 < c < 1$. Again, $g_i \in [0,1]$, is player i's pollution level. The reduced-form payoff function is

$$u_i(M) = \begin{cases} -c(n-|M|) & \forall i \in M \text{ if } |M| \geq 1/c \\ 1 - c(n-|M|) & \forall i \notin M \text{ if } |M| \geq 1/c \\ 1 - cn & \forall i \in N \text{ if } |M| < 1/c \end{cases} \tag{5}$$

for each $M \in \mathcal{N}$.

2.1 Static one-shot game

We review the one-shot game, a building block for the dynamic game. A Nash equilibrium of this participation game is a coalition $M \in \mathcal{N}$ such that[3]

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

Our tie-breaking assumption is that players join a coalition whenever they are indifferent between joining and not joining. Following the convention in 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) whereas the 'if' part implies that it is *externally stable* (non-members do not want to join).[4] The two examples above illustrate the static equilibrium.

Remark 1. In the model of Example 1, there exists a unique integer $m_* \geq 2$ such that M satisfies (6) if and only if $|M| = m_*$. The value of m_* is independent of c, but is weakly decreasing in γ with $\lim_{\gamma\to 1} m_* = n$ and $\lim_{\gamma\to\infty} m_* = 2$. In particular, $m_* = 3$ for $\gamma = 2$ and $m_* = 2$ for all $\gamma > 2$.

Here, the equilibrium size of stable coalitions is unique, decreasing in γ, and converges to 2 as γ increases. A large coalition requires small γ, a measure of the convexity of the abatement cost function. As γ increases, the benefit of cooperation increases but the equilibrium level of cooperation falls. Implausibly, the equilibrium is independent of the marginal damage parameter c. The linear model of Example 2 corrects this defect.

[3] By describing a negotiation process as a non-cooperative simultaneous move game, we are implicitly focusing on the very last minute of the negotiation where each player makes a participation decision with the understanding that the other players will not react to her decision. This story makes more sense in a dynamic setting where the end of negotiation period is explicit.

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

$$u_i\left(M\right) \geq u_i\left(M \setminus \{i\}\right) \quad \forall i \in M \tag{7}$$

and the 'if' part is equivalent to

$$u_i\left(M \cup i\right) < u_i\left(M\right) \quad \forall i \notin M. \tag{8}$$

Remark 2. In the model of Example 2, there exists a unique integer $m_* \geq 2$ such that M satisfies (6) if and only if $|M| = m_*$, the solution to

$$m_* = \min\{n, \lceil 1/c \rceil\}, \tag{9}$$

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

This model also has a unique equilibrium size of stable coalitions, but here it depends on the marginal damage parameter. Since m_* is decreasing in c, only small coalitions arise when c is large. In both examples, there are no large stable coalitions when cooperation is most beneficial. The intuition is familiar: Higher marginal damage make coalition members more willing to reduce their pollution even if the size of the coalition is small, giving all players an incentive to free-ride.

Naturally, in the symmetric setting Condition (6) does not identify the identify the members of a stable coalition. Denote $\mathcal{M} \subset \mathcal{N}$ as the set of equilibrium outcomes, namely,

$$\mathcal{M} := \{M \in \mathcal{N} \mid M \text{ satisfies (6)}\}. \tag{10}$$

In the examples above, $\mathcal{M}$ contains $C^n_{m_*} := \binom{n}{m_*}$ different outcomes of the game. Every stable equilibrium has the same number of members, but their identity varies. The literature does not to take a stand on which coalition in $\mathcal{M}$ materializes. This indeterminacy is innocuous in the static model. However, in a dynamic setting we need to describe players' beliefs about the negotiation outcome.[5]

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. Play-

[5]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.

ers who share a common belief π evaluate their ex-ante payoff as

$$\mathbb{E}_\pi\left[u_i(\tilde{M})\right] := \sum_{M\in\mathcal{M}} u_i(M)\pi_M, \tag{11}$$

where $\mathcal{M} \subset \mathcal{N}$ is defined by (10).

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.[6] The state variable at the beginning of a period is the coalition inherited from the previous period, M_{-1}. (In period 0 the inherited coalition is the null set.) 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 decides to deviate from the existing coalition in the first stage, that coalition dissolves and players move to the second stage where a stable coalition is randomly selected using the probability distribution π. 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.[7] 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}. 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

[6]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.

[7]The direction of any 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.

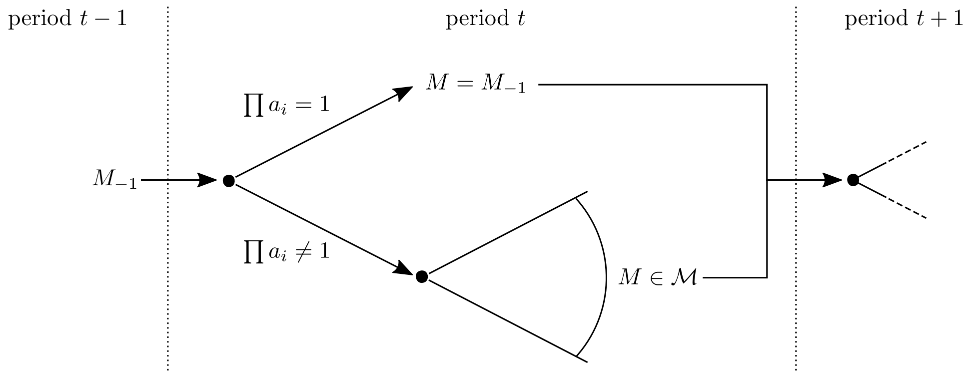

Figure 1: The timing of the game.

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 a Nash equilibrium of the second-stage participation game if and only if it satisfies[8]

$$i \in M \iff u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \geq u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}). \quad (12)$$

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 (12). 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. Thus, we need consider only that contingency. Player i's expected payoff of abandoning M_{-1} and reopening the negotiation process is

$$\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] := \sum_{M \in \mathcal{M}} \left(u_i(M) + \delta V_i(M) \right) \pi_M, \quad (13)$$

where

$$\mathcal{M} := \{ M \in \mathcal{N} \mid M \text{ satisfies } (12) \}. \quad (14)$$

[8]Equations (6) and (12) have the same interpretation. Taking as given the actions of other players at the second stage, members do not want to leave a coalition and non-members do not want to join it.

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

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi[u_i(\tilde{M}) + \delta V_i(\tilde{M})], \tag{15}$$

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 (14). 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 $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N})$ 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 $(\pi_M)_{M\in\mathcal{M}}$ is given by

$$\mathcal{M} = \{M \in \mathcal{N} \mid M \text{ satisfies (12) given } (V_i)_{i\in N}\}; \tag{16}$$

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

$$a_i(M_{-1}) \in \operatorname*{argmax}_{a_i\in\{0,1\}} \left\{ [u_i(M_{-1}) + \delta V_i(M_{-1})]\, a_i \right. \\ \left. + \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] (1 - a_i) \right\}; \tag{17}$$

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

$$V_i(M_{-1}) = \begin{cases} u_i(M_{-1}) + \delta V_i(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M})\right] & \text{otherwise.} \end{cases} \tag{18}$$

Condition (16) 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.[9] Condition (17) 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 (in which case i is indifferent between the two actions). This assumption is stronger than the standard definition of Nash equilibrium, which allows players to choose $a_i = 0$ in such a situation.

[9]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 (12), players realize that the negotiation process can actually result in M, invalidating the original belief which excludes M from its support.

This assumption rules out uninteresting equilibria where a player chooses $a_i = 0$ simply because another player chooses $a_j = 0$.

Condition (18) implies that even non-members can trigger the abandonment of the inherited coalition. We could assume that non-members do not have this veto power, but that modification does not change the equilibrium in the presence of a free-rider problem. Whenever the members of a coalition want to stick to the coalition, so do non-members.

To simplify the analysis, we emphasize a class of equilibria where π treats players symmetrically in the following sense.

Definition 2.2. A common belief π is symmetric if

$$|M| = |M'| \implies \pi_M = \pi_{M'}, \tag{19}$$

meaning that the probability of forming a coalition of a particular size is independent of the identity of its members.

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

3 Results

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

$$u_i(M) > u_i(M \cup \{i\}) \quad \forall i \in N \setminus M \iff |M| \geq m_* \tag{20}$$

and

$$u_i(M) \geq u_i(M \setminus \{i\}) \quad \forall i \in M \iff |M| \leq m_*. \tag{21}$$

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) $u_i(M) < u_j(M)$ for all $i \in M$ and $j \notin M$ as long as $|M| \geq m_*$.

Conditions (20) and (21) imply that there is a unique stable equilibrium, m_*, to the one-shot game. 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. Because the reduced-form payoff functions are assumed to be symmetric, we often denote by u_{in}^m and u_{out}^m 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

We first 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

$$\bar{u}^m := \frac{m}{n} u_{in}^m + \left(1 - \frac{m}{n}\right) u_{out}^m \tag{22}$$

and observe that under Assumption 1

$$u_{in}^m < \bar{u}^m < u_{out}^m \quad \forall m \geq m_* - 1. \tag{23}$$

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 receive a higher payoff, compared to the average payoff in the coalition with size m_*. That is $l^* > m_*$ is defined by

$$u_{in}^{l^*} \geq \bar{u}^{m_*} > u_{in}^{l^*-1}; \tag{24}$$

l^* is well defined 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*

$$a_i(M_{-1}) = \begin{cases} 1 & \text{if } |M_{-1}| \geq l^* \text{ and } i \in M_{-1} \\ 1 & \text{if } |M_{-1}| \geq m_* \text{ and } i \notin M_{-1} \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

together with the symmetric common belief $(\pi_M)_{M\in\mathcal{M}}$ *defined by*

$$\pi_M = 1/C^n_{m_*} \quad \forall M \in \mathcal{M}, \tag{26}$$

where

$$\mathcal{M} := \{M \in \mathcal{N} \,|\, |M| = m_*\}, \tag{27}$$

constitutes an equilibrium if and only if

$$\delta < \delta_{l^*} := \frac{u^{l^*-1}_{out} - u^{l^*}_{in}}{u^{l^*-1}_{out} - \bar{u}^{m_*}} \in (0, 1]. \tag{28}$$

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^* (e.g. the null set), 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 ($\delta < \delta_{l^*}$).[10]

Figure 2 illustrates the result 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 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

[10]This type of equilibrium does not exist for a larger δ because of condition (16), 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_*.

