Pledge-and-Review Bargaining*

Bård Harstad

bard.harstad@econ.uio.no

July 30, 2019

Abstract

This paper develops a novel bargaining game inspired by the Paris climate-change agreement. Each party quantifies its own contribution before the set of pledges must be accepted. With uncertain tolerance for delay, each equilibrium pledge coincides with an asymmetric Nash bargaining solution. The weights on others' payoffs reflect the uncertainty, and, inefficiently, they vary pledge-to-pledge. This bargaining solution is embedded in a dynamic game with endogenous technology, participation, enforcement, and contract terms. The results are consistent with the key differences between the agreements of Kyoto (1997) and Paris (2015) as well as the development from the former to the latter. (*JEL* C78, F55, H87, Q54)

Keywords: Dynamic games, bargaining games, climate change, Paris Agreement, Kyoto Protocol

* I thank the coeditor, Debraj Ray, four anonymous referees, Geir Asheim, Scott Barrett, Ernesto Dal Bo, Faruk Gul, Jon Hovi, Steffen Lippert, Paolo Piacquadio, Santiago Rubio, Leo Simon, Håkon Sælen, Jean Tirole, David Victor, Christina Voigt, Joel Watson, Asher Wolinsky, and audiences in Adelaide (AARES pre-conference), U. Autonoma de Barcelona, U. of Barcelona, the BEET workshop at BI, UC Berkeley, UC3M, UC San Diego, CREST-Ecole Polytechnique, EIEF, ESEM 2018, University of Essex, HEC Paris, Hong Kong Baptist University, Ifo institute, London School of Economics, Manchester University, University of Melbourne, MIT, National Taiwan University, National University of Singapore, University of Oslo, UPF, Princeton University, Queen Mary University, Singapore Management University, Stanford GSB, SURED 2018, Toulouse School of Economics, and WCERE 2018. Marie Karlsen and Johannes Hveem Alsvik provided excellent research assistance and Frank Azevedo helped with the editing.

-The Paris talks were a bit like a potluck dinner, where guests bring what they can.

The New Yorker
December 14, 2015

Pledge-and-review (P&R) bargaining refers to the structure of the negotiation process adopted in Paris, December 2015. Before the countries were expected to sign the climate agreement, each party was asked to submit an intended nationally determined contribution (INDC). For most developed countries, the INDC specified unconditional cuts in the emissions of greenhouse gases being effective from 2020 to 2025 (or to 2030). As an illustration, Table 1 presents the pledges for a sample of developed countries.[1]

Parties	Australia	Canada	EU	New Zealand	Norway	Russia	Switzerland	USA
Pledge	26-28%	30%	40%	30%	40%	25-30%	50%	26-28%

Table 1. The pledges specify emission cuts relative to nationally chosen baselines.

Every five years the parties shall review and make new pledges for another five-year period (Paris Agreement Art. 4.9).

This bargaining procedure is remarkably different from the one used for the Kyoto Protocol of 1997. There, a "top-down" approach was used to pressure governments to cut emissions by (on average) five percent relative to the 1990 levels.[2] By comparison, P&R has been referred to as a "bottom-up" approach since countries themselves determine how much to cut nationally, without making these cuts conditional on other countries' emissions cuts.[3] In political science, Keohane and Oppenheimer (2016:142) fear that: "Many governments will be tempted to use the vagueness of the Paris Agreement, and the discretion that it permits, to limit the scope or intensity of their proposed actions." Not surprisingly, leading economists, such as Gollier and Tirole (2015), conclude that: "The pledge-and-review strategy is completely inadequate."

The P&R procedure also differs from traditional models of bargaining. To shed light on the procedure, this paper introduces and analyzes a novel bargaining game. I provide a characterization of the bargaining

[1] The baseline year is 1990 for the European Union, Russia, and Switzerland, while it is 2005 for Australia, Canada, New Zealand, and the United States. Article 4.4 of the Paris Agreement encourages "economy-wide absolute emission reduction targets," although several developing countries state pledges in terms of emission per GDP and some of these are conditional on receiving transfers. The official list is at http://www4.unfccc.int/ndcregistry but for an overview see http://cait.wri.org/indc/#/.

[2] For example, Bodansky and Rajamani (2017:11) write: "In essence, the Kyoto Protocol was the product of mutual concessions... The USA accepted a much stronger target (minus 7% from 1990 levels) than it had wanted..."

[3] According to the Paris Agreement (Art. 4.2): "Each Party shall prepare, communicate and maintain successive nationally determined contributions that it intends to achieve." So: "Now, instead of setting commitments through centralized bargaining, the Paris approach sets countries free to make their own commitments" (Victor (2015)), and: "Instead of pursuing a [Kyoto-style] top-down agreement with mandated targets, [the organizers] have asked every country to submit a national plan that lays out how and by how much they plan to reduce emissions in the years ahead" (*The New York Times* (Nov. 28, 2015)).

outcome and relate it to the asymmetric Nash bargaining solution (ANBS). The bargaining outcome is thereafter embedded into a dynamic game where parties can both contribute to a public good and invest in their future capacity to contribute (e.g., renewable energy sources). I investigate how the bargaining procedure influences contributions, investments, participation, compliance, and payoffs.

Remarkably, the difference in bargaining procedure (Fact 1), the way I model it, can rationalize four other stylized facts on how the Paris Agreement diverges from the Kyoto Protocol: (2) While relatively few (37) countries faced binding emission targets under Kyoto, nearly every country in the world contribute to the Paris Agreement. (3) The Kyoto Protocol was endogenously chosen in the 1990s, but the participants preferred the P&R procedure in the 2010s. (4) The commitments under the Kyoto Protocol were "legally binding," but the INDCs are not. (5) Despite all these differences between the two agreements, the commitment period length agreed on was five years for both treaties. My results, in comparison, are the following.

(1) *The Bargaining Game.–* The novel feature of P&R bargaining, the way I formalize it, is that each party is permitted to propose its own individual contribution only, rather than a vector of contributions for all the parties. I assume for simplicity (and because there is no formal sequential procedure in climate negotiations) that all parties propose their pledges simultaneously. If some parties find the vector of pledges unacceptable, the procedure starts again.[4] With complete information, the trivial equilibrium of the game coincides with the noncooperative (or business-as-usual) outcome, where every party simply makes a pledge that maximizes that party's utility. However, with sufficiently noisy shocks on the other parties' willingness to decline and delay the agreement, I show that each party's equilibrium contribution level coincides with the quantity that maximizes an asymmetric Nash product, where the weights on other parties' payoffs reflect the extent of uncertainty as well as how the shocks are correlated. Inefficiently, different parties apply different weights. The relative weights on others' payoffs are less than 1/2 for single-peaked and symmetric shock distributions, and they are close to zero when the variance of each shock is small.

Note that this bargaining game is quite general and it might be useful for several other applications beyond climate negotiations. For example, the game can describe a situation in which multiple business partners must negotiate a package, and where each partner is recognized as an expert and as the proposer for only a single dimension of the package: one partner describes the product quality, another offers a strategy for advertisements, while a third manages a set of retailers, for instance. In such expert meetings, it might be unrealistic to assume that a single partner is capable of proposing and describing all the details of interest, as is normally assumed in bargaining theory.

[4]The interpretation is that if n sovereign countries are about to contribute, the overall agreement must be acceptable by all these n countries. Before the 2009 Copenhagen negotiations, when P&R was first attempted, many countries had submitted pledges. However: "Objections by a small group of countries (led by Bolivia, Sudan, and Venezuela) prevented the Copenhagen conference from 'adopting' the Accord ... as a COP decision, which requires consensus (usually defined as the absence of formal objection)" (Bodansky (2010:231; 238)). As a consequence, negotiations were delayed for years.

Although the bargaining game arguably has alternative applications, the assumptions are motivated by recent climate negotiations and thus the theory should be investigated and confronted with Facts (2)-(5), described above. For this investigation, I present a dynamic game in which parties over time contribute to a public good (by cutting emission) as well as invest in their future capacities to contribute (e.g., they invest in renewable energy or green technology). The pledges quantify emissions targets and these are revised and updated periodically. Naturally, the small weights on others' payoffs, associated with P&R, implies that targets are not very ambitious, and thus investments, as well as welfare, are lower than they would have been under the Nash bargaining solution (NBS), which is often used to describe the outcome of the Kyoto Protocol (see the literature review below). This negative result rationalizes the critique of the P&R procedure.

(2) *Participation.*– The negative result is reversed, however, when the decision to participate in the bargaining game is endogenous. Since not much is expected from the participants (when the weights on others' payoffs are small), it is not that costly for a party to participate, and this explains why the equilibrium coalition size is larger with P&R bargaining than with the NBS. The larger coalition size means, in equilibrium, that the *sum* of contributions is larger, the aggregate investments are larger, and so is welfare.

(3) *Institutional Design.*– The comparison of bargaining procedures is more interesting when we take into account that there is an upper boundary ($\overline{n}$) for the number of potential members and that this constraint might bind. Furthermore, with minimum participation constraints, or when the parties are heterogeneous in that a number ($\underline{n}$) of them will participate regardless of the game, then the narrow-but-deep agreement under the NBS can be more attractive. With these constraints, P&R is preferred if and only if $\overline{n}$ is large while $\underline{n}$ is small, I show.

This result is in line with the development from Kyoto to Paris: In the 1990s, there were a large number of developing countries that could not be expected to contribute much to a global climate policy. Over the last twenty years, some of these have become emerging economies that potentially have important roles to play. The number of relevant potential parties, $\overline{n}$, has therefore increased. During the same period, seven countries that initially signed the Kyoto Protocol declared that they did not intend to contribute to Kyoto's second commitment period (IPCC (2014:1025)). This can be interpreted as a smaller $\underline{n}$. Either (or both) of these developments makes P&R more attractive. Thus, the theory is not inconsistent with the fact that the parties preferred the Kyoto Protocol in the 1990s, but P&R in the 2010s.

(4) *Compliance:* If the parties cannot commit to future actions, the pledges must be self-enforcing. As in the repeated games literature, one may require a party to be willing to comply under the threat that others can retaliate. The P&R bargaining outcome is more likely to be self-enforcing than the NBS, I show. The simplest intuition for this result is that when the pledges are less demanding, the temptation to defect is small. If the bargaining outcome is characterized by the NBS, in contrast, the parties might find it necessary to motivate compliance by raising the political cost of defection. In practice, the political

cost can be raised by requiring the emissions cuts to be "legally binding" or enforced by other punitive measures —and both these methods are indeed employed by the Kyoto Protocol, but not by the Paris Agreement.

(5) *Terms of contract:* The optimal contract duration in this model results from a novel trade-off: A long-term contract is unattractive because, after the parties have invested in new capacity, it becomes optimal to negotiate still more ambitious pledges. A short-term contract, however, creates a hold-up problem when the parties anticipate how their investments will influence the next bargaining outcome. The optimal term trades off these two concerns, but this trade-off is shown to be the same under P&R as under NBS, and independent of the number of participants. Thus, if a five-year commitment period was optimal under the Kyoto Protocol, it is indeed optimal also for the Paris Agreement, according to this result. In other words, the theory is consistent with the similarity as well as the differences between the agreements.

Literature.—So far, the P&R procedure has only been informally described. Falkner (2016:1120) writes: "By subjecting domestically determined mitigation pledges to the international review mechanism, the Paris Agreement ensures that the gap between the required level of action and the total sum of national measures becomes the subject of international policy deliberation and coordination."

Keohane and Oppenheimer (2016:148) establish that parties are "seeking to do enough to induce action by others and avoid sanctions against themselves, but not so much that they bear heavy burdens that seriously affect economic growth." They continue (p. 149): "What is less clear is whether the resulting deals will [help] the world limit climate change. We can imagine high-level equilibria of these games that would do so. These equilibria would induce substantial cuts in emissions [but] we can also imagine low-level equilibria...that enables both sides to pursue essentially business as usual." Thus, the Paris Agreement "opens the door to progress on climate but does not assure it" (p. 150).

These observations are consistent with the theory I provide. My formal analysis combines the need for a unanimous support with the fact that pledges are nationally determined: A contributing country can always decide not to deliver if the opponents' pledges are unsatisfying. That outcome would not be renegotiation-proof, however, since opponents then have incentives to improve on the pledges.

By showing that each individual contribution maximizes an asymmetric Nash product, I contribute to the "Nash program," aimed at finding noncooperative games implementing cooperative solution concepts. The Nash demand game (Nash (1953)) intended to implement the NBS, axiomatized by Nash (1950). There is a large subsequent literature investigating the extent to which the Nash demand game implements the NBS.[5] The alternating-offer bargaining game by Rubinstein (1982) also implements the NBS, as shown by Binmore et al. (1986). Although there can be multiple equilibria with more than two players (Sutton (1986); Osborne and Rubinstein (1990)), the NBS is the unique equilibrium if we impose stationarity or

[5] See Binmore et al. (1992), Abreu and Gul (2000), or Kambe, 2000). Some contributions allow for uncertainty (Binmore (1987); Carlsson (1991); Andersson et al. (2018)), as I do for P&R.

reasonable consistency conditions (Krishna and Serrano (1996); Chae and Yang (1994); Asheim (1992)).[6]

For these reasons, the NBS has reasonably been assumed to characterize bargaining outcomes in applied theory, such as in analyses of climate agreements (see the surveys by Calvo and Rubio (2012) and Capparos (2016) or most of the papers discussed below). This assumption is no longer justifiable after the switch to P&R, I show.

My contribution to this literature and to the Nash program is to show that each equilibrium pledge maximizes an asymmetric Nash product. The weights reflect differences in the discount rates in an intuitive way, but also the extent of uncertainty in shocks and the correlation in shocks across the parties. Since the equilibrium weights vary from one party's pledge to another's, the set of contributions is not Pareto optimal.

The dynamic climate change game below relies on standard assumptions shared by Dutta and Radner (2004; 2006), Harstad (2012; 2016), and Battaglini and Harstad (2016), but the present paper is unique in focusing on the bargaining procedure. In contrast to Dutta and Radner (2019), I rule out side payments. The coalition formation game is a standard one when modeling collusion (d'Aspremont et al. (1983); Bloch (2018)) or environmental coalitions (Hoel (1992); Carraro and Siniscalco (1993); Barrett (1994)). The predicted small coalition size is referred to as a "paradox" by Kolstad and Toman (2005) and Nordhaus (2015). With P&R, however, the coalition is more realistically sized.

There is already a well-known trade-off between treaties that are narrow-but-deep vs. broad-but-shallow; see Schmalensee (1998), Barrett (2002), Aldy et al. (2003), or Finus and Maus (2008). The assumptions in this literature have been criticized by political scientists such as Gilligan (2004) and Bernauer et al. (2013), but the present paper shows how the trade-off arises naturally from differences in bargaining procedures. In contrast to Schmalensee (1998), who recommended "broad, then deep," my model can rationalize the reverse, factual development from the deep-but-narrow Kyoto Protocol to the broad-but-shallow Paris Agreement.

More broadly, when investigating the preferences for bargaining procedures under endogenous participation, this paper complements the literatures on how endogenous entry influences the optimal design of mechanisms (McAfee (1993)) such as auctions (Bulow and Klemperer (1996; 2009), for example), and on how the bargaining game influences externalities and efficiency in trade negotiations (Bagwell and Staiger (1999)).

Outline.– The next section formalizes P&R bargaining and characterizes the outcome in Theorems 1 and 2. Section II embeds the bargaining outcome in the climate change game: P&R leads to lower contributions, capacity investments, and welfare, but this negative finding is reversed when participation

[6] In contrast, the ANBS (axiomatized by Harsanyi and Selten (1972); Kalai (1977); Roth (1979)) characterizes the outcome if there are asymmetric discount rates, recognition probabilities, or voting rules (Miyakawa (2008); Britz et al. (2010); Laruelle and Valenciano (2008)). In another related paper, Yildiz (2003) finds an efficient (competitive) allocation when a proposer can only propose a price. Yildiz assumes sequential offers and that the other party can subsequently select any trade quantity given this price.

is endogenous in Section III. Section IV argues that the theory can explain why participants have switched from preferring the NBS to preferring P&R. Section V shows that the P&R pledges are more likely to be self-enforcing than the NBS would be, and Section IV confirms that the optimal commitment period is the same for the two bargaining procedures. The robustness section presents ten extensions and explains why the results survive in all of them. Section IIX concludes. Appendix A proves the theorems, Online Appendix B generalizes them, while Online Appendix C contains other proofs.

I. A Model of Pledge-and-Review Bargaining

This section describes a novel bargaining game and characterizes its outcome. The section can be read independently from the other sections, since the model here may have alternative applications besides climate negotiations. As mentioned, the bargaining game might be appropriate when a number of business partners are negotiating a multidimensional deal, and each partner has expertise and is making the proposal on a single dimension of the package (such as quality, price, delivery time, etc.). The main new feature of the game is that each party is recognized as being responsible for proposing only one dimension of the agreement, even though payoffs depend on the entire vector.

A. *The Bargaining Game*

There are n parties, each endowed with a payoff function $U_i : \mathbb{R}^n \to \mathbb{R}$, $i \in N = \{1, ...n\}$. The bargaining game starts when every party i simultaneously proposes its own dimension, or contribution, $x_i \in \mathbb{R}$. After they observe $\mathbf{x} = (x_1, ..., x_n)$, each party must decide whether to accept. If everyone accepts, each party i receives the payoff $U_i(\mathbf{x})$ and the game ends. If one or more parties declines $\mathbf{x}$, the game is played again. Unanimity is required among the parties that contribute positively, but the n parties can be a subset of all players (see Sections III and IV). A party is assumed to accept whenever indifferent.

Figure 1 presents $\Delta \in (0, \infty)$ as the lag between proposals and acceptance decisions. However, it is equivalent to instead let Δ be the lag between rejections and new offers, as often assumed in bargaining models.

Section II endogenizes the payoff $U_i(\mathbf{x})$ and shows why it can represent the continuation value in a dynamic game in which the parties pledge for a finite number of periods; Section V shows that it isn't even necessary that the parties can commit or that the x_i's be enforced, as long as the x_i's influence continuation payoffs. For now, simply assume (for tractability) U_i to be concave and continuously differentiable. Concavity is natural when a party begins with the most cost-effective types of contributions. Both U_i and x_i are measured relative to the default outcome, which thus is normalized to zero. Furthermore, I start by making the *additional assumptions* $\partial U_i(\cdot)/\partial x_i < (>) 0$ for $x_i > (<) 0$, and $\partial U_j(\cdot)/\partial x_i > 0$, $j \neq i$, so

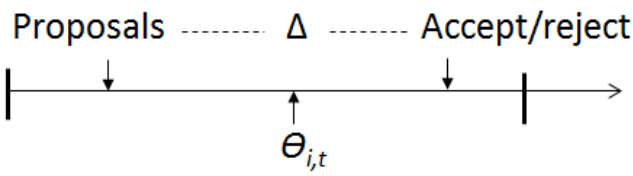

Figure 1: *Bargaining in each period.*

that the x_i's can be interpreted as contributions to a public good beyond the individually rational level. However, the Appendix proves Theorem 2 and a generalization of Theorem 1 without these additional assumptions: See the below "Remarks on Generality."

Party j's discount factor between time t and $t+\Delta$ is $\delta_{j,t}^{\Delta} \leq 1$, but here it is more convenient to refer to the "discount rate" $\rho_{j,t} \equiv \left(1-\delta_{j,t}^{\Delta}\right)/\Delta$.[7] Thus, party j receives $\left(1-\rho_{j,t}\Delta\right)U_j\left(\mathbf{x}^*\right)$ by declining an offer if $\mathbf{x}^*$ can be expected next period. Given $\mathbf{x}^*$, j prefers to accept $\mathbf{x}$ now if:

$$U_j\left(\mathbf{x}\right) \geq \left(1-\rho_{j,t}\Delta\right)U_j\left(\mathbf{x}^*\right). \tag{1}$$

Throughout the paper, I will restrict attention to MPEs (Section V discusses how this restriction can be relaxed). Before proceeding, it will be useful to establish the following benchmark result. If information were perfect and $\rho_{i,t} = \rho_i > 0$, there exists a "trivial equilibrium" consisting of the acceptance strategies (1) and a vector $\mathbf{x}^* = \mathbf{0}$. For any equilibrium candidate in which $U_j\left(\mathbf{x}^*\right) > 0 \forall j$, party i can suggest x_i slightly different from x_i^* without violating (1). Thus, x_i^* must coincide with i's preferred level, $x_i^* = \arg\max_{x_i} U_i\left(x_i, \mathbf{x}_{-i}^*\right)$, which is zero under the additional assumptions. This observation supports the skepticism to P&R, described in the introduction.