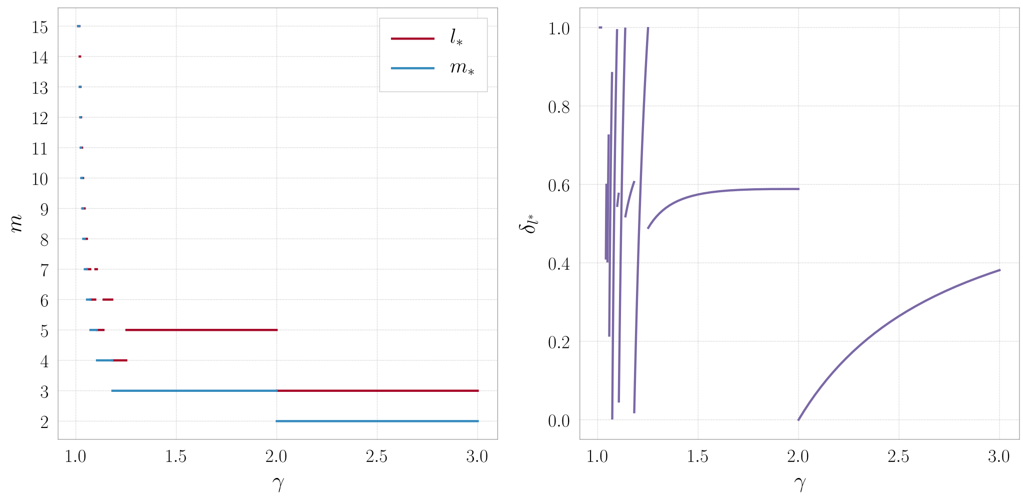

Figure 2: The values of m_* and l^* (left panel) and the threshold value δ_{l^*} of discount factor (right panel) in the model of Example 1. The number of players is set to $n = 15$.

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 2 shows that the value of δ_{l^*} depends on γ in a nontrivial manner, having many discontinuous points associated with discontinuity of m_* and l^*. 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. The implication is therefore similar to the static-version of the same model.

Example 2 suggests a slightly different relation, depicted in Figure 3. 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.

3.2 Equilibrium with multiple coalition sizes

We say that an equilibrium is 'sustainable' if it is both stable and, once achieved, 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. The requirement of permanence means that once

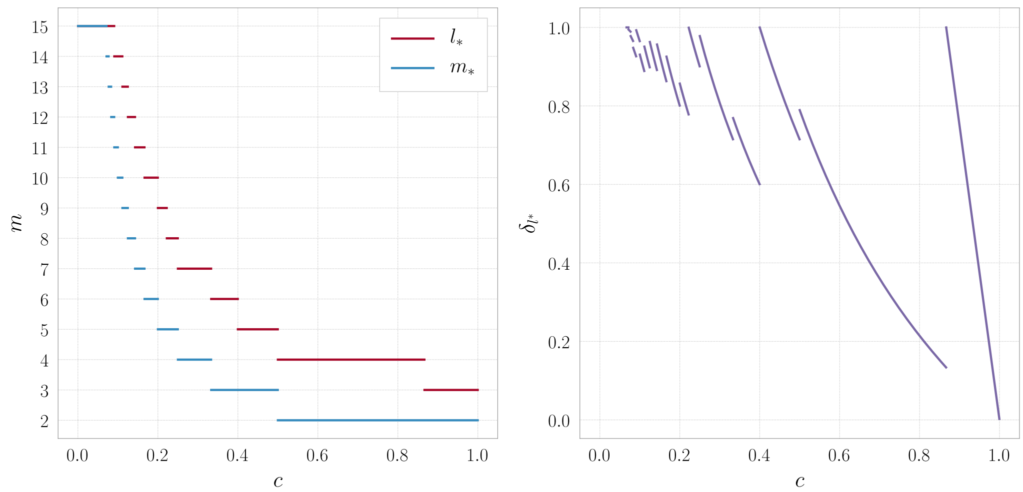

Figure 3: The values of m_* and l^* (left panel) and the threshold value δ_{l^*} of discount factor (right panel) in the model of Example 2. The number of players is set to $n = 15$.

achieved, members are willing to remain in it even though by doing so they give up the possibility of free riding. Members are willing to make this tradeoff only if they are sufficiently patient. The existence of sustainable equilibria therefore requires a large discount factor. Under the conditions of Proposition 3.1 all players want to stick with a sufficiently large equilibrium, one with $m \geq l^*$; but such equilibria are not stable, and therefore cannot be reached at the second-stage negotiation; here there are no sustainable equilibria. Proposition 3.2 characterizes sustainable equilibria for large δ.

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

a) There exists a symmetric common belief $(\pi_M)_{M\in\mathcal{M}}$ with

$$\mathcal{M} = \{M \in \mathcal{N} \,|\, |M| \in \{m_*, m^*\}\}\,, \tag{29}$$

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

$$a_i(M_{-1}) = \begin{cases} 1 & \text{if } |M_{-1}| \geq m^* \text{ and } i \in M_{-1} \\ 1 & \text{if } |M_{-1}| \geq k^* \text{ and } i \notin M_{-1} \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

constitutes an equilibrium.

b) The discount factor δ *is greater than*

$$\delta_{m^*} := \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \max\{\bar{u}^{m_*}, u_{in}^{m^*-1}\}} \in (0,1]. \tag{31}$$

The common belief associated with this equilibrium is given by

$$\pi_M = \begin{cases} \pi^{m^*}/C_{m^*}^n & \text{if } |M| = m^* \\ \left(1 - \pi^{m^*}/C_{m^*}^n\right)/C_{m_*}^n & \text{if } |M| = m_* \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

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

$$\Pi_\delta^{m^*} := \left(\max\left\{0, \frac{1-\delta}{\frac{\bar{u}^m - \bar{u}^{m_*}}{u_{in}^{m-1} - \bar{u}^{m_*}} - \delta}\right\}, \frac{\delta - \frac{u_{out}^{m-1} - u_{in}^m}{u_{out}^{m-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta}\frac{\bar{u}^m - u_{in}^m}{u_{out}^{m-1} - \bar{u}^{m_*}}} \right] \subset (0,1). \tag{33}$$

For a given discount factor, two forces constrain the equilibrium value of m^*. A stable coalition of the second-stage game cannot be too large, otherwise members will not resist the temptation to defect and free ride for a period. However, m^* cannot be too small, because otherwise members would not be willing defect from a smaller coalition in the hope of obtaining m^*. In general, the equilibrium value of m^* is not unique, especially when δ is large. More precisely, for a given δ any integer in the set

$$\{m \in \mathbb{N} | \max\{l^*, m_* + 2\} \leq m \leq n, \delta > \delta_{m^*}\} \tag{34}$$

can be an equilibrium value of m^*. For given δ, the largest element in this set gives a ceiling on the possible equilibrium coalition size; that ceiling increases with the discount factor.

Unlike the equilibrium we discussed in Proposition 3.1, the common belief in Proposition 3.2 is not uniquely determined, although it must lie in the interval given by definition (33). If π^{m^*} is 'too large', members' temptation to deviate from a large coalition is high, because they know that there is a high probability that they will be free-riders to a future large coalition; in that case, m^* is not stable. Large coalitions can not 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^*.

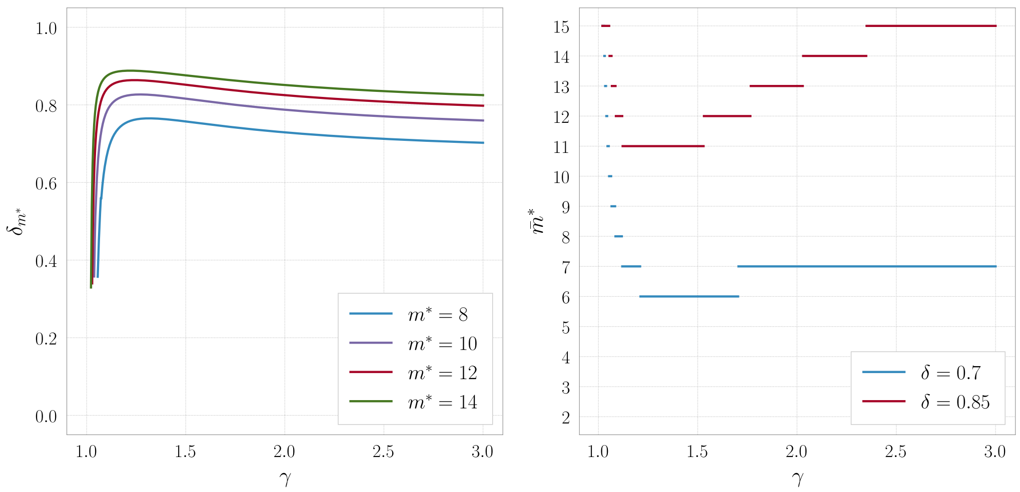

Figure 4: The threshold value δ_{m^*} of discount factor (left) and the largest size $\bar{m}^*$ of stable coalitions (right) in the model of Example 1.

To further illustrate the results, we revisit the two examples above.

Proposition 3.3. *In Example 1, for each* $m^* > \max\{l^*, m_* + 1\}$*, the equilibria described in Proposition 3.2 exist if and only if* δ *is greater than*

$$\delta_{m^*} = \frac{\gamma(m^* - 1)^{\frac{\gamma}{\gamma-1}} - (\gamma - 1)((m^*)^{\frac{\gamma}{\gamma-1}} - 1)}{(m^* - 1)^{\frac{\gamma}{\gamma-1}} - 1}. \tag{35}$$

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

The left panel of Figure 4 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 4 shows this relation more clearly. Here, $\bar{m}^*$ is the ceiling on the size of stable coalitions, defined by $\bar{m}^* = \max\{m^* \leq n \,|\, \delta > \delta_{m^*}\}$. 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 3.4. *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*

$$\delta_{m^*} = 1 - c. \tag{36}$$

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 is difficult to achieve when the marginal damage is large. Proposition 3.4 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 deviate from a coalition is decreasing in 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 would be required to pay if she remains in 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 $1 - c$.