In reality, a party is unlikely to know precisely the condition under which an offer will be accepted. In other words, if $\mathbf{x} \neq \mathbf{x}^*$, the probability that (1) holds may be positive but less than one. Such uncertainty can be derived from shocks in the utility functions, beliefs over the delays following rejections, or the impatience, as originally proposed by Rubinstein (1985). The Online Appendix shows that any such uncertainty can imply that $\mathbf{x} \neq \mathbf{x}^*$ will be rejected with some probability. To fix ideas, this section provides a version of Rubinstein's idea:[8] Assume henceforth that the exact discount rates decision makers will apply for the next period (i.e., the $\rho_{j,t}$'s) are affected by shocks. After all, a policy maker's tolerance for delay is influenced by a number of (con)temporary domestic policy or economy issues that compete for the policy maker's attention. (The impatience may also depend on the current probability of remaining in office, as in Harstad (2020)). Since no-one can foresee all these issues when the pledges are made, perhaps several months in advance, Figure 1 illustrates how the shocks are realized and observed by

[7] If the real discount rate is $\widetilde{\rho}_{j,t}$, the discount factor is $e^{-\widetilde{\rho}_{j,t}\Delta} = \delta_{j,t}^{\Delta}$, so $\rho_{j,t} \equiv \left(1-e^{-\widetilde{\rho}_{j,t}\Delta}\right)/\Delta$ approaches $\widetilde{\rho}_{j,t}$ when $\Delta \to 0$. I refer to $\rho_{j,t}$ as the discount rate even though the identity holds only in the limit.

[8] While I follow Rubinstein (1985) by letting the discount rate be uncertain, there are many differences between our two approaches.

everyone after the offers but before acceptance decisions are made.[9]

Formally, write $\rho_{i,t} = \theta_{i,t}\rho_i$, where ρ_i is i's expected discount rate while $\theta_{i,t}$ is a shock with mean 1. The shocks are jointly distributed with pdf $f(\theta_{1,t}, ..., \theta_{n,t}) \in (0, \infty)$ on support $\prod_{i \in N} \left[0, \overline{\theta}_i\right]$, i.i.d. at each time t, and the marginal distribution of $\theta_{i,t}$ is $f_i(\theta_{i,t}) \equiv \int_{\Theta_{-i}} f(\theta_{1,t}, ..., \theta_{n,t})$, where $\Theta_{-i} \equiv \prod_{j \neq i} \left[0, \overline{\theta}_j\right]$. The shocks can be correlated across the parties, but since the vector of shocks is i.i.d. over time, the game is stationary. Thus, it continues to be meaningful to restrict attention to MPEs.

B. *The Bargaining Solution*

Attention will be restricted to pure-strategy offers. An MPE is then a vector of offers, x^*, combined with a set of strategies for the acceptance stage. The acceptance strategies are straightforward to characterize: If $U_j(\mathbf{x}^*) > 0$, then j accepts $\mathbf{x}$, after learning $\theta_{j,t}$, if and only if:

$$U_j(\mathbf{x}) \geq \left(1 - \theta_{j,t}\rho_j\Delta\right) U_j(\mathbf{x}^*) \Rightarrow \theta_{j,t} \geq \frac{U_j(\mathbf{x}^*) - U_j(\mathbf{x})}{\rho_j \Delta U_j(\mathbf{x}^*)}. \tag{2}$$

When $\theta_{j,t}$ is drawn from a continuous distribution, the probability that j accepts will be continuous in x_i. On the one hand, this continuity can motivate positive contributions: x^* can be supported as a "nontrivial" MPE if the marginal benefit for i by slightly reducing x_i is outweighed by the risk that at least one party might be sufficiently patient to decline the offer and wait for x^*.

On the other hand, party i is tempted to take some risk by reducing x_i, especially when x_i^* is costly to i. But note that there cannot be delay on the equilibrium path: If party i prefers to take risk by offering less today, everyone expects party i will find it optimal to offer less also in the future, given the stationarity of the game. Therefore, no-one gains from rejecting the stationary offer today. The offer will thus not be risky at the equilibrium path: $\mathbf{x} = \mathbf{x}^*$ will be proposed and (2) implies that, as a result, the proposal will be accepted without delay with probability 1. However, the temptation to take (further) risks limits how large x_i^* can be, as shown in the following theorem.

THEOREM 1**:** *Consider a nontrivial MPE in which $U_i(\mathbf{x}^*) > 0 \forall i$. For every $i \in N$:*

$$\begin{aligned} x_i^* &\leq \arg\max_{x_i} \prod_{j \in N} \left(U_j\left(x_i, \mathbf{x}_{-i}^*\right)\right)^{w_j^i}, \text{ where} \\ \frac{w_j^i}{w_i^i} &= \frac{\rho_i}{\rho_j} f_j(0) \, \mathrm{E}\left(\theta_{i,t} \mid \theta_{j,t} = 0\right), \forall j \neq i. \end{aligned} \tag{3}$$

Before explaining and discussing the result, note that the above reasoning does not limit how *small* the equilibrium x_i^*'s can be, because there is no point for i to contribute more than x_i^*, whatever the

[9] This timing is not unreasonable: Since there can be a substantial lag between offers and acceptance decisions, it is natural that policy makers in the meantime learn about how urgent it is for them to conclude the negotiations, or about the attention they instead must give to other policy and economic issues.

equilibrium x_i^* is. (As noted, an MPE vector $\mathbf{x}^*$ is always acceptable given that $\theta_{j,t} \geq 0$.) There can thus be multiple MPEs. To obtain sharper results, it is common to require the equilibrium to be robust to small trembles. If there is a risk that even $\mathbf{x}^*$ will be declined, then i may prefer to reduce the risk by increasing x_i. This logic holds if we impose the following version of trembling-hand perfection.[10]

Definition of Local Perfection: *Consider a perturbed game in which, when the vector of submitted offers is* $\mathbf{x}$, $\mathbf{x}+\epsilon s_t$ *is realized and observed, where* s_t *is a vector of* n *shocks distributed i.i.d. over time, with bounded support, and with strictly positive density on a neighborhood of* $\mathbf{0}$. *Say that* $\mathbf{x}^*$ *is* locally perfect *if* $x_i^* = \lim_{\epsilon \to 0} x_i^*(\epsilon) \, \forall i \in N$, *where* $\mathbf{x}^*(\epsilon)$ *is an equilibrium of the perturbed game.*

THEOREM 2**:** *Consider a locally perfect MPE. Inequality (3) binds for every* $i \in N$.

As a comparison, in the ANBS, each x_i maximizes the *same* asymmetric Nash product:

$$x_i^A = \arg\max_{x_i} \prod_{j \in N} \left(U_j \left(x_i, \mathbf{x}_{-i}^A \right) \right)^{w_j^i}, \tag{4}$$

for some (exogenously) fixed weights. In this case, the vector $\mathbf{x}$ will be Pareto optimal.

Also when (3) binds, following P&R bargaining, the equilibrium x_i^* maximizes an asymmetric Nash product, but *different parties* apply *different weights*. Thus, the set of x_i^*'s is not Pareto optimal. In particular, if $w_j^i / w_i^i < 1$ for every (i,j), $j \neq i$, then it is possible to make every party better off by increasing all contributions relative to $\mathbf{x}^*$.

Theorem 1 also endogenizes the weights[11] and shows how they depend on three things. First, the weight on j's utility is larger if j is expected to be patient relative to i. This is natural (and in line with the papers on bargaining mentioned in Footnote 6): When j is patient, j is more tempted to reject an offer that is worse than what one can expect in the next period, and thus i finds it too risky to reduce x_i, especially when i is likely to be impatient.

Second, the weight on j's payoff is larger when there is more uncertainty regarding j's shock. Of importance is especially the marginal likelihood that j's discount rate is close to 0, so that even a small reduction from x_i^* involves some risk that j will decline. If the shock were bounded away from zero, then $w = 0$ (unless $\Delta \to 0$, as shown in Online Appendix B).

Third, the weight on j's payoff is less for a small $\mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0)$, which measures i's expected shock on the discount rate given that j's shock is small. Intuitively, if i can be expected to have a small discount rate exactly when j has so, then it matters less that j declines an offer in this circumstance. When the delay matters less, i does not find it necessary to offer a lot. This suggests that a party i may pay less attention to the payoffs of those who face shocks that are positively correlated with i's shock.

[10] Simon (1987) offers a similar (more general) definition when he extends to infinite games the reasoning by Myerson (1978), i.e., that the trembles in Selten (1975) should be smaller for costlier errors. Simon and Stinchcombe (1995) show that this reasoning can be adapted to infinite games in multiple ways.

[11] Theorem 1 only endogenizes the relative weights, w_j^i / w_i^i, but this is sufficient since $\arg\max_{x_i} \prod_{j \in N} \left(U_j \left(x_i, \mathbf{x}_{-i}^* \right) \right)^{w_j^i}$ stays unchanged if every weight w_j^i is multiplied by the same positive number.

C. *Simplifications and Generalizations*

Theorem 2 has several important consequences.

COROLLARY 1**:** *Suppose all parties have the same preferences and shock distributions,* $f_i = f_\theta \forall i \in N$. *(i) In a locally perfect MPE, the equilibrium offers are:*

$$\begin{aligned} x_i^* &= \arg\max_{x_i}[U_i\left(x_i, \mathbf{x}_{-i}^*\right) + w\sum_{j\neq i} U_j\left(x_i, \mathbf{x}_{-i}^*\right)], \text{ where} \\ w &= f_\theta\left(0\right) E\left(\theta_{i,t} \mid \theta_{j,t} = 0\right) \ \forall i, j. \end{aligned}$$

(ii) The weight is $w \leq \frac{1}{2}\ \forall i, j \neq i$, *if* $f_\theta\left(\cdot\right)$ *is single-peaked and symmetric and shocks are uncorrelated.*[12]

Combining the two parts, the corollary suggests that the weight on other parties' payoffs is less than $\frac{1}{2}$ of the weight on i's payoff when x_i is proposed, if the preferences are similar. If uncertainty vanishes, such that the pdf $f_\theta\left(.\right)$ concentrates around its mean, then $f_\theta\left(0\right) \to 0$, so $w \to 0$, and x_i^* must approach the level in the trivial equilibrium, as when there is no bargaining.

Example E.– As an illustration, consider the situation in which i benefits linearly from the others' contributions, while each party's contribution cost is quadratic,

$$U_i\left(x_i, \mathbf{x}_{-i}^*\right) = \alpha\sum_{j\neq i} x_j - \beta x_i^2/2 + \gamma, \tag{E}$$

for some positive constants α, β, and γ. Corollary 1 implies:

$$x_i^* = w\left(n-1\right)\alpha/\beta.$$

The grey area in Figure 2 illustrates the set of equilibria permitted by Theorem 1 when $n = 2$ and $w = \frac{1}{2}$. (The pair of dashed curved lines corresponds to $w = 1$, and the NBS is illustrated by x_1^S and x_2^S.) In this example, it is easy to check that all equilibria satisfying (3) with strict inequalities are Pareto dominated by the equilibrium in which the inequalities bind (i.e., x_1^* and x_2^*) if $w < \sqrt{3} - 1 \approx 0.73$. Thus, focusing on equilibria that are not Pareto dominated might in some cases replace the restriction to local perfection.[13]

Remarks on Generality.– (i) Note that the theorems do not require Δ to be small. (ii) The inequality in (3) will bind for nontrivial equilibria even without requiring local perfection, if we, as an alternative, introduce trembles on the supports of the discount rates.[14] (iii) Although I above made the additional

[12] To see this, note that if $f_\theta\left(0\right) > 1/2$, then, when $f_\theta\left(\cdot\right)$ is single-peaked and symmetric around the mean of one, $\int_0^2 f_\theta\left(\theta_j\right) d\theta_j > 1$, violating the definition of a pdf. If the shocks are not correlated, then $E(\theta_{i,t} \mid \theta_{j,t} = 0) = 1$.

[13] I thank Asher Wolinsky for making this observation.

[14] To be precise, Theorem 2 holds without imposing local perfection if we instead assume that the support of $\theta_{j,t}$ is $\left[\epsilon\underline{\theta}_j, \overline{\theta}_j\right]$, rather than $\left[0, \overline{\theta}_j\right]$, where $\underline{\theta}_j < 0$, $\epsilon > 0$, and $\epsilon \downarrow 0$. The interpretation of a negative discount rate may be that, in some circumstances, a party prefers to delay signing agreements because of other urgent economic/policy issues that require the decision makers' attention. If the lower boundaries approach zero in the limit (as $\epsilon \to 0$), then there will be no delay on the equilibrium path.

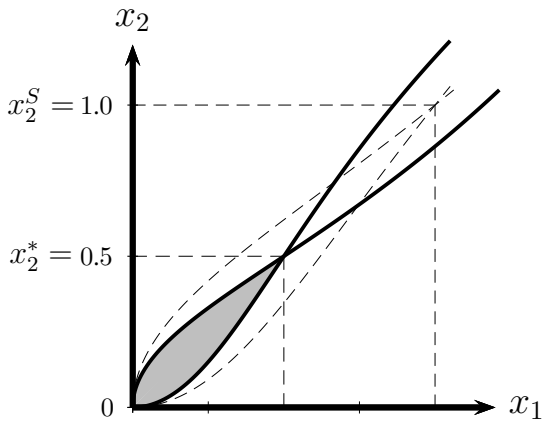

Figure 2: *There are multiple equilibrium contribution levels in Example E, but they are all smaller than both the efficient level ($\mathbf{x}^S$) and the locally perfect equilibrium ($\mathbf{x}^*$).*

assumptions that $\partial U_i(\cdot)/\partial x_i < 0$ and $\partial U_j(\cdot)/\partial x_i > 0$, $j \neq i$, these assumptions are not needed for Theorem 2, and a generalization of Theorem 1 is proven in Appendix A without these additional assumptions. (iv) Online Appendix B shows how P&R leads to positive contributions also when $f_j(0) = 0$ if $\Delta \to 0$.

Remark on Sufficiency.– Condition (3) is necessary for $\mathbf{x}^*$ to be an MPE, but it may not be sufficient. Whether the second-order condition for an optimal deviation for i holds globally depends on the f_j's. If $n = 2$, a sufficient condition for the second-order condition to hold is that f_j be weakly increasing, as when $\theta_{j,t}$ is uniformly distributed, for example.[15]

D. *Remarks on Nash's Demand Game and Bargaining Solution*

The P&R bargaining outcome is in stark contrast to the Nash bargaining solution, predicting that the x_i's would follow from (4) with $w_j = 1 \forall j$. The NBS is frequently used to describe multilateral bargaining outcomes, such as the Kyoto Protocol, partly because it results from standard noncooperative bargaining games, including the Nash demand game (NDG). Interestingly, this result follows as a corollary to Theorem 2 because P&R is a generalization of the NDG.[16]

In the NDG, each player is demanding an ex post utility level or, equivalently, a variable (x_i) that dictates i's ex post utility, $v_i(x_i)$. The vector of demands is feasible with probability $p(\mathbf{x})$, so i's expected payoff is (see Binmore (1987)):

$$U_i(x_i, \mathbf{x}_{-i}) = v_i(x_i)\, p(\mathbf{x}). \tag{5}$$

With (5), a binding (3) can be rewritten as follows:

[15] It is then easy to see from equation (29) in the Appendix that the second-order condition holds.

[16] I am grateful to Jean Tirole and a referee for the motivation for this subsection.

COROLLARY 2: *Consider pledge-and-review bargaining and suppose every payoff is given by (5). If* x^* *is a locally perfect MPE, then*:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \prod_{i\in N} v_i\left(x_i\right)^{\varrho_i} p\left(\mathbf{x}\right)^{\varpi} \tag{6}$$

$$= \arg\max_{\mathbf{x}} \prod_{i\in N} v_i\left(x_i\right)^{\varrho_i} \text{ s.t. } p\left(\mathbf{x}\right) = p\left(\mathbf{x}^*\right), \textit{ where} \tag{7}$$

$$\varrho_i = \frac{w_i^i / \sum_{j\in N} w_j^i}{\sum_{k\in N}\left(w_k^k / \sum_{j\in N} w_j^k\right)} \text{ and } \varpi = \frac{1}{\sum_{k\in N}\left(w_k^k / \sum_{j\in N} w_j^k\right)}.$$

Outcome (7) coincides with the NBS if the weights (w_j^i/w_i^i) are equal and if the uncertainty on the feasibility constraint vanishes, in the sense that $p\left(\mathbf{x}\right)$ is close to 0 or 1 for almost every $\mathbf{x}$. Then, it is intuitive that $\mathbf{x}^*$ must be close to an $\mathbf{x}$ that ensures $p\left(\mathbf{x}\right) \approx 1$, and so the constraint $p\left(\mathbf{x}\right) = p\left(\mathbf{x}^*\right)$ in (7) simply requires $\mathbf{x}$ to be feasible. The mapping from the NDG to the NBS is in several respects generalized by Corollary 2: The mapping holds if there is a finite delay when $\mathbf{x}$ is not approved unanimously, before the game can be played again; the result also permits stochastic discount rates and shows that these are irrelevant for the mapping if parties are symmetric. The intuition for why w is irrelevant is that, given the sharp threshold characterized by $p\left(\mathbf{x}\right)$ when uncertainty vanishes, i's preferred x_i coincides with the efficient level, given the other x_j's.

When the weights (w_j^i/w_i^i) are heterogeneous, (7) characterizes an ANBS: The bargaining power index (ϱ_i) is larger for those parties who are likely to be patient or who face less uncertainty regarding the opponents' discount rates, it can be shown.

But when the uncertainty on the feasibility constraint vanishes, U_i becomes discontinuous in x_j, violating the assumption in Section I(A). If each $U_i\left(\mathbf{x}\right)$ is instead continuous in every x_j, as when the uncertainty on the feasibility constraint is *not* vanishing, then, with P&R bargaining, $\mathbf{x}^*$ does not coincide with the NBS. Instead, (6) shows that i places less weight on the (collective) risk if $w_j^i/w_i^i < 1$.

Similarly, if the parties do not negotiate utility levels, but contribution levels, as in Example E, then $U_i\left(\mathbf{x}\right)$ is likely to be continuous in all the x_j's. This continuity makes the P&R outcome inefficient.

II. A Dynamic Contribution Game

Section I took as given the players' objective functions (the $U_i\left(\cdot\right)$'s). To better understand the implications of pledge-and-review, the next subsection presents a tractable climate change model to endogenize how these payoffs depend on pledges. Subsection B solves the model and Subsection C explains how the above P&R outcome can be combined with this model to determine the pledges. Section III endogenizes participation, Section IV endogenizes the choice of bargaining game, and Section VI derives the optimal commitment period length. To be consistent with the analysis above, I continue to limit attention to MPEs, but Section V discusses how this restriction can be relaxed. Since the purpose of this

analysis is to rationalize Facts 1-5 on Paris vs. Kyoto, rather than to focus on generality, it is sufficient to rely on linear-quadratic per-period utility functions. Section VII is nevertheless discussing how the model can be generalized in ten different ways without affecting the results.

A. *The Climate Policy Model*

The model describes a situation in which the parties can contribute to a public good as well as invest in their future capacities to contribute. In equilibrium, the negotiated contribution levels will influence how much the parties will invest, but past investments will also influence the future contribution levels. Although the model can be applied to other public good settings, it fits especially well to analyze climate policies. As required by the Paris Agreement (Art. 4.9): "Each Party shall communicate a nationally determined contribution every five years." Apparently, "The idea is that this short time frame would give countries the opportunity to regularly capture scientific and technological developments in their official targets."[17] The Stern Review (2006) also pointed out that new technology would be crucial to mitigate climate change. However, the treaties say that "technology needs must be nationally determined, based on national circumstance and priorities" (§114 of the 2010 Cancun Agreement). For the model to be consistent with this practice, emissions cuts are negotiable and contractible, while technology investments are not. (The assumption is relaxed in Section VII.)

In each period t, the utility for a party is the sum of three parts. First, if each party i contributes or abates the quantity $q_{i,t}$, the sum of abatements has the value $a\sum_{i\in N} q_{i,t}$ to each party. This linearity assumption is made for simplicity, but it is common also because it is a reasonable approximation when it comes to climate change.[18] An additional benefit of this linearity is that we can easily allow for a stock of greenhouse gases that accumulates over time, without changing the analysis, since a can be interpreted as the present discounted cost of emitting another unit of emission into the atmosphere, when we anticipate that this unit may contribute to climate change for decades.[19]

The second term in the utility function specifies the cost of contributing to the public good. For example, suppose a country can consume energy from both fossil fuels ($g_{i,t}$) and renewables ($R_{i,t}$). If the total consumption of energy is less than i's bliss point, $B_{i,t}$, then i experiences a disutility that is quadratic in the difference: $\frac{b}{2}\left(B_{i,t}-[g_{i,t}+R_{i,t}]\right)^2$. This disutility can be written as $\frac{b}{2}\left(q_{i,t}-R_{i,t}\right)^2$, when $q_{i,t}$ represents a cut in emissions relative to i's bliss point (i.e., when $q_{i,t}\equiv B_{i,t}-g_{i,t}$).[20]

[17] https://www.carbonbrief.org/explainer-the-ratchet-mechanism-within-the-paris-climate-deal

[18] As Golosov et al. (2014:78) write: "Linearity is arguably not too extreme a simplification, since the composition of a concave S-to-temperature mapping with a convex temperature-to-damage function may be close to linear."

[19] To see this, suppose party i emits $g_{i,t}$ and the pollution stock is $G_t=\sigma G_{t-1}+\sum_{i\in N} g_{i,t}$, where $\sigma\in[0,1]$ measures the fraction of the past stock that survives to the next period. If parameter $h>0$ measures each party's per-period marginal environmental harm from the stock G_t, then the present discounted harm of another unit of emission is $h/(1-\sigma\delta)$ for each party. Consequently, $a\equiv h/(1-\sigma\delta)$ measures the present discounted benefit from abating a marginal unit.

[20] For this interpretation of the model, $g_{i,t}=B_{i,t}-q_{i,t}$ might be negative if $q_{i,t}$ is very high. I do not impose any

Of course, also for other public good situations, it will naturally be especially costly for i to contribute a lot relative to i's capacity level, as represented by the stock $R_{i,t}$.

Each party can over time add to the capacity $R_{i,t}$ by investing $r_{i,t}$. The investment cost is assumed to be convex and quadratic and it constitutes the third term in the per-period utility function:

$$\begin{aligned} u_{i,t} &= a\sum_{j\in N} q_{j,t} - \frac{b}{2}\left(q_{i,t} - R_{i,t}\right)^2 - \frac{c}{2}r_{i,t}^2, \text{ where} \\ R_{i,t+1} &= R_{i,t} + r_{i,t}, \end{aligned} \tag{8}$$

and where a, b, and c are positive constants. The parties can have heterogeneous bliss points for consumption and initial technology levels ($R_{i,1}$) but, for simplicity, the parties are assumed to be identical in other respects. At the beginning of each period t, party i intends to maximize i's continuation value, which is:

$$V_{i,t} = u_{i,t} + \mathrm{E}\delta_{i,t}^{\Delta} V_{i,t+1} = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_{i,\tau},$$

Consequently, a party deciding on contributions and investments at the beginning of a period finds it optimal to restrict attention to $\delta \equiv \mathrm{E}\delta_{i,t}^{\Delta}$, even with i.i.d. shocks on the discount factors (as in Section I).

BAU.– As a benchmark, consider the noncooperative Markov-perfect equilibrium (MPE) without any treaty, i.e., the "business as usual" (BAU) equilibrium. At every time t, when i takes as given $R_{i,t}$, the marginal abatement cost equals the marginal benefit for party i:

$$b\left(q_{i,t}^{BAU} - R_{i,t}\right) = a \Leftrightarrow q_{i,t}^{BAU} = R_{i,t} + \frac{a}{b}.$$

Consequently, the investment level does not influence i's future contribution cost, but only i's contribution levels in every future period. Party i's preferred investment level is thus:

$$r_{i,t}^{BAU} = \frac{\delta}{1-\delta}\frac{a}{c}.$$

With this, it is straightforward to derive party i's continuation value in BAU, $V_{i,t}^{BAU}$.[21]

The first-best outcome is given by the exact same equations if just a is replaced by na. In both cases, the second-order conditions trivially hold.

Pledges.– Now, consider the situation that arises after the parties have committed to contribute *more* than the BAU levels. In particular, suppose i has agreed to contribute $x_i \geq 0$ units, beyond i's BAU level, for each of the next T periods. I will continue to restrict attention to Markov-perfect strategies for

constraint $g_{i,t} \geq 0$ because (a) of simplicity, (b) $g_{i,t} < 0$ is in reality feasible with carbon-capture and storage technologies, (c) $g_{i,t} \geq 0$ will not bind if $B_{i,t}$ is growing sufficiently fast over time, and (d) it should be possible to interpret $q_{i,t}$ as (unbounded) contributions to a public good, more generally. See Harstad (2012) for how one may deal with the constraint $g_{i,t} \geq 0$ in a similar (although somewhat different) model without affecting the results qualitatively.

[21] As proven in the Online Appendix:

$$V_{i,t}^{BAU} = \frac{a}{1-\delta}\sum_{j\in N} R_{j,t} + \frac{a^2}{1-\delta}\left(n - \frac{1}{2}\right)\left(\frac{1}{b} + \frac{1}{c}\left[\frac{\delta}{1-\delta}\right]^2\right).$$

the investment levels. Clearly, the commitment x_i is payoff-relevant and it might motivate i to invest $y_{i,t}$ units in addition to the BAU level. Total contributions and investments can then be written as:

$$q_{i,t} \equiv q_{i,t}^{BAU} + x_i \text{ and } r_{i,t} \equiv r_{i,t}^{BAU} + y_{i,t}. \tag{9}$$

The analysis focuses on the choices of x_i and $y_{i,t}$, since these also pin down $q_{i,t}$ and $r_{i,t}$, given BAU.

Timing.– The timing is as illustrated in Fig. 1 except that, at the beginning of each period, the parties simultaneously set the $q_{i,t}$'s and the $r_{i,t}$'s. (The outcome would be the same if these decisions were made sequentially.) If no agreement has been made regarding the contributions, party i is free to set any $q_{i,t}$ and $r_{i,t}$. In each of the T periods after an agreement has been made, the pledge x_i pins down $q_{i,t}$ but party i is free to set $r_{i,t}$ or, equivalently, $y_{i,t}$.

The results below hold whether or not the parties negotiate new pledges after every T-period commitment period. To distinguish the two cases, the index $\iota \in \{0,1\}$ takes the value of 1 if a new commitment period will be negotiated every T period, but $\iota = 0$ if one returns to BAU after the T-period commitment period. (Note that we have $\iota = 1$ for the Paris Agreement since new pledges will be set every five years).

B. *Equilibrium Investments*

Given the above equations, party i's continuation value can be written as a function of the x_i's and the $y_{i,t}$'s. After the pledges have been agreed on, party i's problem is to choose the investment levels over the next T periods. This boils down to a standard optimal control problem which is solved in Online Appendix C. The exact solution for the investment levels is presented here:

LEMMA 1**:** *For each $i \in N$, $t \in \{1,..,T\}$, and $\iota \in \{0,1\}$, equilibrium investments are linear in x_i:*

$$\begin{aligned}
y_{i,t} &= x_i \left(k_1 m_1^{t-1} \left[1 - m_1 \right] - k_2 m_2^{t-1} \left[m_2 - 1 \right] \right) \text{, where} \\
m_1 &\equiv \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c} \right) - \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta}} \in (0,1)\,, \\
m_2 &\equiv \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c} \right) + \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta}} > 1, \\
k_1 &\equiv \frac{m_2^{T-1}\left(m_2 - 1 \right)}{m_1^{T-1}\left(1 - m_1 \right) + m_2^{T-1}\left(m_2 - 1 \right)} \in (0,1)\,, \text{ and} \\
k_2 &\equiv \frac{m_1^{T-1}\left(1 - m_1 \right)}{m_1^{T-1}\left(1 - m_1 \right) + m_2^{T-1}\left(m_2 - 1 \right)} = 1 - k_1 \in (0,1)\,.
\end{aligned}$$

Naturally, if i is committed to contribute a lot, in that x_i is large, then i invests more. It is also easy to check that $y_{i,t}$ increases in T, decreases in t, and reaches zero when $t = T$:

$$y_{i,T} = 0.$$

In the final period, a party invests exactly the same amount as in BAU (i.e., whether one expects to negotiate new pledges in the next period ($\iota = 1$) or not ($\iota = 0$)). The intuition is related to the hold-up

problem: One more technology unit in the next period can —without any other change in investment or contribution cost —raise the total contribution level by one unit then and forever after. The party that invested captures $1/n$ of this benefit, just as in BAU. This intuition also explains why the equilibrium investment at any point in time is the same whether future agreements are expected (i.e., $\iota = 1$) or not ($\iota = 0$). An important implication is that in every period in which the parties have not yet agreed to any pledge, contribution and investment levels are just as in BAU: $x = 0$ and $y_{i,t} = 0$. This outcome is therefore the default outcome when the parties negotiate the pledges.

Lemma 1 states that technology and investment levels will be linear functions of x_i. We can substitute these functions into i's utility function and write party i's continuation value (i.e., the present discounted value of the future utility levels) as a function, $V_{i,1}(\mathbf{x})$, that is quadratic in x_i. Given the benchmark continuation value without an agreement, $V_{i,1}^{BAU}$, we are especially interested in the additional payoff with the pledges: $U_i(\mathbf{x}) \equiv V_{i,1}(\mathbf{x}) - V_{i,1}^{BAU}$. As proven in Online Appendix C, the additional payoff $U_i(\mathbf{x})$ simplifies to Example E, introduced in Section I, with α, β and γ functions of a, b, c, δ, and T.

LEMMA 2**:** *Party i's continuation value, relative to BAU, can be written as in Example E:*

$$\begin{aligned} U_i(\mathbf{x}) &= \alpha \sum_{j \neq i} x_j - \frac{\beta}{2} x_i^2 + \gamma, \quad \textit{where} \\ \alpha &\equiv \frac{a}{1-\delta}\left[1 - \delta^T\left(k_1 m_1^{T-1} + k_2 m_2^{T-1}\right)\right], \\ \beta &\equiv \sum_{t=1}^{T} \delta^{t-1}\left[b\left(k_1 m_1^{t-1} + k_2 m_2^{t-1}\right)^2 + c\left(k_1 m_1^{t-1}\left[1 - m_1\right] - k_2 m_2^{t-1}\left[m_2 - 1\right]\right)^2\right], \textit{ and} \\ \gamma &\equiv \delta^T U_i(\mathbf{x}^*)\,\iota. \end{aligned} \tag{E}$$

C. *Equilibrium Contributions*

We can now combine the models of Section I and Section II. As explained, Lemma 1 implies that the decisions are just as in BAU in every period before the parties have agreed. As soon as the parties have accepted $\mathbf{x}$, i faces the continuation value $U_i(\mathbf{x})$ in addition to the default payoff $V_{i,t}^{BAU}$. The facts that the climate game is dynamic, involves investments, and specifies contributions for T periods are taken into account in the summary statistic $U_i(\mathbf{x})$.

With these payoffs (relative to the default), we can apply Theorem 2 to determine the P&R outcome for x. Since the $U_i(\mathbf{x})$'s are also symmetric, we can also draw on Corollary 1 to obtain:

$$x_i^* = \arg\max_{x_i}\left[U_i(\mathbf{x}) + w\sum_{j \neq i} U_j(\mathbf{x})\right] = w(n-1)\,\alpha/\beta. \tag{10}$$

The smaller w is, the smaller are the x_i^*'s, and the smaller are all investment levels. Both effects make the parties worse off, relative to a situation in which $w = 1$. By combining (E) and (10), we can see that $U_i(\mathbf{x}^*)$ increases in w for every $w < 1$.

PROPOSITION 1**:** *A smaller $w \leq 1$ reduces contributions, investments, and therefore payoffs:*

$$U_i(\mathbf{x}^*) = \frac{\alpha^2 (n-1)^2}{\beta \left(1 - \delta^T \iota\right)} w \left(1 - \frac{w}{2}\right). \tag{11}$$

Fact 1.– As explained in the introduction, the Paris Agreement on climate change calls for P&R, while the top-down negotiations associated with the Kyoto Protocol can be approximated by the NBS. Given this factual difference, Proposition 1 is consistent with the criticism mentioned in the introduction as well as with negative experimental evidence on P&R (Barrett and Dannenberg (2016)). The following sections show that the picture will be more nuanced when we endogenize participation.

III. Participation

This section endogenizes the coalition size and studies how it depends on the bargaining procedure. In line with the literature discussed in the introduction, and according to Nordhaus (2015:1344), "the standard approach in environmental economics" when modelling coalitions begins with a participation stage at which every potential party, $i \in \{1, ..., \overline{n}\}$, decides whether to participate in the coalition. These decisions are made simultaneously and everyone expects that participants will continue by playing the game analyzed in Section II. Given the restriction to MPEs, free riders will simply follow their dominant BAU strategy and set $x_i = 0$.

It is most natural (and common) to focus on pure-strategy equilibria at the participation stage, and doing so pins down the equilibrium coalition size, n. I start by ignoring the constraint $n \leq \overline{n}$ as well as a possible minimum participation threshold, $\underline{n}$, but these constraints are extensively discussed in Section IV. I also begin assuming that the participation decision is made once and for all, but Section VII explains why the results continue to hold when this assumption is relaxed.

A. *Equilibrium Participation*

Since coalition members ends up contributing more than the level that would maximize their own utility, there is a cost of participating in the coalition. For a member to be willing to participate, the benefit of participating must outweigh this cost. The benefit of participating is that other participants will internalize (a fraction w of) the utility of one additional coalition member.

Given Lemma 2, which states that the parties' payoffs can be summarized as (E), I henceforth restrict attention to these payoff functions. The equilibrium payoff for each of the n participants is given by (11). If one of these parties instead free rides, the free rider's payoff will be $\alpha (n-1) w (n-2) \alpha / \beta \left(1 - \delta^T \iota\right)$,

since each of the other $n-1$ parties will now contribute $w(n-2)\alpha/\beta$ every T period (if $\iota = 1$). By comparison, participation is beneficial if:

$$U_i(\mathbf{x}^*) = \frac{\alpha^2(n-1)^2}{\beta\left(1-\delta^T\iota\right)}w\left(1-\frac{w}{2}\right) \geq \frac{\alpha^2(n-1)(n-2)}{\beta\left(1-\delta^T\iota\right)}w \Rightarrow n \leq 1+\frac{2}{w}. \tag{12}$$

The size n cannot be too great since then individual contributions would be so large and so costly that free riding would be preferable. For a coalition to be stable, (12) must hold for the equilibrium n, but it must fail for any larger n (since, otherwise, nonmembers would also like to participate). Thus, it is useful to employ the function $\lfloor\cdot\rfloor$, mapping its argument to the largest weakly smaller integer.

PROPOSITION 2**:** *The equilibrium coalition size is decreasing in* w:

$$n = \lfloor 1+2/w \rfloor.$$

Note that $n = 3$ if $w = 1$, as when applying the NBS. This "small-coalition paradox" is well known in the literature, which also discusses the trade-off between "narrow and deep" vs. "broad and shallow" coalitions (see the literature review). With P&R bargaining, w is small and a coalition member is not expected to contribute a lot. The cost of participation is then small, and participation is attractive for a larger set of n's.

Since the number of participants must be an integer, n is a step function that decreases in w. When comparing bargaining procedures, we are interested in large rather than small differences in w. Thus, it is not unreasonable to abstract from the fact that n must be an integer and to use the approximation

$$n \approx n(w) \equiv 1+2/w. \tag{13}$$

With this approximation, the product $(n-1)w$ is a constant that is pinned down when a (marginal) member must be indifferent between free-riding and participating.

B. *Equilibrium Contributions - Revisited*

When $(n-1)w$ stays constant as w is reduced, $x_i^* = (n-1)w\alpha/\beta$ also remains constant, and so does every investment level $y_{i,t}$. Since the individual contributions are invariant in w, while n is decreasing in w, the sum of payoffs will be larger when w is small. A participant's payoff is also larger when w is small: this is evident when the endogenous n, as described by (13), is combined with the utility (11). This gives:

$$U_i^* = \frac{4\alpha^2}{\beta\left(1-\delta^T\iota\right)}\left(\frac{1}{w}-\frac{1}{2}\right). \tag{14}$$

COROLLARY 3**:** *With endogenous participation, approximated by n(w), Proposition 1 is reversed: A smaller* w *increases aggregate contributions, investments, and welfare.*

Fact 2.– While only 37 countries promised emission cuts for the Kyoto Protocol's first commitment period, 195 countries have pledged to contribute to the Paris Agreement. This fact is consistent with Proposition 2 since the Paris Agreement is associated with P&R and thus a smaller w. The result that x_i continues to be large even if w is small (because n increases) might shed some light on why the pledges in Table 1 are substantial, despite the P&R bargaining game.

Remark on Robustness.– The result that x_i stays unchanged when n varies endogenously with w hinges on the functional forms. If the continuation value had ended up being:

$$\alpha \sum_{j \neq i} x_j - \frac{\beta}{2} x_i^{\varphi}, \text{ with } \varphi > 1,$$

then one can show that: n always decreases in w; x_i decreases in w if $\varphi < 2$ but increases in w if $\varphi > 2$; U_i^* always decreases in w, when $w \in [0, 1]$. Thus, the rationalization of Fact 2 survives this generalization.

IV. When to Choose Pledge-and-Review

With reasonable modifications, the preference concerning bargaining procedure involves a trade-off. This section discusses realistic constraints on n in order to compare participants' payoffs under a low $w = \underline{w}$ (say, P&R) and a large $w = \overline{w} > \underline{w}$ (with the NBS, $w = 1$).

A. *Maximum Participation*

The world consists of a finite number ($\overline{n}$) of countries that can participate in a coalition. If $\overline{n} < n(\overline{w}) < n(\underline{w})$, where $n(\cdot)$ is defined by (13), then both bargaining games (characterized by $\underline{w}$ or $\overline{w}$) induce full participation. In this case, $\overline{w}$ is preferable, according to Proposition 1. If, instead, $n(\overline{w}) < n(\underline{w}) < \overline{n}$, the upper boundary on n is nonbinding and $\underline{w}$ is preferable, according to Corollary 3. A trade-off arises when $n(\overline{w}) < \overline{n} < n(\underline{w})$, since then participation is larger, but individual contributions smaller, when w is small. In this case, a sufficiently large $\overline{n}$ is necessary to ensure that a participant's payoff is larger under $\underline{w}$.

The exact condition follows when comparing a participant's utility, as given by equation (11), for the two cases. The payoff is larger when $w = \underline{w}$ than when $w = \overline{w}$ if:

$$\frac{\alpha^2 (\overline{n}-1)^2}{\beta\left(1-\delta^T \iota\right)} \underline{w}^2 \left(\frac{1}{\underline{w}} - \frac{1}{2}\right) > \frac{\alpha^2 (n(\overline{w})-1)^2}{\beta\left(1-\delta^T \iota\right)} \overline{w}^2 \left(\frac{1}{\overline{w}} - \frac{1}{2}\right) \Rightarrow \frac{\overline{n}-1}{n(\overline{w})-1} > \Omega, \text{ where}$$

$$\Omega \equiv \sqrt{\frac{\overline{w}(1-\overline{w}/2)}{\underline{w}(1-\underline{w}/2)}} \in \left(1, \frac{\overline{w}}{\underline{w}}\right).$$

B. *Minimum Participation*

Most international treaties specify minimum participation thresholds that must be met for the treaty to enter into force. This threshold was the same for the Kyoto Protocol and the Paris Agreement. In isolation, the effect of such a threshold, $\underline{n}$, is that $n = \max\{\underline{n}, n(w)\}$ since none of $\underline{n}$ participants prefer to free-ride when the consequence will be the BAU outcome.[22]

There are also other forces that can lead to a minimum participation level. After all, countries are more heterogeneous in reality than permitted in the model above. If the willingness to participate varied across countries, n would not decline as fast as predicted by Proposition 2. A simple way of capturing this heterogeneity is to assume that a number $\underline{n}$ is committed to participate regardless of w. The reason these parties are committed can be outside the model, but one may think of existing international treaties on non-climate issues such as international trade or regulatory politics. To be specific, European Union member countries cannot easily opt out of an environmental agreement unilaterally.