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 5 the possible equilibrium combinations of m^* and π^{m^*} for different values of δ. Here, $m_* = 3$, $l^* = 5$, and $\delta_{l^*} = 0.588$. So, if δ is smaller than 0.588, the unique symmetric equilibrium is characterized by Proposition 3.1. Along the equilibrium path,

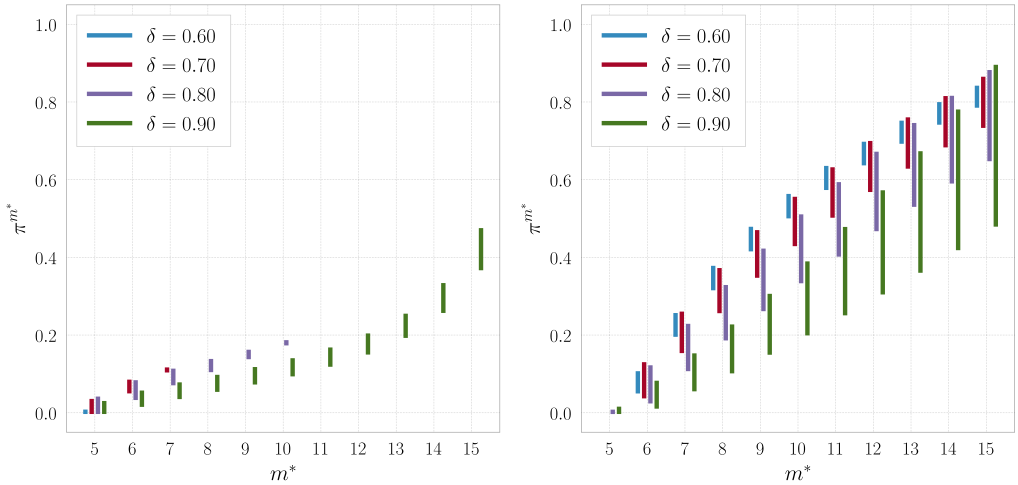

Figure 5: The equilibrium beliefs in Example 1 (left) and Example 2 (right). For each m^*, each bar represents the range $\Pi_\delta^{m^*}$ of possible values of π^{m^*} for different δ. The number of players is set to $n = 15$ for both cases. In Example 1 we set $\gamma = 2$ and in Example 2 we set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

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 instance, coalitions of size $m^* = 5$ are stable as well as those of size $m_* = 3$. 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 of size 5 with a positive but small probability, (less than or equal to 0.005). Once a coalition of size 5 is formed, players will stick with it. In this case, 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 5 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; the stable coalition has $m_* = 3$ members. By Proposition 3.4, $\delta_{m^*} = 1 - c = 0.525$ 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. Howe4ver, 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, the symmetric equilibrium is always characterized by either Proposition 3.1 or Proposition 3.2 for a given value of δ. 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 of size 3 or they may end up with another equilibrium where a significantly larger coalition, including the grand coalition, is possible. Moreover, in Example 1, while coalitions of a medium size can be easily included in the set of stable outcomes, very large coalitions are still impossible to achieve unless the discount factor is sufficiently close to 1. Example 2, by 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 a 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 5, 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 $m^* = 5$. The model of Example 2 has even greater multiplicity of equilibria.

Multiplicity helps explain the fact that international environmental agreements have sometimes failed and sometimes succeeded to attract a large number of participants when cooperation is important. Our model suggests that the actual outcome may depend on self-fulfilling beliefs generated by the political climate. In an optimistic environment, countries may believe that large coalitions are possible, leading to a good outcome. If there is little political momentum to solve the problem, in contrast, countries may believe that only small coalitions are possible, making it impossible to achieve a larger coalition.

Additional selection criteria can narrow the set of equilibria. For example, it might be reasonable to assume that an equilibrium is more plausible when the associated range $\Pi_\delta^{m_*}$ of possible equilibrium value of π^{m^*} is quite wide. If the economy experiences a shock that shifts the common belief to its ϵ-neighborhood, the updated value of π^{m^*} may jump out of the admissible range unless $\Pi_\delta^{m_*}$ is wide enough. Or one may simply argue that whenever multiple values for m^* are possible, the one associated with the largest interval $\Pi_\delta^{m_*}$ will most likely materialize because narrower $\Pi_\delta^{m_*}$ requires more precise belief coordination among otherwise uncoordinated players. This argument suggests that the most likely value of m^* is given by

$$m^* \in \underset{m}{\operatorname{argmax}}\{|\max \Pi_\delta^m - \inf \Pi_\delta^m|\}, \tag{37}$$

which normally pins down m^*. It is not possible to characterize such m^* in general, but for a sufficiently large discount factor, we have a sharp prediction.

Proposition 3.5. *Under Assumption 1, there always exists $\delta^* \in (0,1)$ such that*

$$\underset{m}{\operatorname{argmax}}\{|\max \Pi_\delta^m - \inf \Pi_\delta^m|\} = \{n\} \tag{38}$$

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

If the discount factor is sufficiently large, players most likely reach an equilibrium where the larger stable coalition is the grand coalition. Figure 5 illustrates this result; here the range $\Pi_\delta^{m^*}$ is the widest for $m^* = n$ when δ is close to 1. When players are sufficiently patient, they keep reopening the negotiation process until they achieve the grand coalition. 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.

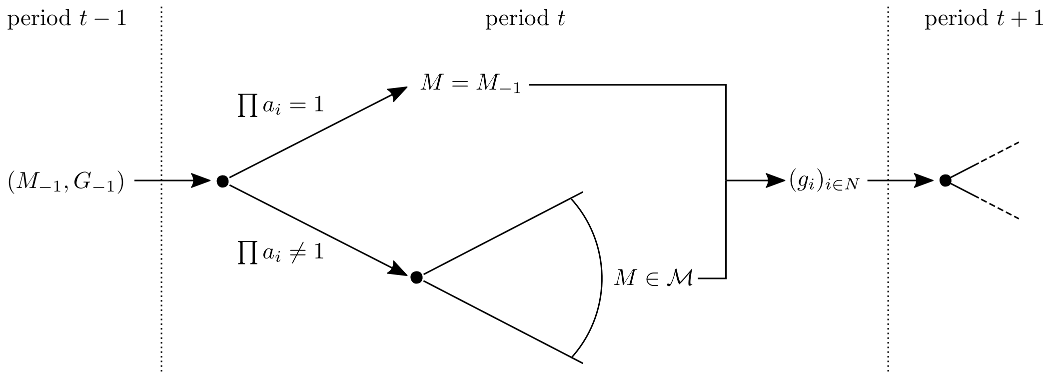

Figure 6: The timing of the structural game.

4.1 The model

To establish the isomorphism, we define a structural (as distinct from reduced-form) model, one characterized by a list $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$; as above, $\delta \in (0,1)$ is the discount factor and $N := \{1, 2, \dots, n\}$ is the set of all players. The function $\Phi_i(\boldsymbol{g}_t, G_t)$ determines the period payoff; the vector $\boldsymbol{g}_t := (g_{1,t}, \dots, g_{n,t})$ contains the players' emissions, which affect the evolution of the stock G_t, a public bad such as greenhouse gasses. The equation of motion for G is

$$G_t = F(\boldsymbol{g}_t, G_{t-1}) \tag{39}$$

for some function F. Player i's discounted present-value payoff at period t is

$$\sum_{s=t}^{T+t} \delta^{s-t}\Phi_i(\boldsymbol{g}_s, G_s), \text{with } T \leq \infty. \tag{40}$$

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. Each player's strategy is characterized by a pair of policy functions, a function $a_i(M_{-1}, G_{-1}) \in \{0, 1\}$ that determines the player's action in the first stage, and a real-valued function $g_i(M, G_{-1})$, that determines her contribution to G at the end of each period. Figure 6 depicts the timing of the game.

Let $V_i(M_{-1}, G_{-1})$ be player i's continuation value conditional on the coalition M_{-1} and the level of the public bad G_{-1} inherited from the preceding period. 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

$$
i \in M \iff \begin{aligned} &\hat{\Phi}_i(M \cup \{i\}, G_{-1}) + \delta \hat{V}_i(M \cup \{i\}, G_{-1}) \\ &\quad \geq \hat{\Phi}_i(M \setminus \{i\}, G_{-1}) + \delta \hat{V}_i(M \setminus \{i\}, G_{-1}), \end{aligned} \tag{41}
$$

where

$$
\hat{\Phi}_i(M, G_{-1}) := \Phi_i(\boldsymbol{g}(M, G_{-1}), F(\boldsymbol{g}(M, G_{-1}), G_{-1})) \tag{42}
$$

and

$$
\hat{V}(M, G_{-1}) := V_i(M, F(\boldsymbol{g}(M, G_{-1}), G_{-1})). \tag{43}
$$

Condition (41) is the structural analogue of (12). With this notation, we can define the equilibrium of structural models for $T \leq \infty$. The equilibrium continuation values, policy functions, and beliefs in any such equilibrium may depend on T. We are interested in the point-wise limit of this sequence of equilibria, as $T \to \infty$. We define this limit using:

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

$$
\mathcal{M}(G_{-1}) = \{M \in \mathcal{N} \mid M \text{ satisfies (41) given } (V_i)_{i \in N},\ (g_i)_{i \in N}, \text{ and } G_{-1}\}; \tag{44}
$$

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

$$
\begin{aligned} a_i(M_{-1}, G_{-1}) \in \underset{a_i \in \{0,1\}}{\operatorname{argmax}} \Big\{ &\left[\hat{\Phi}_i(M_{-1}, G_{-1}) + \delta \hat{V}_i(M_{-1}, G_{-1})\right] a_i \\ &+ \mathbb{E}_\pi \left[\hat{\Phi}_i(\tilde{M}, G_{-1}) + \delta \hat{V}_i(\tilde{M}, G_{-1})\right] (1 - a_i) \Big\}; \end{aligned} \tag{45}
$$

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

$$
\begin{aligned} (g_i(M, G_{-1}))_{i \in M} \in \ &\underset{(g_i)_{i \in M}}{\operatorname{argmax}} \sum_{i \in M} \{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i(M, F(\boldsymbol{g}, G_{-1}))\} \\ &\text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \notin M, \end{aligned} \tag{46}
$$

$$
\begin{aligned} g_i(M, G_{-1}) \in \ &\underset{g_i}{\operatorname{argmax}} \{\Phi_i(\boldsymbol{g}, F(\boldsymbol{g}, G_{-1})) + \delta V_i(M, F(\boldsymbol{g}, G_{-1}))\} \\ &\text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \in N \setminus \{i\} \end{aligned} \qquad \forall i \notin M; \tag{47}
$$

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

$$V_i(M_{-1}, G_{-1}) = \begin{cases} \hat{\Phi}_i(M_{-1}, G_{-1}) + \delta\hat{V}_i(M_{-1}, G_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}, G_{-1}) = 1 \\ \mathbb{E}_\pi\left[\hat{\Phi}_i(\tilde{M}, G_{-1}) + \delta\hat{V}_i(\tilde{M}, G_{-1})\right] & \text{otherwise.} \end{cases} \tag{48}$$

This definition is a straightforward extension of Definition 2.1. Unlike the equilibria of reduced-form models, $\mathcal{M}$ and a might depend on G_{-1}.[11]

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 model representation of a 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

$$\Phi_i(\boldsymbol{g}_t, G_t) = \phi_i(\boldsymbol{g}_t) - cG_t \tag{49}$$

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

$$F(\boldsymbol{g}_t, G_{t-1}) = f(\boldsymbol{g}_t) + \sigma G_{t-1} \tag{50}$$

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 the DSGE model of Golosov et al. (2014) and in the more generalized climate-economy model of Traeger (2015).