The minimum participation threshold $\underline{n}$ is relevant only if $\underline{n} > n(\overline{w})$. If also $\underline{n} > n(\underline{w})$, the number of participants is always $\underline{n}$ and then the larger w is optimal, according to Proposition 1. To isolate the trade-off associated with $\underline{n}$, suppose $n(\overline{w}) < \underline{n} < n(\underline{w}) < \overline{n}$. In this case, only $\underline{n}$ parties participate under $\overline{w}$, while participation under $\underline{w}$ is given by $n(\underline{w})$. By comparison, a participant's payoff can be larger under $\underline{w}$ if and only if $\underline{n}$ is sufficiently small.

The exact condition follows when we use the utility function (11) to compare the two cases. The payoff is larger when $w = \underline{w}$ than when $w = \overline{w}$ if:

$$\frac{\alpha^2 \left(n(\underline{w}) - 1\right)^2}{\beta\left(1 - \delta^T \iota\right)} \underline{w}^2 \left(\frac{1}{\underline{w}} - \frac{1}{2}\right) > \frac{\alpha^2 \left(\underline{n} - 1\right)^2}{\beta\left(1 - \delta^T \iota\right)} \overline{w}^2 \left(\frac{1}{\overline{w}} - \frac{1}{2}\right) \Rightarrow \frac{n(\underline{w}) - 1}{\underline{n} - 1} > \Omega.$$

C. *The Preferred Bargaining Game*

Figure 3 illustrates that payoffs are non-monotonic in w. Clearly, it is possible that both the minimum and the maximum participation levels bind at the same time. This happens if $n(\overline{w}) < \underline{n} < \overline{n} < n(\underline{w})$. In

[22] In a pure-strategy equilibrium, one can correctly anticipate which countries will participate. When $n \geq \underline{n}$ binds, the $\underline{n}$ participants must clearly agree unanimously, as assumed in Section I. Also if the equilibrium coalition size satisfies $n > \underline{n}$, the agreement must be acceptable by all n contributors, if they represent sovereign countries. In this case, a subgroup of the n has the option of concluding the agreement even if party j, say, declined $\mathbf{x}$: But sovereign j would thereafter set $x_j = 0$, the subgroup would be worse off (the equilibrium $U_i(\mathbf{x}^*)$ increases in n), so this threat would not be credible. The alternative strategy to first negotiate and sign, then not ratify, is limited by the Vienna Convention on the Law of Treaties (Article 18), which states: "A State is obliged to refrain from acts which would defeat the object and purpose of a treaty when...it has signed the treaty..." Consequently, almost every country that signed the Kyoto Protocol (or the Paris Agreement) has also ratified; the US being the famous exception.

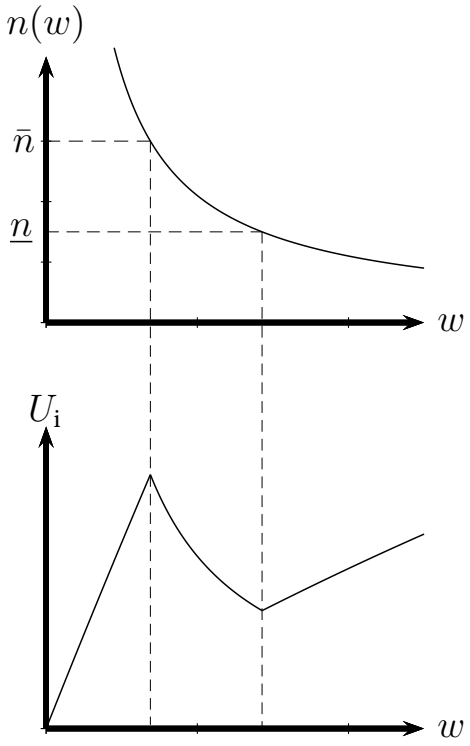

Figure 3: *Participation and participants' payoffs are strictly decreasing in w only when $n(w) \in (\underline{n}, \overline{n})$.*

this case, there is full participation under $\underline{w}$, but only $\underline{n}$ parties participate under $\overline{w}$. In this situation, $\underline{w}$ is preferred when $\overline{n}$ is large and $\underline{n}$ is small.

The utility function (11) shows that $\underline{w}$ gives a higher payoff than $\overline{w}$ if:

$$\frac{\alpha^2 (\overline{n}-1)^2}{\beta\left(1-\delta^T \iota\right)} \underline{w}^2 \left(\frac{1}{\underline{w}} - \frac{1}{2}\right) > \frac{\alpha^2 (\underline{n}-1)^2}{\beta\left(1-\delta^T \iota\right)} \overline{w}^2 \left(\frac{1}{\overline{w}} - \frac{1}{2}\right) \Rightarrow \frac{\overline{n}-1}{\underline{n}-1} > \Omega.$$

The three conditions can be combined in the following way.

PROPOSITION 3**:** *Participants prefer pledge-and-review bargaining (i.e., to switch from $w = \overline{w}$ to $w = \underline{w} < \overline{w}$) if $\overline{n}$ is large while $\underline{n}$ is small. The exact condition is:*

$$\frac{\min\{\overline{n}-1, 2/\underline{w}\}}{\max\{\underline{n}-1, 2/\overline{w}\}} > \Omega. \tag{15}$$

This condition is drawn as the solid line in Figure 4. If there is a larger number of potential parties, or if fewer of them are committed to participate, we move in the direction of the arrow in the figure. Then, the "shallow" agreement ($\underline{w}$) becomes preferred even though the "deep" agreement was preferred given a smaller number of potential parties and a larger number of committed parties.

Fact 3.– From Kyoto to Paris: One may argue that both these developments (i.e., a larger $\overline{n}$ and a smaller $\underline{n}$) are in line with changes in world politics over the last couple of decades. Today we have a large number of emerging economies which in the 1990s were developing countries that could not be

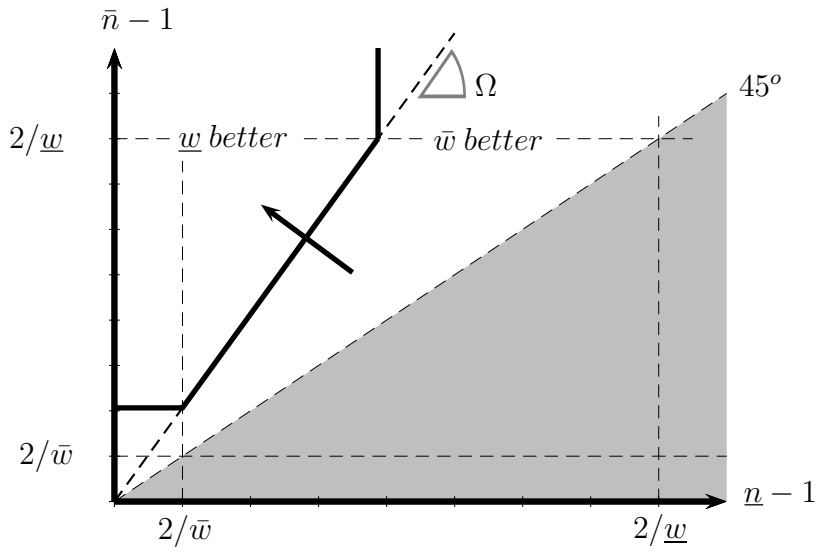

Figure 4: *Participants prefer to switch to pledge-and-review ($\underline{w}$) above the solid line.*

expected to contribute much to an international climate change treaty. For the model, this development implies that the number of relevant parties, n, has increased.

During the same period, seven of the original Annex I countries, who initially signed the Kyoto Protocol, announced that they would not contribute to the Kyoto Protocol's second commitment period.[23] These withdrawals may be interpreted as a reduction in the number of committed countries, $\underline{n}$. For either reason (or both), the switch to P&R is consistent with the theory of this paper.

It is straightforward to show that the *uncommitted* countries prefer the broad agreement only when:

$$\min\left\{\overline{n}-1, 2/\underline{w}\right\} > \sqrt{\frac{\max\left\{\overline{w}\left(\underline{n}-1\right)\underline{n}, 4/\overline{w}+2\right\}}{\underline{w}\left(1-\underline{w}/2\right)}},$$

which is stronger than (15). The theory can thus rationalize why developing countries preferred to continue with the Kyoto Protocol.[24] With this disagreement in mind, we can conclude that the original set of participants prefer to switch to P&R too soon, that is, for a larger set of parameters than the set under which such a switch increases global welfare. Analogously, if the new potential members were pivotal in the decision on treaty design, they would accept P&R too late or too seldom, relative to the decision that is optimal if the original members' payoffs are taken into account.

[23] According to the IPCC (2014:1025), "a number of Annex I countries (Belarus, Canada, Japan, New Zealand, Russia, the United States, and Ukraine) decided not to participate in the second commitment period."

[24] Bodansky et al. (2017:202) write: "Developing countries, for which the Kyoto model has obvious attractions because they are exempt from emissions targets, were keen to extend the protocol for a second and future commitment periods. Kyoto Annex B parties, in contrast, were reluctant to do so, for some countries because of Kyoto's prescriptive architecture, and for others because they did not want to be subject to emissions targets if the US, China, and other large emitters were not."

V. Enforcement and Compliance

Is is elsewhere in this paper assumed that the parties are able to commit to the pledges for T periods. However, given the incentive to free ride, discussed in Sections III and IV, it is reasonable to also be concerned with the temptation to contribute less at the time when other participants are expected to deliver on their promises. This section contributes to the literature on self-enforcing agreements (see, for example, Barrett (1994; 2002); Dutta and Radner (2004; 2006); Harstad et al. (2019), and the references therein) by showing when and why certain procedures, such as pledge-and-review, are more likely than others to be self-enforcing. To study self-enforcement, this section deviates from the rest of the paper by considering non-Markovian history-dependent strategies such as trigger strategies.[25]

Since decisions are made simultaneously, a party that "defects" by not contributing will be able to enjoy the benefit from other participants' contributions in that period. When n is exogenously given, individual contributions are relatively small when w is small. The temptation to defect may thus also be small when w is small. When n is endogenous, a smaller w motivates a larger n and that, in turn, implies that it is more important for a party that cooperation continues. In both cases, it is intuitive that the incentive constraint is more likely to hold when w is small, as under P&R bargaining.

To illustrate this intuition, suppose the parties revert to BAU (i.e., the noncooperative MPE) forever as soon as one party has defected by contributing less than pledged.[26] Section VII and the Online Appendix permit the punishment to last for any length $l \leq \infty$ of periods and to be triggered by any probability $\phi \in (0, 1]$. The result below holds qualitatively for every $l > 0$ and $\phi > 0$.

Note that a finitely long agreement cannot be self-enforcing if $\iota = 0$, i.e., if one returns to the BAU outcome after period T. In such a situation there would be no incentive to comply in period T, and thus not in period $T - 1$, etc. This observation can rationalize why the Paris Agreement specifies that new pledges must be set every five year (i.e., $\iota = 1$) and why I henceforth restrict attention to $\iota = 1$.

The dynamic game in Section II is different from a repeated game because past investments influence BAU and thus all future contributions. When party i invests $y_{i,t}$, then i's contribution will increase by $y_{i,t}$ in every future period if the parties revert to BAU. Since every $y_{i,t}$ is largest at the beginning of the commitment period, the temptation to defect is also largest in the beginning. Then, the payoff if i defects (by not contributing) is as expressed on the right-hand side in the following inequality. This payoff must

[25] Although space constraints prevent a full characterization of SPEs, note that if $\delta \to 1$ then "folk theorems" imply that a large set of outcomes, including the first best (characterized in Section II(A)), can be supported as SPEs. For a smaller δ, the best SPE is likely to be distorted in ways already investigated by Harstad et al. (2019).

[26] On the one hand, it is possible to sustain as SPEs harsher punishments than the reversion to BAU. With harsher punishments, a treaty would be self-enforcing under a larger set of circumstances than those derived below. On the other hand, if parties could renegotiate punishments, then a treaty would be self-enforcing for a smaller set of parameters.

be smaller than i's equilibrium payoff, on the left-hand side:

$$\begin{aligned} U_i(\mathbf{x}^*) &= \frac{\alpha^2 (n-1)^2}{\beta\left(1-\delta^T\right)} w\left(1-\frac{w}{2}\right) \geq a\left(\sum_{j\neq i} x_j + \frac{\delta}{1-\delta}\sum_{j\neq i} y_{j,1}\right) \Leftrightarrow \\ w &\leq \widehat{w} \equiv 2 - 2\left[1-\delta\left(k_1 m_1 + k_2 m_2\right)\right] \frac{a\left(1-\delta^T\right)}{\alpha(1-\delta)}. \end{aligned} \tag{16}$$

The implication follows when we substitute for the equilibrium $y_{i,1}$, x_i, and α and rewrite. The condition is easier to satisfy when w is small, i.e., if the bargaining procedure is characterized by P&R rather than by the NBS, for example.

Interestingly, n drops out from the inequality, and thus n does not influence whether the bargaining outcome will be self-enforcing. It follows that condition (16) is robust to whether n is exogenous (as in Section II) or endogenous (as in Section III). Technically, the invariance follows because both the cost of the individual contribution and the benefit from the others' contributions are proportional to $(n-1)^2$.

PROPOSITION 4**:** *Regardless of whether participation is exogenous or endogenous, the bargaining outcome is self-enforcing if and only if $w \leq \widehat{w}$, defined by (16).*

If w is so large that the incentive constraint is violated, then the parties must find additional ways of raising the cost of noncompliance. In reality, there are several ways of increasing these costs, since the exact wording in an international treaty influences the political and reputational costs if one later defects. Although there exists no world government ready to enforce contracts, it is not irrelevant whether a treaty is called "legally binding." IPCC (2014:1020) explains that "a more legally binding commitment ... signals a greater seriousness by states ... These factors increase the costs of violation (through enforcement and sanctions at international and domestic scales, the loss of mutual cooperation by others, and the loss of reputation and credibility in future negotiations)."

Fact 4.– The pledges are not legally binding under the Paris Agreement, but: "the Kyoto Protocol represents a much harder, more prescriptive approach, including legally binding, quantified emissions limitation targets" (Bodansky and Rajamani (2017:22)). This difference between the two agreements is consistent with Proposition 4. Since the Paris Agreement applies P&R bargaining, where w is smaller, it is possible that the incentive constraint holds for this agreement without making it legally binding. In this case, the parties would strictly prefer non-binding commitments if there were tiny costs associated with legally bindingness (e.g., the observed emission level might be only partly under the government's control, etc.).

Of course, when one can raise the cost of noncompliance by modifying the legal status of the agreement, then countries will comply on the equilibrium path regardless of the bargaining procedure. In line with this prediction, the remaining "thirty-six Kyoto parties [after Canada pulled out] were in full compliance with their first commitment period targets" (Bodansky and Rajamani (2017:31)).

VI. The Commitment Period Length

The results above hold for any commitment period length. The optimal T, from the participants' point of view, trades off the effect on investments with the benefit that newly developed technology can strengthen the commitments sooner when T is small. On the one hand, the larger the length of the commitment period is, the larger the equilibrium investments are at every point in time (this can be seen from Lemma 1). This comparative static was explained by the classic hold-up problem. On the other hand, after investments have been made, it is ex post optimal for all parties to start bargaining soon again to take advantage of the newly developed capacity.[27]

The trade-off when it comes to deciding on T is independent of w and n in the model above. When n is exogenous, then a party's payoff is given by equation (11). When n is instead endogenous, a party's payoff is given by (14). Every participant's preferred T is the same in either case:

$$T^* = \arg\max_T U_i(\mathbf{x}^*) = \arg\max_T U_i^* = \arg\max_T \frac{\alpha^2}{\beta\left(1-\delta^T\iota\right)},$$

where α and β are functions of T, as described by Lemma 2.[28]

PROPOSITION 5**:** *The optimal commitment period length, T^*, is independent of n and w, and of whether n and w are endogenous or exogenous.*

Fact 5.– Given the many differences between the Kyoto Protocol and the Paris Agreement, the two are surprisingly similar regarding how frequently commitments must be updated. Pledges under the Paris Agreement must be updated every five years, and the Kyoto Protocol's first commitment period was also five years (2007-2012). This similarity is consistent with Proposition 5, stating that the optimal commitment period length is the same, despite the many other differences between the two treaties.

Not only this result, but also the mechanism driving it seems to match well with reality. OECD's (2018:5) first argument for a five- rather than a ten-year commitment period is: "More regular opportunities to make technical and fundamental adjustments to NDCs as well as to incorporate effects of technology..."[29]

[27] The combined trade-off is new to the literature, but the hold-up problem associated with a small T is already recognized: see, e.g., Beccherle and Tirole (2011) or Harstad (2016). Harris and Holmstrom (1987) observed that a small T is beneficial since it permits a rigid contract to be updated when the external environment changes.

[28] If $c \to \infty$, or $b/c \to 0$, investments are irrelevant. With $\iota = 1$, $\alpha^2/\beta\left(1-\delta^T\iota\right) \to a^2/b\left(1-\delta\right)$, independent of T. In this limit, every T is equally good. If instead $\iota = 0$, i.e., the agreement is followed by the BAU outcome, then $T = \infty$ is optimal.

[29] It is reasonable that Kyoto's second commitment period would also have been five years, if the parties had not anticipated that a new global treaty would be effective from 2020. According to Bodansky et al. (2017:203), in 2011, "Parties disagreed on several issues including: the length of the commitment period– whether it should be five years (like the first commitment period) or eight years (to coincide with the scheduled launch of the 2015 agreement)." In 2012, "the eight-year duration of the second commitment period was chosen so as to end when the Paris Agreement's NDCs were expected to take effect, and thus to avoid a commitment gap" (p. 205).

VII. Robustness

The model above is simple and stylized yet able to rationalize Facts 1-5, discussed above. This rationalization is quite robust in that it continues to hold for a number of model modifications. This section explains (and Online Appendix C proves) that Propositions 1-5 hold even if the parties negotiate investment levels or emission taxes (or both) instead of (or in addition to) the x_i's. The x_i's can also be time dependent, and the investment levels might be decided by profit-maximizing firms, without changing the propositions. The results are also quite robust to changes in timing.

(i) Pledging to Invest.– Some of the INDCs in the Paris Agreement specify national targets for renewable energy.[30] This possibility can be captured by letting parties decide on the $y_{i,t}$'s instead of on the x_i's. As discussed in the Online Appendix, it is straightforward to analyze this scenario: when the $y_{i,t}$'s, but not the $x_{i,t}$'s, are pinned down, then i's choice of $x_{i,t}$ will satisfy $b\left(x_{i,t} - Y_{i,t}\right) = 0$, just as in BAU. If the investment pledge must be time independent (y_i) throughout a commitment period, then i's continuation value can be written as in Example (E), where x_i is replaced by y_i, although the definitions of α and β will be different. The proofs of Propositions 1-5 are thus similar to earlier proofs.

In fact, i's continuation value will be separable in the $\mathbf{y}_t$'s, where $\mathbf{y}_t = (y_{1,t}, ..., y_{n,t})$. Consequently, we can apply Corollary 1 when parties negotiate $\mathbf{y}_t$, while keeping fixed the investment levels for other periods. Corollary 1 will imply that the P&R outcome for $y_{i,t}$ is:

$$y_{i,t}^* = (n-1)\, w \frac{\delta a/c}{1-\delta}. \tag{17}$$

Since this $y_{i,t}$ is time independent, there is no loss for the parties if they restrict attention to time-independent investment levels. For these reasons, the length of the commitment period (T) will not influence payoffs, and any T is here equally good, regardless of the levels of n and w.

(ii) Pledging on Emission Taxes.– It is also straightforward to allow parties to pledge on domestic emission taxes, instead of on emission cuts. With an emission tax $z_{i,t}$, it is natural that consumption of fossil fuel be given by the condition in which the marginal benefit of consuming (or the marginal cost of abating) equals the tax: $b\left(x_{i,t} - Y_{i,t}\right) = z_{i,t}$. When parties are free to decide their investment levels, they will invest just as in BAU, so $y_{i,t} = 0$. If the emission tax level must be time independent (z_i) throughout the commitment period, then i's continuation value can be written as in Example (E), where x_i is replaced by z_i, although the definitions of α and β differ. Again, the proofs of Propositions 1-5 are similar to earlier proofs.

In fact, i's continuation value will be separable in the z_t's, where $z_t = (z_{1,t}, ..., z_{n,t})$. Consequently, we can apply Corollary 1 when parties negotiate z_t, while keeping fixed the emission taxes for other periods.

[30] For example, China pledges to increase the share of non-fossil fuels in its primary energy consumption to around 20 percent, while India pledges to produce about 40 percent of its electric power from non-fossil-fuel-based energy resources by 2030. For a recent overview, see http://cait.wri.org/indc/#/.

Corollary 1 will imply that the P&R outcome for $z_{i,t}$ is:

$$z_{i,t}^{*} = (n-1)\, wa. \tag{18}$$

Since this $z_{i,t}$ is time independent, there is no loss for parties if they restrict attention to time-independent emission taxes. For these reasons, the length of the commitment period (T) will not influence payoffs, and any T is equally good, regardless of the levels of n and w.

As a side remark, it is worth noting that the choice of instrument (i.e., whether parties should negotiate x_i's, y_i's, or z_i's) is also independent of n and w. As proven in the Online Appendix, negotiating investment levels is better for all parties than negotiating emission taxes if and only if investments are inexpensive and the future is important:[31]

$$\frac{1}{\delta} < 1 + \sqrt{\frac{b}{c}}.$$

(iii) Pledging on Investment Levels and Emission Taxes.– Party i's continuation value is separable in the $\mathbf{y}_t$'s and the z_t's, it can be shown. Thus, Corollary 1 can be applied for each instrument separately, while keeping the other fixed. With this procedure, the outcome is given by the combination of (17) and (18). In this case, we have a "complete contract" since, given the negotiated investment levels (and thus the $Y_{i,t}$'s), the emission taxes pin down the contribution levels.

(iv) Pledging on Investment Levels and Contribution Levels.– Once the investment levels (and thus the $Y_{i,t}$'s) are pinned down, negotiating $z_{i,t} = b\,(x_{i,t} - Y_{i,t})$ is equivalent to negotiating $x_{i,t}$. Thus, Scenario (iii) leads to the same outcome as that which occurs when parties can negotiate every investment level and every contribution level. As before, the choice of T is irrelevant, regardless of the n and w levels.[32]

(v) Time-dependent Contribution Levels.– In Scenario (iv), one best choice of T is $T = \infty$. With $T = \infty$, it is actually irrelevant that parties have negotiated investment levels in addition to contribution levels. The irrelevance follows because, once the $x_{i,t}$'s are given for every time, there is no externality associated with the $y_{i,t}$'s and, hence, every party will have incentives to invest optimally, without any need to negotiate $y_{i,t}$. As is shown in the Online Appendix, the equilibrium time-dependent contribution level is:

$$x_{i,t}^{*} = (n-1)\, w\frac{a}{b} + (n-1)\, w\frac{a}{c}\frac{\delta}{1-\delta}t.$$

Given this pledge, party i prefers to invest as in (17), ensuring that the marginal benefit from consuming (and from cutting emissions) is $b\left(x_{i,t}^{*} - ty_{i,t}^{*}\right) = (n-1)\, wa$, which coincides with $z_{i,t}^{*}$ in (18).

In this situation, it is clear that parties are strictly better off with $T = \infty$ than with $T < \infty$, since, with any finite T, equilibrium $y_{i,t}$'s are lower (and less efficient) than the $y_{i,t}$ that would follow in Scenario

[31] The comparison to the situation in which the x_i's are negotiated is more complex, however.

[32] If parties can negotiate time-independent x_j's and y_j's, which must stay constant throughout the commitment period, then the parties would strictly prefer $T = 1$. With $T = 1$, the outcome will be the same as with time-dependent policies (Scenario (iv) and Scenario (iii)), while $T > 1$ would be less efficient. In contrast to the discussion on the optimal T, in Section V, there is no need to have a large T when the first-period investment level can be negotiated, since agreeing on $y_{i,1}$ circumvents the hold-up problem.

(iv), which coincides with the equilibrium $y_{i,t}$'s when $T = \infty$. When referring to the trade-off discussed in the previous section, there is here no reason to reduce T in order to update the pledges when the pledges can be time dependent. It is thus optimal with $T = \infty$ to mitigate the hold-up problem.

(vi) Firms Invest.– All three Scenarios (iii)-(v) implement the complete contract outcome, i.e., as when all $y_{i,t}$ and $x_{i,t}$ are negotiated according to P&R. The same outcome can be achieved if parties negotiate $x_{i,t}$ at time t, for $T = 1$, while letting firms invest. If so, the pledged $x_{i,t}$ will satisfy $b(x_{i,t} - Y_{i,t}) = (n-1)wa$, which thus also characterizes the marginal willingness to pay for another unit of $Y_{i,t}$ at time t. Thus, the present discounted value of a unit invested today is $\delta(n-1)wa/(1-\delta)$, while the marginal investment cost is $cy_{i,t}$. The two are equalized when profit-maximizing price-taking firms decide on $y_{i,t}$ and, then, the result is (17), just as when the parties negotiate the investment levels directly. In this situation, it is clear that parties are strictly better off with $T = 1$ than with $T > 1$ (unless the contribution levels are time dependent). Firms, unlike governments, are not discouraged by the nations' hold-up problem when new pledges are negotiated.[33]

(vii) The Timing of T.– Proposition 4 showed that every participant agreed on the choice of T and that this choice was independent of n and w. Thus, the choice of T remains the same whether participants decide on T after the participation stage, before the bargaining-choice stage, or in between the two. The timing of T influences neither the equilibrium level of n nor the preference regarding w.

(viii) Multiple Participation Stages.– Propositions 2-5 continue to hold if there is a participation stage before pledges are negotiated at the beginning of every commitment period (i.e., every T period). Participating is then attractive if and only if (12) holds, just as before. The identity of the n participants is also the same in every commitment period in an MPE, implying that every participant's continuation value is given by (14). Thus, the proofs of Propositions 2-5 continue to hold.

(ix) Multiple Bargaining-choice Stages.– Propositions 2-5 also hold if w, as well as n, are endogenously chosen at the beginning of every commitment period, for the same reasons as in Scenario (viii). In fact, if some parameters (such as $\underline{n}$ and/or $\overline{n}$) change every T period, then Propositions 2-5 characterize the outcome, and Proposition 3 characterizes the best bargaining procedure, for every commitment period, regardless of the parameter values after period T. This generalization implies that Proposition 3 can indeed rationalize a change from one procedure to another, if $\underline{n}$ and/or $\overline{n}$ has changed.[34]

(x) Limited Punishments.– When the self-enforcement constraint was discussed, Proposition 4 relied on the assumption that if one party defected, then all parties would play BAU forever after. On the one hand, one may argue that it is optimistic to assume that a defection can be observed with probability $\phi = 1$. On the other hand, one may also argue that, if cooperation has broken down, then parties might renegotiate to start cooperating again. To capture these concerns to some extent, the proof of Proposition 4 permits defection to be punished with a reversion to BAU for $l \leq \infty$ periods with probability $\phi \leq 1$

[33] If each government can subsidize/tax the firms' investments, it can implement its preferred choice of investment, as described in the previous sections. Then, even the exact equations in Sections II-VI stay unchanged, one can argue.

[34] The analysis would have been more complicated, however, if parameters changed also within commitment periods.

(while, with probability $1-\phi$, there is no punishment). The incentive constraint is then:

$$w \leq 2 - 2\left[\frac{1-\delta\left(k_1 m_1 + k_2 m_2\right)}{(1-\delta)\left(1-\delta\left(\left[1-\phi+\phi\delta^l\right]\right)\right)}\right]\frac{a\left(1-\delta^T\right)}{\alpha}.$$

A smaller ϕ or l strengthens the incentive constraint, but Proposition 4 holds for all $\phi \in (0,1]$ and $l > 0$.

These generalizations can be summarized in the following proposition (proven in the Online Appendix).

PROPOSITION 6**:** *Propositions 1-5 continue to hold if parties pledge-and-review bargain:*
(i) investment levels instead of $\mathbf{x}$;
(ii) emission taxes instead of $\mathbf{x}$*;*
(iii) investment levels and emission taxes instead of $\mathbf{x}$*;*
(iv) investment levels and $\mathbf{x}$ *instead of only* $\mathbf{x}$;
(v) a time profile $\{\mathbf{x}_t\}_{t=1}^{\infty}$ *instead of a time-independent* $\mathbf{x}$;
(vi) $\mathbf{x}$*, while profit-maximizing price-taking firms invest;*
(vii) T *after the* n *stage, or before* n *but after the* w *stage.*
(viii) Propositions 2-5 continue to hold if there is a participation stage before every commitment period.
(ix) Propositions 3-5 continue to hold if both w *and* n *are decided on every commitment period.*
(x) Proposition 4 holds if defection leads to BAU for $l \in (0,\infty]$ *periods with probability* $\phi \in (0,1]$.

As discussed, the optimal level of T varies across the scenarios, but for every scenario the optimal T is independent of the bargaining procedure. Obviously, the optimal T, as well as the other results, may depend on many things that are outside of this model, such as policy makers' ability to commit to the distant future, or the ability to predict the optimal level of contributions many years in advance. Propositions 1-4 have thus been derived for any fixed T, and they hold for every T.

It is possible to extend the model in many other directions as well.[35] This paper has simplified tremendously partly because the additional insight of some of the generalizations would overlap with results in earlier papers and partly because these extensions are evidently not necessary to rationalize the five facts on how the Paris Agreement compares with the Kyoto Protocol. On the contrary, if the modified model did predict that T should be a function of w or n, or that any of the other propositions would change, then it would not be supported by Facts 1-5.

[35] In Harstad (2016), relying on the NBS, both pollution and shocks on the marginal environmental harm accumulate over time. The shocks make it hard to predict optimal pledges and they motivate a small T, while the hold-up problem motivates a large T, especially when there are large technological spillovers. In Battaglini and Harstad (2016), n and subsequently T are set endogenously before every commitment period. Then, participants may prefer a small T if n is small, since a small T facilitates the admission of new members sooner. Since a small T also leads to hold-up problems, countries are motivated to participate to ensure a large T. Acemoglu et al. (2012) permit investments in dirty as well as in green technology, Dutta and Radner (2018) transfers from the North to the South, and Martimort and Sand-Zantman (2016) a mechanism-design approach.

IIX. Conclusions

This paper presents a model and an analysis of pledge-and-review bargaining. The novelty of this bargaining game is that each party proposes how much to contribute independently —not conditional on what other parties pledge —before the parties agree to the vector of pledges. If there is some uncertainty regarding what other parties are willing to accept, for example due to shocks on the short-term discount rate, then contributions will be larger if there is a substantial variance in these shocks. In equilibrium, each party's contribution level is as described by an asymmetric Nash bargaining solution, where the weights on others' payoffs reflect the distribution and correlation of shocks. Since the weights vary from pledge to pledge, the collection of pledges is not Pareto optimal. The inefficiency may arise in business as well as political negotiations.

The P&R bargaining game has been associated with the 2015 Paris Agreement on climate change, and it makes the agreement's approach rather different from the top-down approach that characterized the 1997 Kyoto Protocol (often approximated by the NBS). The analysis uncovered that (1) the factual difference in bargaining procedure can rationalize four other facts on how the Paris Agreement differs from the Kyoto Protocol: (2) Since P&R bargaining is not very demanding, it attracts a larger number of participants. This result can explain why many more countries took on emission cuts in Paris than in Kyoto. (3) Since raising participation is the main benefit of P&R, it is preferable if and only if there is a large number of relevant players. This logic can explain why P&R was preferred in the 2010s, after several developing economies had become emerging economies, whereas the Kyoto Protocol's top-down procedure was chosen in the 1990s. (4) Since P&R is not very demanding, the equilibrium pledges are more likely to be self-enforcing than is the NBS. This result is consistent with the fact that the Kyoto Protocol's emission cuts were legally binding, whereas they are not for the Paris Agreement. (5) Despite all these differences, the theory is also consistent with the fact that the commitment period's length has been the same for the two agreements.

Although the paper focuses on a positive analysis, the reader may instinctively search for normative lessons. Pledge-and-review bargaining might not be as inadequate as it at first appears to be; it can actually be preferable to the alternative when participation is endogenous. However, if participation can be encouraged by other means, then a more demanding conditional-offer bargaining game becomes preferable. Consequently, the benefit of offering "club benefits" (such as the lower tariffs in Nordhaus (2015)) is not, ultimately, that participation will increase, but that parties can choose a more ambitious bargaining procedure without fearing that participation will fall by too much.

REFERENCES

Abreu, D., and F. Gul (2000): "Bargaining and Reputation," *Econometrica* 68(1): 85-117.

Acemoglu, D., P. Aghion, L. Bursztyn, and D. Hemous (2012): "The Environment and Directed Technical Change," *American Economic Review* 102(1):131-66.

Aldy, J. E., S. Barrett, and R. N. Stavins (2003): "Thirteen plus one: a comparison of global climate policy architectures," *Climate Policy* 3(4): 373-97.

Andersson, O., C. Argenton, and J. W. Weibull (2018): "Robustness to Strategic Uncertainty in the Nash Demand Game," *Mathematical Social Sciences* 91:1-5.

Asheim, G. (1992): "A Unique Solution to n-Person Sequential Bargaining," *Games and Economic Behavior* 4: 169-81.

Bagwell, K., and R. W. Staiger (1999): "An Economic Theory of Gatt," *American Economic Review* 89(1): 215-48.

Barrett, S. (1994): "Self-enforcing international environmental agreements," *Oxford Economic Papers*, 46: 878-94.

Barrett, S. (2002): "Consensus Treaties," *Journal of Institutional and Theoretical Economics* 158(4): 529-47.

Barrett, S., and A. Dannenberg (2016): "An experimental investigation into 'pledge and review' in climate negotiations," *Climatic Change* 138(1): 339-51.

Battaglini, M., and B. Harstad (2016): "Participation and Duration of Environmental Agreements," *Journal of Political Economy* 124(1): 160-204.

Beccherle, J., and J. Tirole (2011): "Regional Initiatives and the Cost of Delaying Binding Climate Change Agreements," *Journal of Public Economics* 95: 1339-48.

Bernauer, T., A. Kalbhenn, V. Koubi, and G. Spilker (2013): "Is there a 'Depth versus Participation' Dilemma in International Cooperation?" *Review of International Organization* 8 (4): 477-97.

Binmore, K. (1987): "Nash Bargaining Theory (II)," *The Economics of Bargaining*, ed. by K. Binmore and P. Dasgupta (ed.). Cambridge: Basil Blackwell.

Binmore, K., M. J. Osborne, and A. Rubinstein (1992): "Non-Cooperative Models of Bargaining," *Handbook of Game Theory*, Volume 1, ed. by R. J. Aumann and S. Hart, Elsevier Science Publishers.

Binmore, K., A. Rubinstein, and A. Wolinsky (1986): "The Nash Bargaining Solutin in Economic Modelling," *The RAND Journal of Economics* 17(2): 176-88.

Bloch, F. (2018): "Coalitions and networks in oligopolies," *Handbook of Game Theory and Industrial Organization*, ed. by. L. Corchon and M. Marini, Edward Elgar.

Bodansky, D. (2010): "The Copenhagen Climate Change Conference: A Postmortem," *American Journal of International Law* 104(2): 230-40.

Bodansky, D., J. Brunnee, and L. Rajamani (2017): *International Climate Change Law*, Oxford University Press.

Bodansky, D., and L. Rajamani (2017): "Evolution and Governance Architecture of the Climate Change Regime," forthcoming in *Global Climate Policy: Actors, Concepts, and Enduring Challenges*, ed. by U. Luterbacher and D. Sprinz, MIT Press.

Britz, V., P. J. Herings, and A. Predtetchinski (2010): "Non-cooperative Support for the Asymmetric Nash Bargaining Solution," *Journal of Economic Theory* 145: 1951-67.

Bulow, J., and P. Klemperer (1996): "Auctions Versus Negotiations," *American Economic Review* 86(1): 180-94.

Bulow, J., and P. Klemperer (2009): "Why Do Sellers (Usually) Prefer Auctions?" *American Economic Review* 99(4): 1544-75.

Calvo, E., and S. J. Rubio (2012): "Dynamic Models of International Environmental Agreements: A Differential Game Approach," *International Review of Environmental and Resource Economics* 6: 289-339.

Caparrós, A. (2016): "Bargaining and International Environmental Agreements," *Environmental and Resource Economics* 65(1): 5-31.

Carlsson, H. (1991): "A Bargaining Model Where Parties Make Errors," *Econometrica* 59(5): 1487-96.

Carraro, C., and D. Siniscalco (1993): "Strategies for the International Protection of the Environment." *Journal of Public Econonomics* 52(3): 309–28.

Chae, S., and J. Yang (1994): "An N-person pure bargaining game", *Journal of Economic Theory* 62(1): 86-102.

d'Aspremont, C., A. Jacquemin, J. J. Gabszewicz, and J. A. Weymark (1983): "On the stability of collusive price leadership," *The Canadian Journal of Economics* 16(1): 17–25.

Dutta, P. K., and R. Radner (2004): "Self-enforcing climate-change treaties," *Proceedings of the National Academy of Science* 101: 4746-51.

Dutta, P. K., and R. Radner (2006): "A Game-Theoretic Approach to Global Warming," *Advances in Mathematical Economics* 8: 135-53.

Dutta, P. K., and R. Radner (2019): "The Paris Accord Can Be Effective if the Green Climate Fund is Effective," mimeo, Columbia University.

Falkner, R. (2016): "The Paris Agreement and the new logic of international climate politics," *International Affairs* 92(5): 1107-25.

Finus, M., and S. Maus (2008): "Modesty may pay!" *Journal of Public Economic Theory* 10: 801–26.

Gilligan, M. J. (2004): "Is There a Broader-Deeper Trade-off in International Multilateral Agreements?" *International Organization* 58 (3): 459-84.

Gollier, C., and J. Tirole (2015): "Making Climate Agreements Work," *The Economist*, guest blog, June 1st: https://www.economist.com/free-exchange/2015/06/01/making-climate-agreements-work

Golosov, M., J. Hassler, P. Krusell, and A. Tsyvinski (2014): "Optimal Taxes on Fossil Fuel in General Equilibrium," *Econometrica* 82(1): 41-88.

Harris, M., and B. Holmstrom (1987): "On The Duration of Agreements," *International Economic Review* 28(2): 389-406.

Harsanyi, J., and R. Selten (1972): "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information," *Management Science* 18(5) part 2: 80-106.

Harstad, B. (2012): "Climate Contracts: A Game of Emissions, Investments, Negotiations, and Renegotiations," *Review of Economic Studies* 79(4): 1527-57.

Harstad, B. (2016): "The Dynamics of Climate Agreements," *Journal of the European Economic Association* 14(3): 719-52.

Harstad, B. (2020): "Technology and Time Inconsinstency," *Journal of Political Economy*, forthcoming.

Harstad, B., F. Lancia, and A. Russo (2019): "Compliance Technology and Self-Enforcing Agreements," *Journal of the European Economic Association* 17(1): 1-30.

Hoel, M. (1992): "International environmental conventions: the case of uniform reductions of emissions," *Environmental and Resource Economics* 2(2): 141-59.

IPCC (2014): *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Kalai, E. (1977): "Non-symmetric Nash Solutions and Replication of 2-Person Bargaining," *International Journal of Game Theory* 6(3): 129-33.

Kambe, S. (2000): "Bargaining with Imperfect Commitment," *Games and Economic Behavior* 28: 217-37.

Keohane, R. O., and M. Oppenheimer (2016): "Paris: Beyond the Climate Dead End through Pledge and Review," *Politics and Governance* 4(3): 42-51.

Kolstad, C. D., and M. Toman (2005): "The Economics of Climate Policy," *Handbook of Environmental Economics* 3: 1562-93.

Krishna, V., and R. Serrano (1996): "Multilateral Bargaining", *Review of Economic Studies* 63(1): 61-80.