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

[11] 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 $T \in \{1, 2, \ldots, \infty\}$ and $M \in \mathcal{N}$, there exists a unique vector $\boldsymbol{g}^T(M) = (g_1^T(M), \ldots, g_n^T(M))$ that solves

$$\max_{(g_i)_{i \in M}} \sum_{i \in M} \left\{ \phi_i(\boldsymbol{g}) - c \frac{1 - (\delta\sigma)^T}{1 - \delta\sigma} f(\boldsymbol{g}) \right\} \text{ given } (g_j^T(M))_{j \in N \setminus M}, \tag{51}$$

and

$$\max_{g_i} \left\{ \phi_i(\boldsymbol{g}) - c \frac{1 - (\delta\sigma)^T}{1 - \delta\sigma} f(\boldsymbol{g}) \right\} \text{ given } (g_j^T(M))_{j \in N \setminus \{i\}} \quad \forall i \notin M, \tag{52}$$

simultaneously.

Under Assumptions 2 and 3, for each $T \in \{1, 2, \ldots, \infty\}$, we can define a mapping that transforms a structural model $\langle \delta, N, (\Phi_i)_{i \in N}, F \rangle$ into a reduced-form model $\langle \delta, N, (u_i^T)_{i \in N} \rangle$ where the period payoff function is

$$u_i^T(M) := \phi_i(\boldsymbol{g}^T(M)) - c \frac{1 - (\delta\sigma)^T}{1 - \delta\sigma} f(\boldsymbol{g}^T(M)) \quad \forall M \in \mathcal{N}. \tag{53}$$

For our purpose the mapping for $T = \infty$ is particularly relevant.

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

This proposition ensures that any equilibrium of the reduced-form model associated with a 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. 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* $((\pi_M)_{M \in \mathcal{M}}, (a_i)_{i \in N}, (g_i)_{i \in N})$ *is a limit equilibrium of the structural model* $\langle \delta, N, (\Phi_i)_{i \in N}, F \rangle$*, then* $((\pi_M)_{M \in \mathcal{M}}, (a_i)_{i \in N})$ *is an equilibrium of the 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.[12]

[12] 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 $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$ where

$$\Phi_i(\boldsymbol{g}, G) = -\frac{1}{2}(\bar{g}_i - g_i)^2 - cG \tag{54}$$

and

$$F(\boldsymbol{g}, G_{-1}) = \sigma G_{-1} + \sum_{i\in N} g_i. \tag{55}$$

With these functional forms, Assumptions 2 and 3 are both satisfied. In particular, we have

$$g_i^T(M) = \begin{cases} \bar{g}_i - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}|M| & \forall i \in M \\ \bar{g}_i - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma} & \forall i \notin M. \end{cases} \tag{56}$$

and

$$u_i^T(M) = \begin{cases} -c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\left\{\sum_{i\in N}\bar{g}_i - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\left(|M|^2 - |M| + n - \frac{1}{2}|M|^2\right)\right\} & \forall i \in M \\ -c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\left\{\sum_{i\in N}\bar{g}_i - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\left(|M|^2 - |M| + n - \frac{1}{2}\right)\right\} & \forall i \notin M \end{cases} \tag{57}$$

for each $T \in \{1, 2, \ldots, \infty\}$. Propositions 4.1 and 4.2 then show that it suffices to analyze the equilibrium of the associated reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$. Notice that this model, after being transformed into the reduced-form model, coincides with 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

$$\sum_{s=t}^{\infty} \delta^{s-t} \ln(C_{i,t}), \tag{58}$$

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

$$K_{i,t} = Y_{i,t} - C_{i,t}. \tag{59}$$

The production function is given by

$$Y_{i,t} = \Omega(G_t) A_{i,t-1} K_{i,t-1}^{\kappa} H_i(N_{i,t}^1, \dots, N_{i,t}^L) \quad \text{with} \quad \sum_{l=1}^{L} N_{i,t}^l = 1, \tag{60}$$

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 $(N_{i,t}^1, \dots, N_{i,t}^L)$ via

$$g_{i,t} = E_i(N_{i,t}^1, \dots, N_{i,t}^L) \tag{61}$$

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

$$G_t = F(\boldsymbol{g}_t, G_{t-1}) \tag{62}$$

for some function $F(\cdot)$.

We can simplify this structural model provided that

$$H_i^*(g_i) := \max_{N_i^1, \dots, N_i^L} \left\{ H_i(N_i^1, \dots, N_i^L) \,\middle|\, E_i(N_i^1, \dots, N_i^L) \leq g_i \right\} \tag{63}$$

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:

$$Y_{i,t} = \Omega(G_t) A_{i,t-1} K_{i,t-1}^{\kappa} H_i^*(g_{i,t}). \tag{64}$$

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

$$\begin{aligned}\sum_{s=t}^{\infty} \delta^{s-t} \ln(C_{i,s}) = & \frac{\kappa}{1-\delta\kappa} \ln(K_{i,t-1}) + \frac{1}{1-\delta\kappa} \sum_{s=t}^{\infty} \delta^{s-t} \ln\left(A_{i,s-1}\right) \\ & + \sum_{s=t}^{\infty} \delta^{s-t} \left(\ln(1-s_{i,s}) + \frac{\delta\kappa}{1-\delta\kappa} \ln(s_{i,s}) \right) \\ & + \frac{1}{1-\delta\kappa} \sum_{s=t}^{\infty} \delta^{s-t} \left\{ \ln\left(H_i^*(g_{i,s})\Omega(G_s)\right) \right\}. \end{aligned} \tag{65}$$

The first and the second terms on the right-hand side are both constant 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 $(G_s, g_{i,s})_{s=t}^{\infty}$. 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

$$\sum_{s=t}^{\infty} \delta^{s-t} \ln\left(H_i^*(g_{i,s})\Omega(G_s)\right). \tag{66}$$

Therefore, this model is represented by a structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F \rangle$ 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, can be better off by unilaterally deviating. 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 Schmidt and Kovac (2015), costly delay of an agreement, is explicit. In both models, countries must be able to commit to a long-term agreement.

In our model the new round of negotiation triggered by inheritance of a small coalition takes the place of the 'punishment'. In the second stage of a period, leaving a coalition of cut-off size m^* drives the remaining members to abandon the coalition in the next period. That abandonment makes the original defector strictly worse off. 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 back to mutual 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 doubt that transitory equilibria must satisfy Farrell and Maskin's (1989) definition of renegotiation proof. That definition requires that the equilibria cannot be Pareto ranked. In our setting, for sufficiently large discount factor, the stable set in the second-stage participation game consists of a small and a large coalition. 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 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 (when $m^* < n$). Moreover, once possible changes of coalition members are put on the negotiation 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* equilibrium 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 chicken game, where players make threats to induce others to back down.[13] 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.

[13]Although this fact was recognized early in the literature (Carraro and Siniscalco, 1992, 1993), there seems to be a common misconception that the underlying game is the prisoner's dilemma. Barrett (2003) clearly states that the participation game is a game of chicken, but later in the book where he discusses repeated games in detail, he seems to focus on the case of prisoner's dilemma.

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 could obtain other equilibria by relaxing these restrictions, but those analyzed in this paper would also be included in the larger set of equilibria.

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).[14] 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. This expectation 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.

[14] 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.

6 Conclusion

This paper presents a dynamic model of agreements among sovereign countries. The model's distinctive feature is that countries can abandon the existing agreement whenever doing so is in their national self-interest. This possibility reflects the reality of modern international relations, where countries cannot commit to agreements nor convince others that they are committed. Countries might be able to sign a long-term costly agreements to provide public goods. But countries 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.

Given the difficulty inherent in the relationship among sovereign countries, it is 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. As long as countries 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.' We 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 basic idea underlying our analysis is simple, but worth re-stating here. The exact outcome of the negotiation process of IEAs is inherently uncertain due to the multiplicity of equilibria. This uncertainty opens the possibility that countries keep cooperating once a sufficiently good agreement is established. For this outcome to be an equilibrium, countries must not be too pessimistic. They need to set the bar sufficiently high and also believe that it is possible to clear the hurdle. But excessive optimism would undermine a large existing agreement by making members think that defection is cheap. Cooperation among sovereign countries is possible only when they recognize that it is not easy to achieve. Cooperation requires 'sober optimism' based on a communication during a carefully designed pre-negotiation process.

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 (1992): "The international dimension of environmental policy," *European Economic Review*, 36, 379–387.

——— (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.

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.

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.

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.