Laurelle, A., and F. Valenciano (2008): "Non-Cooperative Foundations of Bargaining Power in Committees and the Shapley-Shubik Index," *Games and Economic Behavior* 63: 341-53.

Martimort, D., and W. Sand-Zantman (2016): "A Mechanism Design Approach to Climate-change Agreements," *Journal of the European Economic Association* 14(3): 669-718.

McAfee, P. (1993): "Mechanism Design by Competing Sellers," *Econometrica* 61(6): 1281-1312.

Miyakawa, T. (2008): "Note on the Equal Split Solution in an n-Person Non-Cooperative Bargaining Game," *Mathematical Social Sciences* 55(3): 281-91.

Myerson, R. (1978): "Refinement of the Nash Equilibrium Concept," *International Journal of Game Theory* 7: 73-80.

Nash, J. (1950): "The Bargaining Problem," *Econometrica* 18: 155-62.

Nash, J. (1953): "Two-Person Cooperative Games," *Econometrica* 21(1): 128-40.
Nordhaus, W. D. (2015): "Climate Clubs: Overcoming Free-riding in International Climate Policy," *American Economic Review* 105(4): 1339-70.
OECD (2018): "Common time frames: Summary of discussions at the March 2018 Climate Change Expert Group Global Forum," Note prepared by the OECD/IEA Climate Change Expert Group.
Osborne, M. J., and A. Rubinstein (1990): *Bargaining and Markets*, Academic Press.
Roth, A. (1979): "Proportional Solutions to the Bargaining Problem," *Econometrica* 47(3): 775-78.
Rubinstein, A. (1982): "Perfect Equilibrium in a Bargaining Model," *Econometrica* 50(1): 97-109.
Rubinstein, A. (1985): "A bargaining model with incomplete information about time preferences," *Econometrica* 53(5): 1151-72.
Schmalensee, R. (1998): "Greenhouse policy architectures and institutions," *Economics and Policy Issues in Climate Change*, ed. by W. D. Nordhaus, Resources for the Future Press, Washington, D. C.
Segal, I. (1999): "Complexity and Renegotiation: A Foundation for Incomplete Contracts," *Review of Economic Studies* 66(1): 57-82.
Selten, R. (1975): "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory* 4: 25-55.
Simon, L. K. (1987): "Local Perfection," *Journal of Economic Theory* 43: 134-56.
Simon, L. K., and M. B. Stinchcombe (1995): "Equilibrium Refinement for Infinite Normal-Form Games," *Econometrica* 63(6): 1421-43.
Stern, N. (2006): *The Economics of Climate Change: The Stern Review*, Cambridge University Press
Sutton, J. (1986): "Non-Cooperative Bargaining Theory: An Introduction," *The Review of Economic Studies* 53(5): 709-24.
Victor, D. (2015): "Why Paris Worked: A Different Approach to Climate Diplomacy," Yale Envir. 360: https://e360.yale.edu/features/why_paris_worked_a_different_approach_to_climate_diplomacy
Yildiz, Muhamet (2003): "Walrasian bargaining," *Games and Economic Behavior* 45(2): 465-87.

Appendix A: Proofs of Theorems

PROOF OF THEOREM 1: As advertised in Section I, the following generalization of Theorem 1 is here proven without the additional assumptions $\partial U_i(\cdot)/\partial x_i < 0$ for $x_i > 0$, and $\partial U_j(\cdot)/\partial x_i > 0$, $j \neq i$.

THEOREM A-1: *If $\mathbf{x}^*$ is a nontrivial MPE in which $U_i(\mathbf{x}^*) > 0 \forall i$, then, for every $i \in N$, we have:*
(a) if $\partial U_i(\mathbf{x}^*)/\partial x_i \leq 0$,

$$-\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{\rho_i U_i(\mathbf{x}^*)} \leq \sum_{j\neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j U_j(\mathbf{x}^*)}\right\} f_j(0)\, \mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0); \tag{19}$$

(b) if $\partial U_i(\mathbf{x}^*)/\partial x_i > 0$,

$$\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{\rho_i U_i(\mathbf{x}^*)} \leq \sum_{j\neq i} \max\left\{0, -\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j U_j(\mathbf{x}^*)}\right\} f_j(0)\, \mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0).$$

With the additional assumptions $\partial U_i(\cdot)/\partial x_i < 0$, $\partial U_j(\cdot)/\partial x_i > 0$, and the constraint $x_i \geq 0$, $j \neq i$, (19) is equivalent to the first-order condition of (3).

PROOF OF PART (a). First, note that in any MPE we must have $U_i(\mathbf{x}^*) \geq 0 \forall i$, since otherwise a party with $U_i(\mathbf{x}^*) < 0$ would reject $\mathbf{x}^*$ in order to obtain the default payoff, normalized to zero. I will search for nontrivial equilibria in which $U_i(\mathbf{x}^*) > 0 \forall i$.

Next, consider an equilibrium $\mathbf{x}^*$, satisfying $U_j(\mathbf{x}^*) > 0 \forall j$. When $\mathbf{x}^*$ is proposed, it will be accepted with probability 1 since $\rho_{j,t} \geq 0$. Therefore, i will never offer $x_i > x_i^*$ when $\frac{\partial U_i(\mathbf{x}^*)}{\partial x_i} \leq 0$, so to check when $\mathbf{x}^*$ is an equilibrium, it is sufficient to consider a deviation by i, $\mathbf{x}^i$, such that $x_i^i < x_i^*$ while $x_j^i = x_j^*$, $j \neq i$.

Acceptable offers.– Let $P(\mathbf{x}^i; \mathbf{x}^*)$ be the probability that at least one $j \neq i$ rejects $\mathbf{x}^i$, and $P_{-j}(\mathbf{x}^i; \mathbf{x}^*)$ the probability that at least one party other than j and i rejects $\mathbf{x}^i$, given an equilibrium $\mathbf{x}^*$.

Since party j's discount factor is $\delta_{j,t}^{\Delta} \equiv 1 - \rho_{j,t}\Delta = 1 - \theta_{j,t}\rho_j\Delta$, $j \neq i$ rejects $\mathbf{x}^i$ iff:

$$\begin{aligned} \left(1 - P_{-j}(\mathbf{x}^i)\right) U_j(\mathbf{x}^i) + P_{-j}(\mathbf{x}^i)\left(1 - \rho_{j,t}\Delta\right) U_j(\mathbf{x}^*) &< \left(1 - \rho_{j,t}\Delta\right) U_j(\mathbf{x}^*) \Longleftrightarrow \\ \theta_{j,t} < \widetilde{\theta}_j(\mathbf{x}^i) &\equiv \max\left\{0, \frac{U_j(\mathbf{x}^*) - U_j(\mathbf{x}^i)}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\}. \end{aligned} \tag{20}$$

Note on the derivative.– Since we only need to consider $x_i < x_i^*$ and U_j is concave, $U_j(\mathbf{x}^*) < U_j(\mathbf{x}^i) \Leftrightarrow \partial U_j(\mathbf{x}^i)/\partial x_i < 0$. In this case, j accepts $\mathbf{x}^i$ with probability 1, $\widetilde{\theta}_j(\mathbf{x}^i) = 0$, and $\partial\widetilde{\theta}_j(\mathbf{x}^i)/\partial x_i = 0$. If, instead, $U_j(\mathbf{x}^*) > U_j(\mathbf{x}^i)$, $\partial\widetilde{\theta}_j(\mathbf{x}^i)/\partial x_i = \left[-\partial U_j(\mathbf{x}^i)/\partial x_i\right]/\rho_j \Delta U_j(\mathbf{x}^*) < 0$. Combined:

$$\frac{\partial\widetilde{\theta}_j(\mathbf{x}^i)}{\partial x_i} = -\max\left\{0, \frac{\partial U_j(\mathbf{x}^i)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} \leq 0.$$

When the joint pdf of shocks $\boldsymbol{\theta}_t = (\theta_{1,t}, ..., \theta_{n,t})$ is represented by $f(\boldsymbol{\theta}_t)$, the probability that every $j \neq i$ accepts $\mathbf{x}^i$ can be written as:

$$\begin{aligned} 1 - P(\mathbf{x}^i) &= G\left(\widetilde{\theta}_1(\mathbf{x}^i), ..., \widetilde{\theta}_{i-1}(\mathbf{x}^i), \widetilde{\theta}_{i+1}(\mathbf{x}^i), ...\widetilde{\theta}_n(\mathbf{x}^i)\right) \\ &\equiv \int_0^{\overline{\theta}_i}\left[\int_{\widetilde{\theta}_1(\mathbf{x}^i)}^{\overline{\theta}_1} ... \int_{\widetilde{\theta}_{i-1}(\mathbf{x}^i)}^{\overline{\theta}_{i-1}} \int_{\widetilde{\theta}_{i+1}(\mathbf{x}^i)}^{\overline{\theta}_{i+1}} ... \int_{\widetilde{\theta}_n(\mathbf{x}^i)}^{\overline{\theta}_n} f(\boldsymbol{\theta}_t)\, d\boldsymbol{\theta}_{-i,t}\right] d\theta_i, \end{aligned} \tag{21}$$

which is a function of $n-1$ thresholds (20). If we take the derivative of (21) w.r.t. x_i^i and use the chain rule, we get:

$$-\frac{\partial P(\mathbf{x}^i)}{\partial x_i} = \sum_{j\neq i} -\max\left\{0, \frac{\partial U_j(\mathbf{x}^i)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} G_j'\left(\widetilde{\theta}_1(\mathbf{x}^i), ..., \widetilde{\theta}_{i-1}(\mathbf{x}^i), \widetilde{\theta}_{i+1}(\mathbf{x}^i), ...\widetilde{\theta}_n(\mathbf{x}^i)\right),$$

and, at the equilibrium, $\mathbf{x}^i = \mathbf{x}^*$, we get:

$$\frac{\partial P(\mathbf{x}^*)}{\partial x_i} = \sum_{j \neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} G_j'(\mathbf{0}) = -\sum_{j \neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} f_j(0),$$

where, as written in the text already, $f_j(0)$ is the marginal distribution of $\theta_{j,t}$ at $\theta_{j,t} = 0$.

Equilibrium offers.– When proposing x_i, party i's problem is to choose $x_i \leq x_i^*$ so as to maximize

$$\left(1 - P\left(\mathbf{x}^i\right)\right) U_i\left(\mathbf{x}^i\right) + P\left(\mathbf{x}^i\right)\left(1 - \mathrm{E}\theta_{i,t}^R \rho_i \Delta\right) U_i(\mathbf{x}^*), \tag{22}$$

where $\mathrm{E}\theta_{i,t}^R$ is the expected $\theta_{i,t}$ conditional on being rejected (this will be more precise in eq. (25)).

To derive the first-order condition w.r.t. x_i^i, suppose i considers a small (marginal) reduction in x_i relative to x_i^*, given by $dx_i = x_i^i - x_i^* < 0$. If accepted, this gives i utility

$$U_i\left(\mathbf{x}^i\right) \approx U_i(\mathbf{x}^*) + dx_i \partial U_i(\mathbf{x}^*)/\partial x_i > U_i(\mathbf{x}^*), \tag{23}$$

but $\mathbf{x}^i$ is rejected with probability

$$P\left(\mathbf{x}^i\right) \approx P(\mathbf{x}^*) + \frac{\partial P(\mathbf{x}^*)}{\partial x_i} dx_i = 0 - \sum_{j \neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} dx_i f_j(0), \tag{24}$$

where each of the $n-1$ terms represents the probability that a $\theta_{j,t}$ is so small that j rejects if x_i is modified by dx_i, i.e., $\Pr\left(\theta_{j,t} \leq \widehat{\theta}_j\right)$ for $\widehat{\theta}_j \equiv \frac{\partial U_j(\mathbf{x}^i)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)} |dx_i|$. Naturally, the probability that more than one party has such a small shock vanishes when $|dx_i| \to 0$ since f is assumed to have no mass point.

If we combine (22), (23), and (24), we find party i's expected payoff when proposing x_i^i. This payoff, written on the left-hand side in the following inequality, is smaller than i's payoff if i proposes x_i^* if and only if:

$$\begin{aligned} &\left(1 + \sum_{j \neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} f_j(0) dx_i\right)\left(U_i(\mathbf{x}^*) + dx_i \frac{\partial U_i(\mathbf{x}^*)}{\partial x_i}\right) \\ &- \sum_{j \neq i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} dx_i f_j(0)\left(1 - E\left(\theta_{i,t} \mid \theta_{j,t} \leq \widehat{\theta}_{j,t}\right) \rho_i \Delta\right) U_i(\mathbf{x}^*) \quad \leq \quad U_i(\mathbf{x}^*), \end{aligned} \tag{25}$$

where $\mathrm{E}\left(\theta_{i,t} \mid \theta_{j,t} \leq \widehat{\theta}_j\right)$ follows from Bayes' rule:

$$\mathrm{E}\left(\theta_{i,t} \mid \theta_{j,t} \leq \widehat{\theta}_j\right) \equiv \frac{\int_0^{\widehat{\theta}_j} \int_{\Theta_{-j}} \theta_{i,t} f(\boldsymbol{\theta}_t) d\theta_j}{\int_0^{\widehat{\theta}_j} \int_{\Theta_{-j}} f(\boldsymbol{\theta}_t) d\theta_j}, \text{ and } \mathrm{E}\left(\theta_{i,t} \mid \theta_{j,t} = 0\right) \equiv \lim_{dx_i \uparrow 0} \frac{\int_0^{\widehat{\theta}_j} \int_{\Theta_{-j}} \theta_{i,t} f(\boldsymbol{\theta}_t) d\theta_j}{\int_0^{\widehat{\theta}_j} \int_{\Theta_{-j}} f(\boldsymbol{\theta}_t) d\theta_j},$$

and, as already defined, $\Theta_{-j} \equiv \prod_{k \neq j} [0, \overline{\theta}_k]$ and $\widehat{\theta}_j \equiv \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)} |dx_i|$.

When both sides of (25) are divided by $|dx_i|$ and $dx_i \uparrow 0$, (25) can be rewritten as the first-order condition (19).

The proof of part (b) is analogous and thus omitted. *Q.E.D.*

PROOF OF THEOREM 2: A continuum of $\mathbf{x}^*$'s can satisfy the equilibrium condition in Theorem A-1. To provide an illustration of this, note that if (19) binds then (19) continue to be satisfied when x_i^* is reduced. The idea of local perfection is to introduce trembles such that equilibrium offers can be rejected (i.e., $P(\mathbf{x}^*) > 0$) and thus we must check that i cannot benefit from marginally increasing *or* decreasing x_i^i from x_i^* to reduce $P\left(\mathbf{x}^i\right)$. With trembles, i will strictly benefit from $dx_i > 0$ when (3) is strict, and thus it must hold with equality at $\mathbf{x}^*$.

The vector s_t is i.i.d. over time according to some cdf, $H(\cdot)$, with is assumed to have a bounded support and $\partial H(\mathbf{0})/\partial s_{i,t} > 0$ on a neighborhood of $\mathbf{0}$. When j considers whether to accept $U_j\left(\mathbf{x}^i + \epsilon \mathbf{s}_t\right)$, j faces

the continuation value $V_j(\mathbf{x}^*)$ by rejecting, where $V_j(\mathbf{x}^*)$ takes into account that $\mathbf{x}^*$ may be rejected in the future (if the future s_t's are sufficiently small):

To write the equation for $V_j(\mathbf{x}^*)$, note that it is the *combination* of the $s_{i,t}$'s and the $\theta_{j,t}$'s that determines whether j rejects $\mathbf{x}^*$: let $\Phi_A(\mathbf{x}^*)$ be the set of $s_{i,t}$'s and $\theta_{j,t}$'s such that every j accepts $\mathbf{x}^*$, while $\Phi_R(\mathbf{x}^*)$ is the complementary set.[36] We then have $P(\mathbf{x}^*) = \Pr\{(\mathbf{s}_t, \boldsymbol{\theta}_t) \in \Phi_R(\mathbf{x}^*)\}$, where $\boldsymbol{\theta}_t = (\theta_{1,t}, ..., \theta_{n,t})$, and:

$$V_j(\mathbf{x}^*) = \mathrm{E}_{\mathbf{s}_t:(\mathbf{s}_t,\boldsymbol{\theta}_t)\in\Phi_A(\mathbf{x}^*)}(1 - P(\mathbf{x}^*))\, U_j(\mathbf{x}^* + \epsilon\mathbf{s}_t) + P(\mathbf{x}^*)\, V_j(\mathbf{x}^*)\, \mathrm{E}_{\theta_{j,t}:(\mathbf{s}_t,\boldsymbol{\theta}_t)\in\Phi_R(\mathbf{x}^*)}\left(1 - \theta_{j,t}\rho_j\Delta\right). \tag{26}$$

The shocks, combined with the possibility to reject, imply that $V_j(\mathbf{x}^*) > 0$ even if $U_j(\mathbf{x}^*) = 0$, so there is no need to assume $U_j(\mathbf{x}^*) > 0 \forall j$ or restrict attention to "nontrivial" MPEs.

With this, party $j \neq i$ rejects $\mathbf{x}^i$ if and only if:

$$\left(1 - P_{-j}\left(\mathbf{x}^i\right)\right) U_j\left(\mathbf{x}^i + \epsilon\mathbf{s}_t\right) + P_{-j}\left(\mathbf{x}^i\right)\left(1 - \rho_{j,t}\Delta\right) V_j(\mathbf{x}^*) < \left(1 - \rho_{j,t}\Delta\right) V_j(\mathbf{x}^*) \Longleftrightarrow$$

$$1 - \theta_{j,t}\rho_j\Delta > \frac{U_j\left(\mathbf{x}^i + \epsilon\mathbf{s}_t\right)}{V_j(\mathbf{x}^*)} \Longleftrightarrow \theta_{j,t} < \widetilde{\theta}_j\left(\mathbf{x}^i\right) \equiv \frac{V_j(\mathbf{x}^*) - U_j\left(\mathbf{x}^i + \epsilon\mathbf{s}_t\right)}{\rho_j\Delta V_j(\mathbf{x}^*)}. \tag{27}$$

Here, $\widetilde{\theta}_j\left(\mathbf{x}^i\right)$ is a function of s_t. To simplify the notation, I assume $\widetilde{\theta}_j\left(\mathbf{x}^i\right) \in \left(0, \overline{\theta}_i\right)$ for every s_t when $dx_i < 0$ is small (this is natural since we are considering small trembles when $\epsilon \to 0$). The probability that every $j \neq i$ accepts can then be written as:

$$\begin{aligned}
1 - P\left(\mathbf{x}^i\right) &= \int_{\mathbf{s}_t} G\left(\widetilde{\theta}_1\left(\mathbf{x}^i\right), ..., \widetilde{\theta}_{i-1}\left(\mathbf{x}^i\right), \widetilde{\theta}_{i+1}\left(\mathbf{x}^i\right), ...\widetilde{\theta}_n\left(\mathbf{x}^i\right)\right) dH(\mathbf{s}_t) \\
&\equiv \int_{\mathbf{s}_t}\int_0^{\overline{\theta}_i}\left[\int_{\widetilde{\theta}_1(\mathbf{x}^i)}^{\overline{\theta}_1} ... \int_{\widetilde{\theta}_{i-1}(\mathbf{x}^i)}^{\overline{\theta}_{i-1}}\int_{\widetilde{\theta}_{i+1}(\mathbf{x}^i)}^{\overline{\theta}_{i+1}} ... \int_{\widetilde{\theta}_n(\mathbf{x}^i)}^{\overline{\theta}_n} f(\boldsymbol{\theta}_t)\, d\boldsymbol{\theta}_{-i,t}\right] d\theta_i dH(\mathbf{s}_t) \Rightarrow \\
-\frac{\partial P\left(\mathbf{x}^i\right)}{\partial x_i} &= \mathrm{E}_{\mathbf{s}_t}\sum_{j\neq i} -\frac{\partial U_j\left(\mathbf{x}^i + \epsilon\mathbf{s}_t\right)/\partial x_i}{\rho_j\Delta V_j(\mathbf{x}^*)} G_j'\left(\widetilde{\theta}_1\left(\mathbf{x}^i\right), ..., \widetilde{\theta}_{i-1}\left(\mathbf{x}^i\right), \widetilde{\theta}_{i+1}\left(\mathbf{x}^i\right), ...\widetilde{\theta}_n\left(\mathbf{x}^i\right)\right).
\end{aligned}$$

The condition under which i does not benefit from a marginal change $dx_i < 0$ is given by an equation that is similar to (25), although we now have to take into account the trembles:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{s}_t:(\mathbf{s}_t,\boldsymbol{\theta}_t)\in\Phi_A(\mathbf{x}^*)}\left(1 - P(\mathbf{x}^*) - \frac{\partial P(\mathbf{x}^*)}{\partial x_i}dx_i\right)\left(U_i(\mathbf{x}^* + \epsilon\mathbf{s}_t) + \frac{\partial U_i(\mathbf{x}^* + \epsilon\mathbf{s}_t)}{\partial x_i}dx_i\right) + \\
&\mathrm{E}_{\mathbf{s}_t}\sum_{j\neq i}\left[\frac{\partial U_j(\mathbf{x}^* + \epsilon\mathbf{s}_t)/\partial x_i}{\rho_j\Delta V_j(\mathbf{x}^*)}dx_i G_j'\left(\frac{V_1(\mathbf{x}^*) - U_1(\mathbf{x}^* + \epsilon\mathbf{s}_t)}{\rho_1\Delta V_1(\mathbf{x}^*)}, \frac{V_2(\mathbf{x}^*) - U_2(\mathbf{x}^* + \epsilon\mathbf{s}_t)}{\rho_2\Delta V_2(\mathbf{x}^*)}, ...; \theta_i\right)\right]. \\
&\mathrm{E}_{\theta_{i,t}|\theta_{j,t}<\widetilde{\theta}_j(\mathbf{x}^i)}\left(1 - \theta_{i,t}\rho_i\Delta\right) V_i(\mathbf{x}^*) \leq V_i(\mathbf{x}^*).
\end{aligned} \tag{28}$$

Since the trembles imply that $P(\mathbf{x}^*) > 0$, i might benefit from reducing this risk and consider a marginal increase $dx_i > 0$. Party i will not benefit from $dx_i > 0$ if (28) holds with the reverse inequality sign ($\geq$), it is easy to show. Consequently, (28) must hold with equality for no marginal deviation to be beneficial to i. (Note that (28) must hold with equality regardless of whether $U_i(\cdot)$ would increase when $dx_i > 0$ or when $dx_i < 0$, so, we do not need the assumptions $\partial U_i(\cdot)/\partial x_j > 0$ for $j \neq i$ and < 0 for $j = i$.)

[36] By referring to (27), below, $\Phi_A(\mathbf{x}^*)$ and $\Phi_R(\mathbf{x}^*)$ are defined as:

$$\begin{aligned}
\Phi_A(\mathbf{x}^*) &= \left\{(\mathbf{s}_t, \boldsymbol{\theta}_t) : \theta_{j,t} \geq \frac{V_j(\mathbf{x}^*) - U_j(\mathbf{x}^* + \epsilon\mathbf{s}_t)}{\rho_j\Delta V_j(\mathbf{x}^*)} \forall j\right\}, \\
\Phi_R(\mathbf{x}^*) &= \left\{(\mathbf{s}_t, \boldsymbol{\theta}_t) : \theta_{j,t} < \frac{V_j(\mathbf{x}^*) - U_j(\mathbf{x}^* + \epsilon\mathbf{s}_t)}{\rho_j\Delta V_j(\mathbf{x}^*)} \text{ for at least one } j\right\}.
\end{aligned}$$

When we let $\epsilon \to 0$, so that the trembles vanish, then we can see from (26) and (27) that $P(\mathbf{x}^*) \to 0$ and $V_j(\mathbf{x}^*) \to U_j(\mathbf{x}^*)$. When these limits are substituted into (28), holding with equality, and we divide both sides by dx_i before we let $dx_i \to 0$ and $\epsilon s_t \to 0$, then (28) can be rewritten as:

$$\frac{\partial U_i(\mathbf{x}^*)}{\partial x_i} + \sum_{j \neq i} \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)} f_j(0) \, \mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0) \, \rho_i \Delta U_i(\mathbf{x}^*) = 0, \tag{29}$$

which coincides with the first-order condition of

$$\arg\max_{x_i} \prod_{j \in N} \left(U_j\left(x_i, \mathbf{x}^*_{-i}\right) \right)^{w^i_j},$$

when $\frac{w^i_j}{w^i_i} = \frac{\rho_i}{\rho_j} f_j(0) \mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0)$, $\forall j \neq i$. *Q.E.D.*

PROOF OF COROLLARY 2: With (5), a binding (3) implies:

$$\begin{aligned}
x^*_i &= \arg\max_{x_i} \prod_{j \in N} \left(v_j(x_j) \, p(\mathbf{x}) \right)^{w^i_j} = \arg\max_{x_i} v_i(x_i) \, p(\mathbf{x})^{\sum_j w^i_j / w^i_i} \\
&= \arg\max_{x_i} v_i(x_i)^{w^i_i / \sum_j w^i_j} \, p(\mathbf{x}) = \arg\max_{x_i} \prod_{j \in N} v_j(x_j)^{w^j_j / \sum_k w^j_k} \, p(\mathbf{x}), \text{ so} \\
\mathbf{x}^* &= \arg\max_{\mathbf{x}} \prod_{j \in N} v_j(x_j)^{w^j_j / \sum_k w^j_k} \, p(\mathbf{x}),
\end{aligned}$$

which can be written as (6), given the definitions ϱ_i and ω. Given $\mathbf{x}^*$, (6) can be rewritten as (7). *Q.E.D.*

For Online Publication
Appendix B: P&R Bargaining and Uncertainty

This online appendix builds on the proof of Theorem A-1 to investigate conditions under which contributions can be positive with P&R bargaining and how the outcome can be characterized under alternative assumptions. In short, I show that contributions can be positive even if $f_j(0) = 0$ if instead either $\Delta \to 0$ or if there is a boundary for how small the reduction in x_i might be. For simplicity, I make the additional assumptions that increasing $x_i > 0$ is costly for i but beneficial for everyone else.

No uncertainty.– I start with the basic situation in which there is no uncertainty on the discount rates. Consider the restriction that $x_i = \Delta_i^x \varsigma$, where ς can be any positive integer. That is, if i reduces x_i from x_i^*, i must reduce x_i by at least the amount Δ_i^x. For example, if x_i must be described by a real number with at most ϑ_i decimals, then $\Delta_i^x = 1/10^{\vartheta_i}$. I am especially interested in the limit $\Delta_i^x \to 0$, so that x_i can approximate any real number. If both $\Delta_i^x \to 0$ and $\Delta \to 0$, $\chi_i \equiv \Delta_i^x/\Delta$ might be a finite and strictly positive number.

If i deviates by offering $x_i^i = x_i^* - \Delta_i^x$, then j rejects if and only if:

$$U_j\left(\mathbf{x}^i\right) < \left(1 - \rho_j \Delta\right) U_j\left(\mathbf{x}^*\right).$$

When Δ_i^x is small, this inequality is approximated as:

$$\begin{aligned} U_j\left(\mathbf{x}^i\right) &= U_j\left(\mathbf{x}^*\right) - \frac{\partial U_j\left(\mathbf{x}^*\right)}{\partial x_i}\Delta_i^x < \left(1 - \rho_j\Delta\right) U_j\left(\mathbf{x}^*\right) \Leftrightarrow \\ &\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{U_j\left(\mathbf{x}^*\right)} > \frac{\rho_j}{\chi_i}. \end{aligned} \tag{30}$$

Thus, for $\mathbf{x}^*$ to be an equilibrium, $\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{U_j(\mathbf{x}^*)}$ cannot be very small for every j, since then every j would have accepted a small reduction in x_i instead of waiting for $\mathbf{x}^*$. However, the condition does not rule out that x_i^* can be above i's preferred level: if j anticipates $\mathbf{x}^* \gg 0$, then j will reject a smaller x_i whenever x_i^* is so small that $\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{U_j(\mathbf{x}^*)}$ is larger than ρ_j/χ_i.

THEOREM B-1: *Consider a situation with no uncertainty. If $U_i(\mathbf{x}^*) > 0 \forall i \in N$, $\mathbf{x}^*$ can be a part of a nontrivial MPE if and only if for every $i \in N$, there exists some $j \neq i$ such that:*

$$1 < \frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j U_j\left(\mathbf{x}^*\right)}\chi_i. \tag{31}$$

Intuitively, with P&R bargaining, i is willing to contribute beyond i's bliss point if j is willing to reject a reduction of Δ_i^x. The number Δ_i^x can be arbitrarily close to zero if also Δ is close to zero. For any $\chi_i \equiv \Delta_i^x/\Delta \in (0, \infty)$, the right-hand side of (31) will grow when the contributions fall since then $\partial U_j\left(\mathbf{x}^*\right)/\partial x_i$ grows while $U_j\left(\mathbf{x}^*\right)$ approaches zero. For sufficiently small (but positive) contributions, (31) holds.

Uncertainty under alternative assumptions.– Section I assumed $\rho_{j,t} = \theta_{j,t}^{\rho}\rho_j$ (although the shock $\theta_{j,t}^{\rho}$ was then referred to as $\theta_{j,t}$). We might also consider the possibility that j's expectation over the lag before the next acceptance stage is $\Delta_{j,t} = \theta_{j,t}^{D}\Delta$, where Δ is the common mean for this expectation, while $\theta_{j,t}^{D}$ is a shock with mean 1. This shock might capture a situation in which the delay or lag before the next proposal stage is unknown and different parties obtain different subjective beliefs regarding what the lag will be. Similarly, with a stochastic $\frac{\partial U_j^{\theta}(\mathbf{x}^*)/\partial x_i}{U_j^{\theta}(\mathbf{x}^*)}$, suppose we can write $\frac{\partial U_j^{\theta}(\mathbf{x}^*)/\partial x_i}{U_j^{\theta}(\mathbf{x}^*)} = \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{U_j(\mathbf{x}^*)}/\theta_{j,t}^{U}$, where $\mathrm{E}\left(1/\theta_{j,t}^{U}\right) = 1$. Here, $\theta_{j,t}^{U}$ can be interpreted as a shock that influences j marginal utility of x_i, j's absolute level of utility, or both. All shocks are realized and observed after offers but before acceptance decisions are made, and all shocks are i.i.d. over time.[37] As will be shown in the proof below, the rejection condition

[37] Admittedly, the sources of the various shocks are here simply black boxes. A more serious future investigation should provide a careful micro-foundation for the shocks and relate them to the primitives of the model as well as to real-world evidence.

becomes uncertain in the presence of *any* of these three shocks (or with two or all three of them): of importance is the product of the three shocks:

$$\theta_{j,t} \equiv \theta^{\rho}_{j,t}\theta^{D}_{j,t}\theta^{U}_{j,t}.$$

The $\theta_{j,t}$'s are assumed to be jointly distributed according to F, as before. Clearly, the support of $\theta_{j,t}$ will include zero as long as zero is included in the support of at least one of the three shocks. I will say that there is no uncertainty if every $\theta^{\rho}_{j,t}$, $\theta^{D}_{j,t}$, and $\theta^{U}_{j,t}$ is deterministic.

The condition under which j rejects, (30), can now be written as:

$$\theta_{j,t} \equiv \theta^{\rho}_{j,t}\theta^{D}_{j,t}\theta^{U}_{j,t} < \widetilde{\theta}_j\left(\mathbf{x}^i\right) \equiv \frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j U_j\left(\mathbf{x}^*\right)}\chi_i, \tag{32}$$

replacing (20). With this definition of $\widetilde{\theta}_j\left(\mathbf{x}^i\right)$, we can define the cdf G just as in (21). This G, which is the probability that (32) fails for every j (i.e., everyone accepts the deviation $\mathbf{x}^i$), is clearly a function of Δ^x_i. Write this function as $G_{i,\mathbf{x}^*}\left(\Delta^x_i\right)$.

As in the proof of Theorem A-1, i seeks to maximize (22). For $\mathbf{x}^*$ to be part of an MPE, i cannot benefit from proposing marginally less. Party i does not benefit from offering the marginal amount Δ^x_i less if and only if:

$$\begin{aligned}\mathrm{E}\left[U_i\left(\mathbf{x}^*\right) - \left(\partial U_i\left(\mathbf{x}^*\right)/\partial x_i\right)\Delta^x_i\right]G_{i,\mathbf{x}^*}\left(\Delta^x_i\right) + \left(1 - G_{i,\mathbf{x}^*}\left(\Delta^x_i\right)\right)U_i\left(\mathbf{x}^*\right)\left(1-\rho_{i,t}\Delta_{i,t}\right) \quad &< \quad \mathrm{E}U_i\left(\mathbf{x}^*\right) \Leftrightarrow \\ \left(-\frac{\partial U_i\left(\mathbf{x}^*\right)/\partial x_i}{U_i\left(\mathbf{x}^*\right)\rho_i}\right)\chi_i \leq \frac{1 - G_{i,\mathbf{x}^*}\left(\Delta^x_i\right)}{G_{i,\mathbf{x}^*}\left(\Delta^x_i\right)}&. \end{aligned} \tag{33}$$

The right-hand side of (33) is a positive number when $\chi_i > 0$ as long as it is possible that $\theta_{j,t}$ is small enough to satisfy (32) for some $j \neq i$. When all contributions fall, the right-hand side of (32) increases and approaches infinity when $U_j\left(\mathbf{x}^*\right) \to 0$, so naturally (32) will be satisfied before all contributions are zero.

THEOREM B-2: *Consider a situation with uncertainty and a nontrivial MPE in which $U_i\left(\mathbf{x}^*\right) > 0 \forall i$. For every $i \in N$, (33) holds.*

The theorem limits how large the contributions can be. However, strictly positive contributions can be supported in equilibrium for the same reason as in Section I: Any deviation by i may be rejected by one of the opponents with a sufficiently large probability. As above, $\Delta_{i,t}$ can be arbitrarily small if also Δ is small. Intuitively, if the contributions and payoffs are small, it doesn't take much for a party to reject an offer if the party, in return, can expect a marginally better offer quite soon. Thus, the threshold $\widetilde{\theta}_j$ is strictly positive and it does not approach zero even if $\Delta^x_i \to 0$, if just $\chi_i \equiv \Delta^x_i/\Delta > 0$. On the contrary, if $\chi_i > 0$, $\widetilde{\theta}_j$ grows without bounds when contributions and payoffs become small.

These results prove that the qualitative result of Section I– that P&R bargaining can lead to positive contributions– does not hinge on the assumption that the discount rate can be arbitrarily close to zero. However, the assumptions in Section I are helpful because the outcome simplifies and it can be related to the ANBS in a way that is not possible under the alternative assumptions considered here.

To see this, the proof below shows that a second-order Taylor expansion of the right-hand side of (33) implies:

$$\begin{aligned}-\frac{\partial U_i\left(\mathbf{x}^*\right)/\partial x_i}{U_i\left(\mathbf{x}^*\right)\rho_i} \quad \leq \quad & \sum_{j\neq i} f_j\left(\mathbf{0}\right)\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j U_j\left(\mathbf{x}^*\right)} \\ & + \frac{\chi_i}{2}\sum_{j\neq i}\sum_{k\neq i}\frac{\partial f_j\left(\mathbf{0}\right)}{\partial\widetilde{\theta}_k}\left(\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j U_j\left(\mathbf{x}^*\right)}\right)\left(\frac{\partial U_k\left(\mathbf{x}^*\right)/\partial x_i}{\rho_k U_k\left(\mathbf{x}^*\right)}\right) \\ & + \chi_i\left(\sum_{j\neq i}\frac{\partial f_j\left(\mathbf{0}\right)}{\partial\widetilde{\theta}_j}\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i/}{\rho_j U_j\left(\mathbf{x}^*\right)}\right)^2.\end{aligned}$$

If $\chi_i \to 0$, the last two terms are zero and we are left with the same condition as in Theorem A-1(a). If instead $f_j(\mathbf{0}) \to 0$, the first term on the right-hand side is zero. The second term is zero if shocks are uncorrelated, and, in that case, we are left with the final term. The inequality can then be written as:

$$-\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{U_i(\mathbf{x}^*)\rho_i} \leq \chi_i \left(\sum_{j \neq i} \frac{\partial f_j(\mathbf{0})}{\partial \widetilde{\theta}_j} \frac{\partial U_j(\mathbf{x}^*)/\partial x_i/}{\rho_j U_j(\mathbf{x}^*)} \right)^2.$$

Here, the right-hand side is positive (and positive contributions can be supported) even if $f_j(\mathbf{0}) = 0$ if just $\partial f_j(\mathbf{0})/\partial \widetilde{\theta}_j > 0$. The fact that the term on the right-hand side is quadratic implies that the outcome cannot easily be related to the ANBS.

COROLLARY B-1: *Consider Example E, $U_i(\mathbf{x}) = \alpha \sum_{j \neq i} x_j - \beta x_i^2/2$, and symmetric $\chi_i = \chi$.*
(i) Suppose there is no uncertainty. Symmetric positive contributions x_i^* can be a part of a nontrivial MPE if and only if:

$$x_i^* \in \left(0, \frac{\alpha(n-1)}{\beta} - \frac{\alpha}{\beta}\sqrt{(n-1)^2 - 2\beta\chi/\alpha\rho}\right).$$

(ii) Suppose there is uncertainty and $\chi \to 0$. If $x_i^* > 0$ is a part of a symmetric nontrivial MPE in which $U_i(\mathbf{x}^*) > 0 \forall i$, then:

$$x_i^* \leq (n-1)\frac{\alpha}{\beta} f^{\theta}(\mathbf{0}).$$

(iii) Suppose there is uncertainty, shocks are uncorrelated, and $f_j(0) \to 0 \forall j$. If $x_i^* > 0$ is part of a symmetric nontrivial MPE in which $U_i(\mathbf{x}^*) > 0 \forall i$, then the second-order Taylor approximation of (33) implies:

$$x_i^* \leq (n-1)\frac{\partial f_j(0)}{\partial \theta_j}\frac{\chi\alpha^2}{2\beta\rho}$$

The comparative static w.r.t. the mean discount rates, for example, is the same as in Section I. The above inequalities also give a new comparative static: If χ is larger (so that the time lag Δ goes to zero very fast relative to how finely one can set x_i), then the upper boundary for the thresholds is larger.