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

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

A.1 Proof of Remark 1

Using (3), it is easy to see that $M \in \mathcal{N}$ is internally stable if and only if

$$\frac{(|M|-1)^{\frac{\gamma}{\gamma-1}}}{|M|^{\frac{\gamma}{\gamma-1}}-1} \leq \left(\frac{\gamma}{\gamma-1}\right)^{-1} \tag{A.1}$$

whereas the external stability requires

$$\left(\frac{\gamma}{\gamma-1}\right)^{-1} < \frac{|M|^{\frac{\gamma}{\gamma-1}}}{(|M|+1)^{\frac{\gamma}{\gamma-1}}-1}. \tag{A.2}$$

Put $\theta := \gamma/(\gamma-1) \in (1,\infty)$ so that θ is strictly decreasing in $\gamma > 1$ with $\lim_{\gamma\to 1}\theta = \infty$ and $\lim_{\gamma\to\infty}\theta = 1$. Define a function

$$\Gamma(x,\theta) := \theta\frac{(x-1)^{\theta}}{x^{\theta}-1} - 1 \quad \forall x > 1 \tag{A.3}$$

so that M is stable if and only if

$$|M| < n \quad \text{and} \quad \Gamma(|M|,\theta) \leq 0 < \Gamma(|M|+1,\theta) \tag{A.4}$$

or

$$|M| = n \quad \text{and} \quad \Gamma(|M|,\theta) \leq 0. \tag{A.5}$$

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

Proof. The claim immediately follows from the fact that

$$\lim_{x\to 1}\Gamma(x,\theta) = -1 < 0 < \theta - 1 = \lim_{x\to\infty}\Gamma(x,\theta) \tag{A.6}$$

and

$$\frac{\partial\Gamma(x,\theta)}{\partial x} = \frac{\theta^2(x-1)^{\theta-1}}{x^{\theta}-1}\left(1 - \frac{x^{\theta}-x^{\theta-1}}{x^{\theta}-1}\right) > 0 \tag{A.7}$$

for all $x > 1$. □

By Lemma A.1, we can implicitly define a mapping $x(\theta) \in (1,\infty)$ which satisfies $\Gamma(x(\theta),\theta) = 0$ for each $\theta \in (1,\infty)$. Put

$$m_* = \min\{n, \lfloor x(\theta) \rfloor\}, \tag{A.8}$$

where $\lfloor x(\theta) \rfloor$ is the greatest integer weakly smaller than $x(\theta)$. Then, M is stable if and only if $|M| = m_*$.

To complete the proof, we shall show that m_* defined by (A.8) is decreasing in γ with $\lim_{\gamma \to 1} m_* = n$ and $\lim_{\gamma \to \infty} = 2$. For this claim to be true, it suffices to prove that $x(\theta)$ is strictly increasing in θ with $\lim_{\theta \to \infty} x(\theta) = \infty$ and $\lim_{\theta \to 1} x(\theta) \in (2, 3)$. We prove this by combining a number of small lemmas.

Lemma A.2. *Define a function*

$$h(x) := 1 + \ln(x-1) - \frac{x}{x-1} \ln(x) \quad \forall x > 1. \tag{A.9}$$

Then $h(x)$ is strictly increasing in $x > 1$ and the equation $h(x) = 0$ has a unique root, which we denote by x_1. Moreover, x_1 lies in the interval $(2, 3)$.

Proof. Note that $\lim_{x \to 1} h(x) = -\infty$, $\lim_{x \to \infty} h(x) = 1$, and

$$h'(x) = \frac{\ln(x)}{(x-1)^2} > 0 \quad \forall x > 1. \tag{A.10}$$

Hence, there is a unique $x_1 > 1$ such that $h(x_1) = 0$. Also notice

$$h(2) = \ln(e/4) < 0 < \ln(2e/3) = h(3), \tag{A.11}$$

which implies that $x_1 \in (2, 3)$. □

Lemma A.3. *Let $x_1 \in (2, 3)$ be defined as in Lemma A.2. Define*

$$H(x, \theta) := 1 - \ln(\theta) + \ln(x^\theta - 1) - \frac{x^\theta}{x^\theta - 1} \ln(x^\theta) \quad \forall \theta > 1, \forall x > 1. \tag{A.12}$$

For each $x > 1$, $H(x, \theta)$ is strictly decreasing in θ. If $x > x_1$, $H(x, \theta) = 0$ has a unique root with respect to θ. If $x \le x_1$, $H(x, \theta) < 0$ for all $\theta > 1$.

Proof. The claim is immediate from the fact that

$$\frac{\partial H(x, \theta)}{\partial \theta} = -\frac{1}{\theta} \left(1 - \frac{x^\theta}{(x^\theta - 1)^2} \ln(x^\theta) \ln(x^\theta) \right) < 0 \tag{A.13}$$

for all $\theta > 1$ and $x > 1$, $\lim_{\theta \to 1} H(x, \theta) = h(x)$ and $\lim_{\theta \to \infty} H(x, \theta) = -\infty$. □

Lemma A.4. *If $\Gamma(x, \theta) = 0$ for some $x > 1$ and $\theta > 1$, then*

$$\frac{\partial \Gamma(x, \theta)}{\partial \theta} = \frac{(x-1)^\theta}{x^\theta - 1} H(x, \theta). \tag{A.14}$$

Proof. Noticing that

$$\Gamma(x,\theta)=0 \iff (x-1)^\theta=(x^\theta-1)/\theta, \tag{A.15}$$

we may write

$$\begin{aligned}\frac{\partial\Gamma(x,\theta)}{\partial\theta} &= \frac{(x-1)^\theta}{x^\theta-1}\left(1+\ln((x-1)^\theta)-\frac{x^\theta}{x^\theta-1}\ln(x^\theta)\right)\\ &= \frac{(x-1)^\theta}{x^\theta-1}\left(1-\ln(\theta)+\ln(x^\theta-1)-\frac{x^\theta}{x^\theta-1}\ln(x^\theta)\right)\\ &= \frac{(x-1)^\theta}{x^\theta-1}H(x,\theta), \end{aligned}\tag{A.16}$$

as claimed. □

Lemma A.5. *Let $x_1\in(2,3)$ be defined as in Lemma A.2. For each $x>x_1$, there exists at least one $\theta>1$ such that $\Gamma(x,\theta)=0$. For each $x\in(1,x_1]$, on the other hand, there is no $\theta>1$ such that $\Gamma(x,\theta)=0$.*

Proof. First observe that $\lim_{\theta\to1}\Gamma(x,\theta)=0$ and $\lim_{\theta\to\infty}\Gamma(x,\theta)=-1$ for any $x>1$. Also,

$$\lim_{\theta\to1}\frac{\partial\Gamma(x,\theta)}{\partial\theta}=1+\ln(x-1)-\frac{x}{x-1}\ln(x)=h(x), \tag{A.17}$$

which, by Lemma A.2, is strictly positive for any $x>x_1$. Hence, for each $x>x_1$, there exists at least one $\theta>1$ such that $\Gamma(x,\theta)=0$.

To prove the second half of the lemma, fix $x\in(1,x_1]$. Then, by Lemma A.2, $\lim_{\theta\to1}\partial\Gamma(x,\theta)/\partial\theta=h(x)$ is strictly negative for any $x<x_1$. Moreover, $\lim_{\theta\to1}\partial\Gamma(x_1,\theta)/\partial\theta=h(x_1)=0$ and

$$\lim_{\theta\to1}\frac{\partial^2\Gamma(x_1,\theta)}{\partial\theta^2}=\frac{(1+\ln(x_1-1))^2}{x_1}-1<0 \tag{A.18}$$

because $x_1\in(2,3)$. Hence, if $\Gamma(x,\theta)=0$ has a root, it must be the case that $\partial\Gamma(x,\theta)/\partial\theta\geq0$ at the smallest root. By Lemma A.4, however, this means that $H(x,\theta)\geq0$, which contradicts Lemma A.3. Therefore, if $x\in(1,x_1]$, there is no $\theta>1$ such that $\Gamma(x,\theta)=0$. □

Lemma A.6. *Let $x_1\in(2,3)$ be defined as in Lemma A.2. For each $x>x_1$, there exists a unique $\theta>1$ such that $\Gamma(x,\theta)=0$.*

Proof. Fix $x>x_1$. The existence follows from Lemma A.5. To prove the uniqueness, suppose by way of contradiction that there exist multiple θ's that satisfy

$\Gamma(x,\theta)=0$. Let θ_x be the smallest such θ and let $\theta'_x > \theta_x$ be the second smallest. Then, since $\lim_{\theta\to 1}\partial\Gamma(x,\theta)/\partial\theta = h(x) > 0$ by Lemma A.2, we know that $\partial\Gamma(x,\theta_x)/\partial\theta \le 0$. There are two cases to consider.

CASE 1. Consider the case where $\partial\Gamma(x,\theta_x)/\partial\theta < 0$. By Lemma A.4, this means that $H(x,\theta_x) < 0$, which together with Lemma A.3 implies that $H(x,\theta'_x) < 0$ because $\theta'_x > \theta_x$. By Lemma A.4, this in turn implies that $\partial\Gamma(x,\theta'_x)/\partial\theta < 0$. But since θ'_x is the second smallest root of $\Gamma(x,\theta)=0$, it must be the case that $\partial\Gamma(x,\theta'_x)/\partial\theta \ge 0$, a contradiction.

CASE 2. Consider the other case where $\partial\Gamma(x,\theta_x)/\partial\theta = 0$. If $\partial^2\Gamma(x,\theta_x)/\partial\theta^2 \le 0$, exactly the same argument as in CASE 1 yields a contradiction. If $\partial^2\Gamma(x,\theta_x)/\partial\theta^2 > 0$, on the other hand, pick $x' \in (x_1, x)$. Note that $\Gamma(x',\theta_x) < \Gamma(x,\theta_x) = 0$ because $\Gamma(x,\theta)$ is strictly increasing in x by Lemma A.1. Hence, by making x' sufficiently close to x, one can find $\theta_{x'} < \theta_x$ and $\theta'_{x'} > \theta_x$ such that $\Gamma(x',\theta_{x'}) = \Gamma(x',\theta'_{x'}) = 0$ and $\partial\Gamma(x',\theta_{x'})/\partial\theta < 0 < \partial\Gamma(x',\theta'_{x'})/\partial\theta$. But again, applying the same argument as in CASE 1 yields a contradiction then.

We therefore conclude that for each $x > x_1$, the equation $\Gamma(x,\theta)=0$ has at most one root with respect to θ. $\square$

By combining Lemmas A.1 and A.6, we know that the mapping $x(\theta)$ defined by $\Gamma(x(\theta);\theta)$ is bijective from $(1,\infty)$ to (x_1,∞). To see that $x(\theta)$ is strictly increasing, observe

$$\Gamma(3,2) = \frac{2^2}{3^2-1} = \frac{1}{2}, \tag{A.19}$$

which implies $x(2)=3$. Moreover, since

$$\Gamma(3,3) = \frac{8}{26} < \frac{1}{3} = \Gamma(x(3),3), \tag{A.20}$$

we must have $x(3) > 3 = x(2)$. Hence, $x(\theta)$ must be strictly increasing in θ. Since $x(2)=3$, it then follows that $x(\theta) < 3$ for all $\theta \in (1,2)$. This implies that $m_* = 2$ for all $\gamma > 2$ because $\theta = 2$ if and only if $\gamma = 2$.

A.2 Proof of Remark 2

To prove the 'if' part, put $m_* := \min\{n, \lceil 1/c \rceil\}$ and fix M such that $|M| = m_*$. If $n \le \lceil 1/c \rceil$, or $1/c \le n < 1/c + 1$, then $m_* = n$ and therefore

$$u_{in}^{|M|} - u_{out}^{|M|-1} = cm_* - 1 = cn - 1 \ge 0, \tag{A.21}$$

meaning that M is internally stable. Since there is no outsiders of $M = N$, it follows that M is stable. If $n > \lceil 1/c \rceil$, or $n \geq 1/c - 1$, then $m_* = \lceil 1/c \rceil$ and therefore $1/c \leq m_* < 1/c + 1$. It follows that

$$u_{in}^{|M|} - u_{out}^{|M|-1} = cm_* - 1 \geq 0 \tag{A.22}$$

and

$$u_{out}^{|M|} - u_{in}^{|M|+1} = 1 - c > 0, \tag{A.23}$$

meaning that M is both internally and externally stable.

To prove the 'only if' part, let M be a stable coalition. it is immediate from (5) that 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 $1/c \leq |M| < 1/c + 1$. If $M = N$ and $|M| < 1/c + 1$, it must be the case that $n \leq \lceil 1/c \rceil$ and therefore $|M| = |N| = n = \min\{n, \lceil 1/c \rceil\} = m_*$. If $M \neq N$ and $1/c \leq |M| < 1/c+1$, on the other hand, $n \geq |M|+1 > \lceil 1/c \rceil = |M|$ and therefore $|M| = \lceil 1/c \rceil = \min\{n, \lceil 1/c \rceil\} = m_*$. This completes the proof.

A.3 Proof of Proposition 3.1

Let us first prove the following lemma.

Lemma A.7. *Given the strategy profile* (25)*, M satisfies* (12) *only if $|M| \in \{m_*, l^*\}$.*

Proof. Let M be a coalition that satisfies (12). First, suppose $|M| \leq l^* - 1$. Then,

$$V_i(M \cup \{i\}) = V_i(M \setminus \{i\}) \quad \forall i \in M. \tag{A.24}$$

So the internal stability required in (12) suggests that

$$u_{in}^{|M|} \geq u_{out}^{|M|-1}, \tag{A.25}$$

which, by definition of m_*, is only possible when $|M| \leq m_*$. But $|M| < m_*$ is impossible because the external stability in (12) suggests that

$$u_{in}^{|M|+1} < u_{out}^{|M|}, \tag{A.26}$$

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

On the other hand, $|M|$ cannot be greater than l^*. To see this, suppose $|M| \geq l^* + 1$. Then

$$\begin{cases} V_i(M \cup \{i\}) = \frac{1}{1-\delta} u_{in}^{|M|} \\ V_i(M \setminus \{i\}) = \frac{1}{1-\delta} u_{out}^{|M|-1} \end{cases} \qquad \forall i \in M \tag{A.27}$$

and hence the internal stability required in (12) suggests that

$$u_{in}^{|M|} \geq u_{out}^{|M|-1}, \tag{A.28}$$

which is only possible when $|M| \leq m_*$. But this contradicts the fact that $l^* > m_*$ and therefore $|M|$ cannot be greater than l^*. It follows that either $|M| = m_*$ or $|M| = l^*$. $\square$

We can now prove the 'if' part of the proposition. Suppose that the discount factor δ satisfies

$$\delta < \delta_{l^*} := \frac{u_{out}^{l^*-1} - u_{in}^{l^*}}{u_{out}^{l^*-1} - \bar{u}^{m_*}} \in (0, 1]. \tag{A.29}$$

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

$$V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_i(M_{-1}) & \text{if } |M| \geq l^* \\ \frac{1}{1-\delta} \bar{u}^{m_*} & \text{otherwise} \end{cases} \tag{A.30}$$

solve (18). It follows that

$$u_i(M_{-1}) + \delta V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_i(M_{-1}) & \text{if } |M_{-1}| \geq l^* \\ u_i(M_{-1}) + \frac{\delta}{1-\delta} \bar{u}_i^{m_*} & \text{otherwise,} \end{cases} \tag{A.31}$$

and

$$\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] = \frac{1}{1-\delta} \bar{u}^{m_*}, \tag{A.32}$$

which means that

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \iff u_i(M_{-1}) \geq \bar{u}^{m_*}. \tag{A.33}$$

Notice that

$$u_i(M_{-1}) \geq \bar{u}^{m_*} \iff u_{in}^{|M_{-1}|} \geq \bar{u}^{m_*} \iff |M_{-1}| \geq l^* \tag{A.34}$$

for $i \in M_{-1}$ and

$$u_i(M_{-1}) \geq \bar{u}^{m_*} \iff u_{out}^{|M_{-1}|} \geq \bar{u}^{m_*} \iff |M_{-1}| \geq m_* \tag{A.35}$$

for $i \notin M_{-1}$. Hence, given $(\pi_M)_{M \in \mathcal{M}}$ and $(V_i)_{i \in N}$, the policy functions $(a_i)_{i \in N}$ defined by (25) actually satisfy (17).

To complete the proof of the 'if' part, we shall show that given $(V_i)_{i \in N}$, M satisfies (12) 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 such that $|M| = m_*$. Then for each $i \in M$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{m_*} + \frac{\delta}{1-\delta} \bar{u}_i^{m_*} \\ &\geq u_{out}^{m_*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned} \tag{A.36}$$

where the inequality follows from the definiton of m_*. For each $i \notin M$, on the other hand, since $m_* + 1 = l^*$, we have

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &< u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M) + \delta V_i(M), \end{aligned} \tag{A.37}$$

where the inequality is due to (A.29). This means that coalitions of size m_* satisfy (12). We need to prove that none of the other coalitions satisfy (12). But Lemma A.7 suggests that a coalition is stable only if its size if m_* or l^*. Thus, all we need to do is to show that coalitions of size l^* cannot satisfy (12). In fact, coalitions of size l^* are not internally stable because for each $i \in M$ with $|M| = l^*$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &< u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned} \tag{A.38}$$

where the inequality is again implied by (A.29).

Consider the other case where $l^* > m_* + 1$. In this case, the definition of m_* directly implies that coalitions of size m_* satisfy (12). 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.7, we conclude that M satisfies (12) 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 as in Proposition 3.1 cannot constitute an equilibrium. In particular, we claim that coalitions of size l^* then satisfy (12) if $\delta \geq \delta_{l^*}$. First, coalitions of size l^* are internally stable because for each $i \in M$ with $|M| = l^*$,

$$\begin{aligned} u_i(M) + \delta V_i(M) &= u_{in}^{l^*} + \frac{\delta}{1-\delta} u_{in}^{l^*} \\ &\geq u_{out}^{l^*-1} + \frac{\delta}{1-\delta} \bar{u}^{m_*} \\ &= u_i(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \end{aligned} \tag{A.39}$$

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

$$\begin{aligned} u_i(M \cup \{i\}) + \delta V_i(M \cup \{i\}) &= u_{in}^{l^*+1} + \frac{\delta}{1-\delta} u_{in}^{l^*+1} \\ &< u_{out}^{l^*} + \frac{\delta}{1-\delta} u_{out}^{l^*} \\ &= u_i(M) + \delta V_i(M), \end{aligned} \tag{A.40}$$

where the inequality is due to the fact that $l^* \geq m_*$. This is inconsistent with the common belief defined in Proposition 3.1, which presumes that only coalitions of size m_* satisfy (12).

A.4 Proof of Proposition 3.2

Fix $m^* \geq \max\{l^*, m_* + 2\}$. The following lemma shows that, given the strategy profile $(a_i)_{i\in N}$ defined by (30), coalitions of size m_* are always stable, and if there is any other stable coalition, it must be of size m^*.

Lemma A.8. *Given the strategy profile* (30)*, a) M satisfies* (12) *only if $|M| \in \{m_*, m^*\}$, and b) M satisfies* (12) *whenever $|M| = m_*$.*

Proof. Exactly the same argument as in the proof of Lemma A.7 proves part a). To prove part b), notice that $V_i(M \cup \{i\}) = V_i(M \setminus \{i\})$ for all $i \in N$ if $|M| <$

$m^* - 1$. Since $m^* > m_* + 1$, it follows from the definition of m_* that M satisfies (12) whenever $|M| = m_*$. $\square$

To ease the notation, define

$$\bar{u}(m, \pi^m) := \bar{u}^m \frac{\pi^m}{1 - \delta(1 - \pi^m)} + \bar{u}^{m_*} \left(1 - \frac{\pi^m}{1 - \delta(1 - \pi^m)}\right) \tag{A.41}$$

for each $m \geq m_*$ and for each $\pi^m \in (0, 1)$. Note that

$$\bar{u}^m > \bar{u}(m, \pi^m) > \bar{u}^{m_*} \quad \forall m > m_* \tag{A.42}$$

and $\bar{u}(m_*, \pi^{m_*}) = \bar{u}^{m_*}$ because $\bar{u}^m$ is strictly increasing in m.

We first prove that statement b) in the proposition implies statement a). Suppose $\delta > \delta_{m^*}$. Let $(\pi_M)_{M \in \mathcal{M}}$ be defined as (29) and (32) where we fix $\pi^{m^*} \in \Pi_\delta^{m^*}$. Notice that $\Pi_\delta^{m^*}$ is nonempty as long as $\delta > \delta_{m^*}$. Let $(a_i)_{i \in N}$ be defined by (30) where we choose $k^* \in \{m_*, \ldots, n-1\}$ such that

$$u_{out}^{k^*} \geq \bar{u}(m^*, \pi^{m^*}) > u_{out}^{k^*-1}. \tag{A.43}$$

Under Assumption 1, such a k^* always exists and is unique because

$$u_{out}^{n-1} > u_{in}^n = \bar{u}^n > \bar{u}(m^*, \pi^{m^*}) > \bar{u}^{m_*} > u_{in}^{m_*} \geq u_{out}^{m_*-1} \tag{A.44}$$

and u_{out}^m is strictly increasing in $m \geq m_* - 1$. We shall show that $(\pi_M)_{M \in \mathcal{M}}$ and $(a_i)_{i \in N}$ constitute an equilibrium. With $(\pi_M)_{M \in \mathcal{M}}$ and $(a_i)_{i \in N}$ given, the value functions $(V_i)_{i \in N}$ defined by

$$V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_i(M_{-1}) & \text{if } |M_{-1}| \geq m^* \\ \frac{1}{1-\delta} \bar{u}(m^*, \pi^{m^*}) & \text{otherwise} \end{cases} \tag{A.45}$$

solve (18). It follows that

$$u_i(M_{-1}) + \delta V_i(M_{-1}) = \begin{cases} \frac{1}{1-\delta} u_i(M_{-1}) & \text{if } |M_{-1}| \geq m^* \\ u_i(M_{-1}) + \frac{\delta}{1-\delta} \bar{u}(m^*, \pi^{m^*}) & \text{otherwise,} \end{cases} \tag{A.46}$$

and

$$\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] = \bar{u}(m^*, \pi^{m^*}), \tag{A.47}$$

which means that

$$u_i(M_{-1}) + \delta V_i(M_{-1}) \geq \mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] \iff u_i(M_{-1}) \geq \bar{u}(m^*, \pi^{m^*}). \tag{A.48}$$

Notice that

$$u_i(M_{-1}) \geq \bar{u}(m^*, \pi^{m^*}) \iff u_{in}^{|M_{-1}|} \geq \bar{u}(m^*, \pi^{m^*}) \tag{A.49}$$

for $i \in M_{-1}$ and

$$u_i(M_{-1}) \geq \bar{u}(m^*, \pi^{m^*}) \iff u_{out}^{|M_{-1}|} \geq \bar{u}(m^*, \pi^{m^*}) \tag{A.50}$$

for $i \notin M_{-1}$. Hence, given $(\pi_M)_{M\in\mathcal{M}}$ and $(V_i)_{i\in N}$, the policy functions $(a_i)_{i\in N}$ defined by (30) actually satisfy (17) if

$$u_{in}^{m^*} \geq \bar{u}(m^*, \pi^{m^*}) > u_{in}^{m^*-1} \tag{A.51}$$

and

$$u_{out}^{k^*} \geq \bar{u}(m^*, \pi^{m^*}) > u_{out}^{k^*-1}. \tag{A.52}$$

Notice that (A.51) holds by choice of π^{m^*} and (A.52) is satisfied by construction of k^*. We also need to prove that any coalition of size m_* or m^* satisfies (12) whereas none of the other coalitions do. By Lemma A.7, however, it suffices to show that coalitions of size m^* satisfy (12). The external stability of coalitions of size m^* is trivial. The condition for internal stability may be written as

$$u_{in}^{m^*} \geq (1-\delta) u_{out}^{m^*-1} + \delta \bar{u}(m^*, \pi^{m^*}), \tag{A.53}$$

which is again satisfied as long as π^{m^*} is chosen from $\Pi_\delta^{m^*}$.

Conversely, statement a) in the proposition implies statement b). To see this, suppose that $(\pi_M)_{M\in\mathcal{M}}$ and $(a_i)_{i\in N}$ described in statement a) constitutes an equilibrium. With $(\pi_M)_{M\in\mathcal{M}}$ given, denote by π^{m^*} the probability of a coalition of size m^* is drawn from the distribution, namely, $\pi^{m^*} := \sum_{|M|=m^*} \pi_M$. Since coalitions of size m^* are stable in this equilibrium, they must be internally stable, the condition for which may be written as

$$u_{in}^{m^*} \geq (1-\delta) u_{out}^{m^*-1} + \delta \bar{u}(m^*, \pi^{m^*}). \tag{A.54}$$

Rearranging terms then yields

$$\pi^{m^*} \leq \frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta} \frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}. \tag{A.55}$$

On the other hand, since m^* is used as the optimal cut-off size in the policy function, it must be the case that

$$\mathbb{E}_\pi \left[u_i(\tilde{M}) + \delta V_i(\tilde{M}) \right] > u_i(M_{-1}) + \delta V_i(M_{-1}) \quad \forall i \in M_{-1} \tag{A.56}$$

whenever $|M_{-1}| < m^*$. This in particular implies

$$\bar{u}(m^*, \pi^{m^*}) > u_{in}^{m^*-1}, \tag{A.57}$$

which may be written as

$$\pi^{m^*} > \frac{1-\delta}{\frac{\bar{u}^{m^*} - \bar{u}^{m_*}}{u_{in}^{m^*-1} - \bar{u}^{m_*}} - \delta}. \tag{A.58}$$

Since $\pi^{m^*} \in (0,1)$, combining (A.55) and (A.58) yields

$$\max \left\{ 0, \frac{1-\delta}{\frac{\bar{u}^{m^*} - \bar{u}^{m_*}}{u_{in}^{m^*-1} - \bar{u}^{m_*}} - \delta} \right\} < \min \left\{ 1, \frac{\delta - \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}}{\delta + \frac{\delta}{1-\delta} \frac{\bar{u}^{m^*} - u_{in}^{m^*}}{u_{out}^{m^*-1} - \bar{u}^{m_*}}} \right\}, \tag{A.59}$$

which requires $\delta > \delta_{m^*}$.

A.5 Proof of Proposition 3.3

Putting $\theta = \gamma/(\gamma - 1)$, we may write

$$\frac{\partial \delta_{m^*}}{\partial m^*} \frac{1}{\delta_{m^*}} = \frac{\theta^2 (m^*-1)^{\theta-1} - \theta (m^*)^{\theta-1}}{\theta (m^*-1)^\theta - ((m^*)^\theta - 1)} - \frac{\theta (m^*-1)^{\theta-1}}{(m^*-1)^\theta - 1}. \tag{A.60}$$

Hence, δ_{m^*} is increasing in m^* if

$$(m^*)^{\theta-1} - \theta > 1 - \left(\frac{m^*}{m^*-1} \right)^{\theta-1}. \tag{A.61}$$

Since $\theta > 1$, the right-hand side is negative for any $m^* \geq 2$. The left-hand side is positive as long as $m^* \geq e$. Since $m^* \geq m_* + 1$ by definition and since $m_* \geq 2$

by Proposition 1, (A.61) holds for all m^* and for all θ. We conclude that δ_{m^*} is increasing in m^*.

A.6 Proof of Proposition 3.4

Since $m^* > m_*$,

$$u_{out}^{m^*-1} - u_{in}^{m^*} = 1 - c \tag{A.62}$$

whereas

$$u_{out}^{m^*-1} - u_{in}^{m^*-1} = 1. \tag{A.63}$$

Hence, we have

$$\delta_{m^*} = \frac{u_{out}^{m^*-1} - u_{in}^{m^*}}{u_{out}^{m^*-1} - u_{in}^{m^*-1}} = 1 - c. \tag{A.64}$$

A.7 Proof of Proposition 3.5

Observe that

$$\lim_{\delta \to 1} \left(\max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \right) = 0 \tag{A.65}$$

for any $m^* \neq n$ whereas

$$\lim_{\delta \to 1} \left(\max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \right) = \frac{u_{in}^{n} - \bar{u}^{m_*}}{u_{out}^{n-1} - \bar{u}^{m_*}} > 0. \tag{A.66}$$

It follows that there exists $\delta^* \in (0, 1)$ such that

$$\delta > \delta^* \implies \max \Pi_\delta^n - \inf \Pi_\delta^n > \max \Pi_\delta^{m^*} - \inf \Pi_\delta^{m^*} \quad \forall m^* \neq n, \tag{A.67}$$

which proves the statement of the proposition.

A.8 Proof of Proposition 4.1

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

$$i \in M \iff u_i^\infty(M \cup \{i\}) + \delta V_i(M \cup \{i\}) \geq u_i^\infty(M \setminus \{i\}) + \delta V_i(M \setminus \{i\}), \tag{A.68}$$

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

$$a_i(M_{-1}) \in \operatorname*{argmax}_{a_i\in\{0,1\}} \Bigg\{ \left[u_i^\infty(M_{-1}) + \delta V_i(M_{-1})\right] a_i \\ + \mathbb{E}_\pi \left[u_i^\infty(\tilde{M}) + \delta V_i(\tilde{M})\right] (1-a_i) \Bigg\}, \tag{A.69}$$

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

$$V_i(M_{-1}) = \begin{cases} u_i^\infty(M_{-1}) + \delta V_i(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi \left[u_i^\infty(\tilde{M}) + \delta V_i(\tilde{M})\right] & \text{otherwise,} \end{cases} \tag{A.70}$$

where

$$u_i^\infty(M) = \phi_i(\boldsymbol{g}^\infty(M)) - \frac{c}{1-\delta\sigma} f(\boldsymbol{g}^\infty(M)). \tag{A.71}$$

With (A.68), (A.69), (A.70), (A.71), (51), and (52), it is straightforward to see that $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i^\infty)_{i\in N})$ is an equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$ where the corresponding value functions are given by

$$V_i^\infty(M_{-1}, G_{-1}) := V_i(M_{-1}) - \frac{c}{1-\delta\sigma}\sigma G_{-1}. \tag{A.72}$$

A.9 Proof of Proposition 4.2

We first prove the following lemma.

Lemma A.9. *Under Assumptions 2 and 3, for any $T\in\mathbb{N}$, if $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i)_{i\in N})$ is an equilibrium at the initial period of the T-period version of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$, then $\mathcal{M}$ and a_i are both independent of G_{-1} and the policy function g_i coincides with g_i^T defined in Assumption 3. The value function associated with the T-period model is given by*

$$V_i(M_{-1}, G_{-1}) = v_i(M_{-1}) - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\sigma G_{-1} \tag{A.73}$$

for some function v_i.

Proof. Consider a finite-period version of the model where the number of remaining periods is $T\in\mathbb{N}$. Let us begin with the case with $T=1$ where (M_{-1}, G_{-1}) is given. Let $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i)_{i\in N})$ be an equilibrium of the one-period structural model. We shall show that $\mathcal{M}$ and a_i are both independent of G_{-1}, the policy function g_i coincides with g_i^1, and the value function is given by

$$V_i(M_{-1}, G_{-1}) = v_i(M_{-1}) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma}\sigma G_{-1} \tag{A.74}$$

for some function v_i.

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

$$\begin{aligned} (g_i(M, G_{-1}))_{i\in M} \in \; & \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \textstyle\sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - c\left[\sigma G_{-1} + f(\boldsymbol{g})\right]\right\} \\ & \text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \notin M, \end{aligned} \tag{A.75}$$

and

$$\begin{aligned} g_i(M, G_{-1}) \in \; & \underset{g_i}{\operatorname{argmax}} \left\{\phi_i(\boldsymbol{g}) - c\left[\sigma G_{-1} + f(\boldsymbol{g})\right]\right\} \\ & \text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \in N \setminus \{i\} \end{aligned} \qquad \forall i \notin M \tag{A.76}$$

for each $M \in \mathcal{N}$. By Assumption 3, this means that $(g_i(M, , G_{-1}))_{i\in N} = (g_i^1(M))_{i\in N}$, independent of G_{-1}. This in turn implies that $\mathcal{M}$ must be the collection of all M such that

$$\begin{aligned} i \in M \iff & \phi_i(\boldsymbol{g}^1(M\cup\{i\})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(M\cup\{i\}))\right] \\ & \geq \phi_i(\boldsymbol{g}^1(M\setminus\{i\})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(M\setminus\{i\}))\right] \\ \iff & u_i^1(M\cup\{i\}) - c\sigma G_{-1} \geq u_i^1(M\setminus\{i\}) - c\sigma G_{-1} \\ \iff & u_i^1(M\cup\{i\}) \geq u_i^1(M\setminus\{i\}), \end{aligned} \tag{A.77}$$

where

$$u_i^1(M) = \phi_i(\boldsymbol{g}^1(M)) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma} f(\boldsymbol{g}^1(M)). \tag{A.78}$$

This means that $\mathcal{M}$ is independent of G_{-1}. The policy functions $(a_i)_{i\in N}$ are also independent of G_{-1} because they must solve

$$\begin{aligned} a_i^1(M_{-1}) \in \underset{a_i\in\{0,1\}}{\operatorname{argmax}} & \Bigg\{ \left[\phi_i(\boldsymbol{g}^1(M_{-1})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(M_{-1}))\right]\right] a_i \\ & + \mathbb{E}_\pi\left[\phi_i(\boldsymbol{g}^1(\tilde{M})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^1(\tilde{M}))\right]\right](1-a_i) \Bigg\} \\ = \underset{a_i\in\{0,1\}}{\operatorname{argmax}} & \left\{ u_i^1(M_{-1})a_i + \mathbb{E}_\pi\left[u_i^1(\tilde{M})\right](1-a_i) \right\}. \end{aligned} \tag{A.79}$$

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

$$V_i(M_{-1}, G_{-1}) = v_i(M_{-1}) - c\frac{1-(\delta\sigma)^1}{1-\delta\sigma}\sigma G_{-1}, \tag{A.80}$$

where

$$v_i(M_{-1}) := \begin{cases} u_i^1(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi\left[u_i(\tilde{M})\right] & \text{otherwise.} \end{cases} \tag{A.81}$$

This proves that the statement of lemma is true for $T = 1$.

Now we shall prove the statement of lemma for all $T \in \mathbb{N}$. Suppose, as an induction hypothesis, that the statement is true for some $T \in \mathbb{N}$. Denote by $((\pi_M)_{M\in\mathcal{M}^T}, (a_i^T)_{i\in N}, (g_i^T)_{i\in N})$ an equilibrium at the initial period of the T-period structural model. By assumption, the associated value function is given by

$$V_i^T(M_{-1}, G_{-1}) = v_i^T(M_{-1}) - c\frac{1-(\delta\sigma)^T}{1-\delta\sigma}\sigma G_{-1} \tag{A.82}$$

for some function v_i^T. Let $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i)_{i\in N})$ be an equilibrium at the initial period of the $T+1$-period structural model. We shall show that $\mathcal{M}$ and a_i are both independent of G_{-1}, the policy function g_i coincides with g_i^{T+1}, and the value function is given by

$$V_i(M_{-1}, G_{-1}) = v_i(M_{-1}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1} \tag{A.83}$$

for some function v_i.

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

$$\begin{aligned}(g_i(M, G_{-1}))_{i\in M} \in \ & \underset{(g_i)_{i\in M}}{\operatorname{argmax}} \textstyle\sum_{i\in M} \left\{\phi_i(\boldsymbol{g}) - c\left[\sigma G_{-1} + f(\boldsymbol{g})\right] + \delta V_i^T(M, \sigma G_{t-1} + f(\boldsymbol{g}))\right\} \\ & \text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \notin M,\end{aligned} \tag{A.84}$$

and

$$\begin{aligned}g_i(M, G_{-1}) \in \ & \underset{g_i}{\operatorname{argmax}} \left\{\phi_i(\boldsymbol{g}) - c\left[\sigma G_{-1} + f(\boldsymbol{g})\right] + \delta V_i^T(M, \sigma G_{t-1} + f(\boldsymbol{g}))\right\} \\ & \text{s.t.} \quad g_j = g_j(M, G_{-1}) \quad \forall j \in N \setminus \{i\}\end{aligned} \qquad \forall i \notin M \tag{A.85}$$

for each $M \in \mathcal{N}$. By Assumption 3, this means that $(g_i(M, G_{-1}))_{i\in N} = (g_i^{T+1}(M))_{i\in N}$, independent of G_{-1}. This, together with (A.82), implies that $\mathcal{M}$ must be the collection of all M which satisfies

$$
\begin{aligned}
i \in M \iff\ & \phi_i(\boldsymbol{g}^{T+1}(M \cup \{i\})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^{T+1}(M \cup \{i\}))\right] \\
&\quad + \delta V_i^T(M \cup \{i\}, \sigma G_{-1} + f(\boldsymbol{g}^{T+1}(M \cup \{i\}))) \\
&\geq \phi_i(\boldsymbol{g}^{T+1}(M \setminus \{i\})) - c\left[\sigma G_{-1} + f(\boldsymbol{g}^{T+1}(M \setminus \{i\}))\right] \\
&\quad + \delta V_i^T(M \setminus \{i\}, \sigma G_{-1} + f(\boldsymbol{g}^{T+1}(M \setminus \{i\}))) \\
\iff\ & u_i^{T+1}(M \cup \{i\}) + \delta v_i^T(M \cup \{i\}) \geq u_i^{T+1}(M \setminus \{i\}) + \delta v_i^T(M \setminus \{i\}),
\end{aligned} \tag{A.86}
$$

where

$$
u_i^{T+1}(M) = \phi_i(\boldsymbol{g}^1(M)) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma} f(\boldsymbol{g}^{T+1}(M)). \tag{A.87}
$$

Therefore, $\mathcal{M}$ must be independent of G_{-1}. Also, the policy functions $(a_i)_{i\in N}$ must solve

$$
\begin{aligned}
a_i(M_{-1}, G_{-1}) \in \underset{a_i \in \{0,1\}}{\operatorname{argmax}} & \left\{ \left(u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1} \right) a_i \right. \\
& \left. + \left(\mathbb{E}_\pi \left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M}) \right] - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1} \right)(1-a_i) \right\} \\
= \underset{a_i \in \{0,1\}}{\operatorname{argmax}} & \left\{ \left[u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1}) \right] a_i \right. \\
& \left. + \mathbb{E}_\pi \left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M}) \right](1-a_i) \right\},
\end{aligned} \tag{A.88}
$$

meaning that a_i is independent of G_{-1}. We can compute the associated value functions $(V_i^{T+1})_{i\in N}$ as

$$
V_i(M_{-1}, G_{-1}) = v_i(M_{-1}) - c\frac{1-(\delta\sigma)^{T+1}}{1-\delta\sigma}\sigma G_{-1}, \tag{A.89}
$$

where

$$
v_i^{T+1}(M_{-1}) := \begin{cases} u_i^{T+1}(M_{-1}) + \delta v_i^T(M_{-1}) & \text{if } \prod_{j\in N} a_j^{T+1}(M_{-1}) = 1 \\ \mathbb{E}_\pi \left[u_i^{T+1}(\tilde{M}) + \delta v_i^T(\tilde{M}) \right] & \text{otherwise.} \end{cases} \tag{A.90}
$$

Therefore, the statement of the lemma is true for $T+1$ as well. The induction argument then completes the proof. □

Now we are ready to prove the proposition. Let $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i)_{i\in N})$

be a limit equilibrium of structural model $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$. Then, by Lemma A.9, the corresponding value functions $(V_i)_{i\in N}$ are given by

$$V_i(M_{-1}, G_{-1}) = v_i^\infty(M_1) - c\frac{1}{1-\delta\sigma}\sigma G_{-1} \tag{A.91}$$

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

Since $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N}, (g_i)_{i\in N})$ is an equilibrium of $\langle \delta, N, (\Phi_i)_{i\in N}, F\rangle$, it satisfies (44), (45), and (48). It follows that $\mathcal{M}$ is the collection of all $M \in \mathcal{N}$ that satisfies

$$i \in M \iff u_i^\infty(M\cup\{i\}) + \delta v_i^\infty(M\cup\{i\}) \geq u_i^\infty(M\setminus\{i\}) + \delta v_i^\infty(M\setminus\{i\}), \tag{A.92}$$

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

$$\begin{aligned} a_i(M_{-1}) \in \operatorname*{argmax}_{a_i\in\{0,1\}} \Bigg\{ & \left[u_i^\infty(M_{-1}) + \delta v_i^\infty(M_{-1})\right] a_i \\ & + \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta v_i^\infty(\tilde{M})\right](1-a_i)\Bigg\}; \end{aligned} \tag{A.93}$$

and the functions $(v_i^\infty)_{i\in N}$ solve

$$v_i^\infty(M_{-1}) = \begin{cases} u_i^\infty(M_{-1}) + \delta v_i^\infty(M_{-1}) & \text{if } \prod_{j\in N} a_j(M_{-1}) = 1 \\ \mathbb{E}_\pi\left[u_i^\infty(\tilde{M}) + \delta v_i^\infty(\tilde{M})\right] & \text{otherwise.} \end{cases} \tag{A.94}$$

This means that $((\pi_M)_{M\in\mathcal{M}}, (a_i)_{i\in N})$ is an equilibrium of reduced-form model $\langle \delta, N, (u_i^\infty)_{i\in N}\rangle$ where the corresponding value functions are $(v_i^\infty)_{i\in N}$.