PROOF OF THEOREM B-2 AND COROLLARY B-1: From (21) we can define:

$$\begin{aligned} G'_{i,\mathbf{x}^*} &\equiv \frac{dG_{i,\mathbf{x}^*}(0)}{d\Delta_i^x} = \sum_{j \neq i} \frac{\partial G(\mathbf{0})}{\partial \widetilde{\theta}_j} \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}, \text{ and} \qquad (34) \\ G''_{i,\mathbf{x}^*} &\equiv \frac{d^2 G_{i,\mathbf{x}^*}(0)}{(d\Delta_i^x)^2} = \sum_{j \neq i} \frac{\partial G(\mathbf{0})}{\partial \widetilde{\theta}_j} \frac{\partial^2 U_j(\mathbf{x}^*)/(\partial x_i)^2}{\rho_j \Delta U_j(\mathbf{x}^*)} \\ &\quad + \sum_{j \neq i}\sum_{k \neq i} \frac{\partial^2 G(\mathbf{0})}{\partial \widetilde{\theta}_j \partial \widetilde{\theta}_k} \left(\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right)\left(\frac{\partial U_k(\mathbf{x}^*)/\partial x_i}{\rho_k \Delta U_k(\mathbf{x}^*)}\right). \end{aligned}$$

Consider a second-order Taylor expansion of the right-hand side of (33), $\frac{1-G}{G}$. To derive this, note that:

$$\begin{aligned} \frac{d}{d\Delta_i^x}\left(\frac{1-G}{G}\right) &= \frac{-G'G - (1-G)G'}{G^2} = \frac{-G'}{G^2}, \text{ and} \\ \frac{d^2}{(d\Delta_i^x)^2}\left(\frac{1-G}{G}\right) &= \frac{-G''G^2 + 2G'G'G}{G^4} = \frac{-G'' + 2G'G'/G}{G^2}. \end{aligned}$$

Therefore, the second-order Taylor expansion of the right-hand side of (33) is given by:

$$\frac{1 - G_{i,\mathbf{x}^*}(\Delta_i^x)}{G_{i,\mathbf{x}^*}(\Delta_i^x)} \approx \frac{1 - G_{i,\mathbf{x}^*}(0)}{G_{i,\mathbf{x}^*}(0)} + \frac{-G'_{i,\mathbf{x}^*}}{(G_{i,\mathbf{x}^*}(0))^2}\Delta_i^x + \frac{(\Delta_i^x)^2}{2}\left(\frac{-G''_{i,\mathbf{x}^*} + 2\left(G'_{i,\mathbf{x}^*}\right)^2/G_{i,\mathbf{x}^*}(0)}{(G_{i,\mathbf{x}^*}(0))^2}\right).$$

The first term is zero since $G_{i,\mathbf{x}^*}(0)=1$. If we substitute in for $G'_{i,\mathbf{x}^*}$ and $G''_{i,\mathbf{x}^*}$ using (34), we get:

$$\begin{aligned}\frac{1-G_{i,\mathbf{x}^*}(\Delta_i^x)}{G_{i,\mathbf{x}^*}(\Delta_i^x)} \approx & -\Delta_i^x\sum_{j\neq i}\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}\frac{\mathrm{E}\left(\partial U_j(\mathbf{x}^*)/\partial x_i\right)/U_j(\mathbf{x}^*)}{\rho_j\Delta} \\ & -\frac{(\Delta_i^x)^2}{2}\sum_{j\neq i}\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}\frac{\mathrm{E}\left(\partial^2 U_j(\mathbf{x}^*)/(\partial x_i)^2\right)/U_j(\mathbf{x}^*)}{\rho_j\Delta} \\ & -\frac{(\Delta_i^x)^2}{2}\sum_{j\neq i}\sum_{k\neq i}\frac{\partial^2 G(\mathbf{0})}{\partial\widetilde{\theta}_j\partial\widetilde{\theta}_k}\left(\frac{\mathrm{E}\left(\partial U_j(\mathbf{x}^*)/\partial x_i\right)/U_j(\mathbf{x}^*)}{\rho_j\Delta}\right)\left(\frac{\mathrm{E}\left(\partial U_k(\mathbf{x}^*)/\partial x_i\right)/U_k(\mathbf{x}^*)}{\rho_k\Delta}\right) \\ & +(\Delta_i^x)^2\left(\sum_{j\neq i}\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}\frac{\mathrm{E}\left(\partial U_j(\mathbf{x}^*)/\partial x_i\right)/U_j(\mathbf{x}^*)}{\rho_j\Delta}\right)^2.\end{aligned} \tag{35}$$

Note that the second term is zero when $\Delta_i^x\to 0$, even if $\Delta_i^x/\Delta\to\chi_i>0$.

If $\Delta_i^x/\Delta\to 0$, the third and fourth terms in (35) also become zero, so we are left with only the first term. When this term is substituted into (33), we arrive at

$$-\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{U_i(\mathbf{x}^*)\rho_i}\leq\sum_{j\neq i}\left(-\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}\right)\frac{\mathrm{E}\left(\partial U_j(\mathbf{x}^*)/\partial x_i\right)/U_j(\mathbf{x}^*)}{\rho_j},$$

which is the same condition as in Theorem A-1(a) since $-\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}=f_j(0)$.

If instead $-\frac{\partial G(\mathbf{0})}{\partial\widetilde{\theta}_j}=f_j(0)\approx 0$, so that the density of the shocks on $\widetilde{\theta}_j$ is zero when $\widetilde{\theta}_j\to 0$, then the first and fourth terms in (35) become zero, and we are left with only the third term. When we substitute this term into (33), and divide both sides by $\frac{\Delta_i^x}{\Delta}$, (33) becomes:

$$-\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{U_i(\mathbf{x}^*)\rho_i}\leq\frac{\chi_i}{2}\sum_{j\neq i}\sum_{k\neq i}\left(-\frac{\partial^2 G(\mathbf{0})}{\partial\widetilde{\theta}_j\partial\widetilde{\theta}_k}\right)\left(\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j U_j(\mathbf{x}^*)}\right)\left(\frac{\partial U_k(\mathbf{x}^*)/\partial x_i}{\rho_k U_k(\mathbf{x}^*)}\right),$$

where $-\frac{\partial^2 G(\mathbf{0})}{\partial\widetilde{\theta}_j\partial\widetilde{\theta}_k}=\frac{\partial f_j(0)}{\partial\theta_k}$. If the shocks are not correlated, $\frac{\partial f_j(0)}{\partial\theta_k}=0$ when $k\neq j$, and this inequality simplifies to:

$$-\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{U_i(\mathbf{x}^*)\rho_i}\leq\sum_{j\neq i}f'_j(0)\left(\frac{\mathrm{E}\left(\partial U_j(\mathbf{x}^*)/\partial x_i\right)/U_j(\mathbf{x}^*)}{\rho_j}\right)^2\frac{\chi_i}{2},$$

where the right-hand side is positive when some $f_j(\theta_j)$ is strictly convex at $\theta_j=0$. When this inequality is combined with $U_i(\mathbf{x})=\alpha\sum_{j\neq i}x_j-\beta x_i^2/2$, it can be rewritten to Corollary B-1. *Q.E.D.*

FOR ONLINE PUBLICATION

APPENDIX C: THE CLIMATE POLICY MODEL

I will start by reformulating the optimal control problem described in Section II.

LEMMA C-1**:** *Given the actual pledges,* $\mathbf{x}$*, and the future equilibrium pledges,* $\mathbf{x}^*$*, party* i*'s continuation value is* $V_{i,1}(\mathbf{x}) = V_{i,1}^{BAU} + U_i(\mathbf{x})$*, where:*

$$\begin{aligned} U_i(\mathbf{x}) &\equiv \max_{\{y_{i,t}\}_{t=1}^{T}} \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1} + \delta^T U_i(\mathbf{x}^*), \\ Y_{i,t+1} &\equiv Y_{i,t} + y_{i,t}, \text{ and } Y_{i,1} \equiv 0. \end{aligned}$$

The lemma permits the current pledges ($\mathbf{x}$) to be different from those expected in equilibrium in the subsequent commitment period (i.e., $\mathbf{x}^*$). Conveniently, the heterogeneous bliss points and initial technology levels drop out when utility is measured relative to BAU. It is also convenient that the investments' effects on $Y_{j,T+1}$ are captured in terms that do not interact with the future continuation value, $\delta^T U_i(\mathbf{x}^*)$. The additional investments affect the future $V_{i,1}^{BAU}$ but not $U_i(\mathbf{x})$.

PROOF: I will first derive $V_{i,t}^{BAU}$. When we substitute in for $u_{i,t}$, $q_{i,t}^{BAU}$, and $r_{i,t}^{BAU}$ into $U_{i,t}^{BAU} = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_{i,\tau}$, we can rewrite $V_{i,t}^{BAU}$ as:

$$\begin{aligned} V_{i,t}^{BAU} &= \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[a \sum_{j \in N} \left(R_{j,\tau} + \frac{a}{b} \right) - \frac{b}{2} \left(\frac{a}{b} \right)^2 - \frac{c}{2} \left(\frac{\delta}{1-\delta} \frac{a}{c} \right)^2 \right] \\ &= \frac{a}{1-\delta} \sum_{j \in N} R_{j,t} + a \sum_{j \in N} \sum_{\tau=t}^{\infty} \frac{\delta^{\tau+1-t}}{1-\delta} y_{j,\tau} + \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[\left(n - \frac{1}{2} \right) \frac{a^2}{b} - \frac{c}{2} \left(\frac{\delta}{1-\delta} \frac{a}{c} \right)^2 \right] \\ &= \frac{a}{1-\delta} \sum_{j \in N} R_{j,t} + \frac{a^2}{1-\delta} \left(n - \frac{1}{2} \right) \left(\frac{1}{b} + \frac{1}{c} \left[\frac{\delta}{1-\delta} \right]^2 \right). \end{aligned}$$

Similarly, the BAU payoff at time $T+1$ can be written as:

$$V_{i,T+1}^{BAU} = \frac{a}{1-\delta} \sum_{j \in N} \left(R_{j,T+1}^{BAU} + Y_{j,T+1} \right) + \frac{a^2}{1-\delta} \left(n - \frac{1}{2} \right) \left(\frac{1}{b} + \frac{1}{c} \left[\frac{\delta}{1-\delta} \right]^2 \right),$$

where $Y_{i,T+1}$ measures the additional investments, relative to BAU, thanks to the first commitment period. Each party's present-discounted value of $Y_{i,T+1}$ is $a \frac{\delta^T}{1-\delta} \sum_j Y_{i,T+1}$, when evaluated in period 1. This term should be added when we derive the additional utility, relative to BAU, when the n parties commit to $\mathbf{x}$ for T periods at time $t = 1$ (even if the parties thereafter returned to BAU). The additional utility, relative to BAU, is thus:

$$\begin{aligned} &\sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \in N} \left(q_{j,t}^{BAU} + x_j \right) - \frac{b}{2} \left(q_{i,t}^{BAU} + x_i - R_{i,t}^{BAU} - Y_{i,t} \right)^2 - \frac{c}{2} \left(r_{i,t}^{B} + y_{i,t} \right)^2 - u_{i,t}^{BAU} \right] \\ &+ a \frac{\delta^T}{1-\delta} \sum_{j \in N} Y_{j,T+1} \\ = &\sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2} (x_i - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 + a Y_{i,t} - a\delta \frac{Y_{i,t+1} - Y_{i,t}}{1-\delta} \right] + a\delta^T \frac{\sum_{j \in N} Y_{j,T+1}}{1-\delta} \qquad (36) \\ = &\sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2} (x_i - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1}, \end{aligned}$$

where the last equality follows because the three terms with $Y_{i,\tau}$ in (36) sum to zero for each $\tau = \{2, ..., T+1\}$ and because $Y_{i,1} = 0$.

When the parties to *not* play BAU after the first commitment period, then, in order to obtain i's total additional payoff relative to BAU, we must add the additional payoff $\delta^T U_i(\mathbf{x}^*)$, where $U_i(\mathbf{x}^*)$ is the equilibrium additional utility relative to BAU, in order to get $U_i(\mathbf{x})$ in Lemma A-1. *Q.E.D.*

PROOF OF LEMMA 1: Lemma A-1 defines an optimal-control problem with control $y_{i,t}$. Note that the terminal value for $Y_{i,T+1}$ is zero because $U_i(\mathbf{x})$ is measured relative to $V_{i,1}^{BAU}$: this implies that $y_{i,T} = 0$, i.e., the investment level in the final period coincides with the equilibrium investment level in BAU. In other words, there is no *additional* investment in the final period.

When λ_t defines the shadow value of the stock $Y_{i,t}$, evaluated at time 1, the discrete-time Hamiltonian can be written as:[38]

$$H_t = \delta^{t-1}\left[a\sum_{j\neq i} x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2}y_{i,t}^2\right] + \lambda_{i,t+1}(Y_{i,t} + y_{i,t}),$$

with first-order conditions

$$y_{i,t} = \arg\max_{y_{i,t}} H_t = \lambda_{i,t+1}/c\delta^{t-1},$$

adjoint equation

$$\lambda_{i,t+1} - \lambda_{i,t} = -\frac{\partial H_t}{Y_{i,t}} = -\delta^{t-1} b(x_i - Y_{i,t}),$$

and terminal condition

$$\lambda_{i,T+1} = 0 \Leftrightarrow y_{i,T} = 0.$$

Combining the first two conditions and (8), we get the second-order difference equation:

$$\begin{aligned} c\delta^{t-2}(Y_{i,t} - Y_{i,t-1}) - c\delta^{t-1}(Y_{i,t+1} - Y_{i,t}) &= \delta^{t-1}(x_i - Y_{i,t})\, b \Rightarrow \\ -Y_{i,t+1} + \left(\frac{1}{\delta} + 1 + \frac{b}{c}\right) Y_{i,t} - \frac{1}{\delta} Y_{i,t-1} &= x_i b/c, \end{aligned}$$

which has the solution (see, e.g., Sydsaeter and Hammond (1995:751-53)):

$$\begin{aligned} Y_{i,t} &= A_1 m_1^{t-1} + A_2 m_2^{t-1} + x_i, \text{ where} \\ m_1 &= \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right) - \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right)^2 - \frac{4}{\delta}} \in (0,1), \\ m_2 &= \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right) + \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right)^2 - \frac{4}{\delta}} > 1. \end{aligned} \tag{37}$$

The constants A_1 and A_2 can be derived from the initial condition $Y_{i,1} = 0$, implying $A_1 + A_2 = -x_i$, and the terminal condition, $y_{i,T} = 0$, implying

$$\begin{aligned} y_{i,T} &= Y_{i,T+1} - Y_{i,T} = A_1 m_1^T\left(1 - \frac{1}{m_1}\right) - (A_1 + x_i) m_2^T\left(1 - \frac{1}{m_2}\right) = 0 \Rightarrow \\ A_1 &= -\frac{m_2^T\left(1 - \frac{1}{m_2}\right)}{m_1^T\left(\frac{1}{m_1} - 1\right) + m_2^T\left(1 - \frac{1}{m_2}\right)} x_i, \text{ and} \\ A_2 &= -A_1 - x_i = \frac{m_2^T\left(1 - \frac{1}{m_2}\right)}{m_1^T\left(\frac{1}{m_1} - 1\right) + m_2^T\left(1 - \frac{1}{m_2}\right)} x_i - x_i = -\frac{m_1^T\left(\frac{1}{m_1} - 1\right)}{m_1^T\left(\frac{1}{m_1} - 1\right) + m_2^T\left(1 - \frac{1}{m_2}\right)} x_i. \end{aligned}$$

[38] I here apply Pontryagin's maximum principle for discrete time problems. For a general characterization and proof, see, for example, Leonard and van Long (1992:129-33).

With the definitions $k_1 = -A_1 x_i$ and $k_2 = -A_2 x_i$, (37) can be written as in Lemma 1. *Q.E.D.*

PROOF OF LEMMA 2: By substituting in for $y_{i,t}$ and $Y_{i,t}$ into $U_{i,1}(\mathbf{x})$, defined in Lemma A-1, we get:

$$\begin{aligned}
U_i(\mathbf{x}) - \delta^T U_i(\mathbf{x}^*) &= \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2} (x_i - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1} \\
&= \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2} x_i^2 \left(k_1 m_1^{t-1} + k_2 m_2^{t-1} \right)^2 - \frac{c}{2} \left[x_i \left(k_1 m_1^{t-1} [1 - m_1] - k_2 m_2^{t-1} [m_2 - 1] \right) \right]^2 \right] \\
&\quad + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1} \\
&= \alpha \sum_{j \neq i} x_j + \beta x_i^2 / 2, \text{ if just} \\
&\quad \alpha \equiv \sum_{t=1}^{T} \delta^{t-1} a + a \frac{\delta^T}{1-\delta} \frac{Y_{j,T+1}}{x_j} = \frac{a}{1-\delta} \left[1 - \delta^T \left(k_1 m_1^{T-1} + k_2 m_2^{T-1} \right) \right] \text{ and} \\
&\quad \beta \equiv \sum_{t=1}^{T} \delta^{t-1} \left[b \left(k_1 m_1^{t-1} + k_2 m_2^{t-1} \right)^2 + c \left[\left(k_1 m_1^{t-1} [1 - m_1] - k_2 m_2^{t-1} [m_2 - 1] \right) \right]^2 \right]. \quad \textit{Q.E.D.}
\end{aligned}$$

PROOF OF PROPOSITIONS 1-3: The proof of Proposition 1 follows from the earlier Lemmata, while Propositions 2 and 3 follow from the reasoning in the text.

PROOF OF PROPOSITION 4: If i defects by not contributing at time t, then i can still benefit $a \sum_{j \neq i} x_j + \frac{a\delta}{1-\delta} \sum_{j \neq i} y_{j,t}$, since j's investments will raise j's contribution in the future, even when the parties return to BAU. This benefit is largest at $t = 1$, since $y_{j,t}$ is decreasing in $t \in \{1, ..., T\}$, as noticed already.

When defection is punished by a reversion to BAU for $l \leq \infty$ periods with probability $\phi \in (0, 1]$, then compliance (giving payoff U_i^*) is better at time $t = 1$ if:

$$a \sum_{j \neq i} x_j + \frac{a\delta}{1-\delta} \sum_{j \neq i} y_{j,1} + \delta \left(1 - \phi + \phi \delta^l \right) U_i^* \leq U_i^*.$$

When we substitute in for $y_{j,1}$, x_j^*, and U_i^*, this inequality becomes:

$$\begin{aligned}
a \left(\sum_{j \neq i} x_j^* + \frac{\delta}{1-\delta} \sum_{j \neq i} y_{j,1}^* \right) &\leq \left[1 - \delta \left(1 - \phi + \phi \delta^l \right) \right] U_i^* \iff \\
a \left[1 + \frac{\delta}{1-\delta} (1 - k_1 m_1 - k_2 m_2) \right] \sum_{j \neq i} x_j^* &\leq \left[1 - \delta \left(1 - \phi + \phi \delta^l \right) \right] \frac{\alpha^2 (n-1)^2}{\beta \left(1 - \delta^T \right)} w \left(1 - \frac{w}{2} \right) \iff \\
\frac{a \left(1 - \delta^T \right)}{\alpha \left[1 - \delta \left(1 - \phi + \phi \delta^l \right) \right]} \left[\frac{1 - \delta (k_1 m_1 + k_2 m_2)}{1 - \delta} \right] &\leq 1 - \frac{w}{2} \iff \\
w &\leq 2 - 2 \frac{1 - \delta (k_1 m_1 + k_2 m_2)}{(1-\delta) \left[1 - \delta \left(1 - \phi + \phi \delta^l \right) \right]} \frac{a \left(1 - \delta^T \right)}{\alpha},
\end{aligned}$$

which equals (16) when $\phi = 1$ and $l \to \infty$. *Q.E.D.*

PROOF OF PROPOSITION 5: Proposition 5 follows straightforwardly from the equilibrium continuation values, derived above.

PROOF OF PROPOSITION 6:

(i) Contracts on investments: I will first permit the negotiated $\mathbf{y}_t = (y_{1,t}, ..., y_{n,t})$ to be time-dependent, so that $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_T)$ is a matrix. Lemma 1 presents a reformulation of the problem and (when we remove the max-operator) it holds regardless of how the $x_{i,t}$'s and the $y_{i,t}$'s are decided on. When $y_{i,t}$ is committed to, but not $x_{i,t}$, the latter follows straightforwardly from i's maximization problem and, just as in BAU,

$$q_{i,t} - R_{i,t} = a/b \Rightarrow x_i = Y_{i,t}.$$

The continuation value can thus be written as a function of the investments matrix $\mathbf{y}$:

$$\begin{aligned} U_i(\mathbf{y}) &= \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} \sum_{t'=1}^{t-1} y_{j,t'} - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} \sum_{t'=1}^{T} y_{j,t'} + \delta^T U_i(\mathbf{y}^*) \Longleftrightarrow \\ U_i(\mathbf{y}) - \delta^T U_i(\mathbf{y}^*) &= \sum_{t=1}^{T} \left[\alpha_t \sum_{j \neq i} y_{j,t} - \frac{\beta_t}{2} y_{i,t}^2 \right], \text{ where } \alpha_t = \frac{a\delta^t}{1-\delta} \text{ and } \beta_t = \delta^{t-1} c. \end{aligned}$$

If we require a time-independent $y_{j,t} = y_j$, we can write

$$U_i(\mathbf{y}) - \delta^T U_i(\mathbf{y}^*) = \alpha \sum_{j \neq i} y_j - \frac{\beta}{2} y_i^2, \text{ where } \alpha = \delta a \frac{1-\delta^T}{(1-\delta)^2} \text{ and } \beta = \sum_{t=1}^{T} \delta^{t-1} c = c \frac{1-\delta^T}{1-\delta}.$$

Just as before, we can write i's payoff as in Example E. Consequently, the proofs for the other propositions follow the same steps as above. Proposition 1 gives, for example:

$$y_j^* = w(n-1)\alpha/\beta = w(n-1)\frac{\delta a/c}{1-\delta}.$$

Time-dependent investment levels: Since i's payoff is separable in the $y_{j,t}$'s, we can apply Corollary 1 for each $\mathbf{y}_t$, if we fix the investment levels for the other periods, in order to get:

$$y_{j,t}^* = w(n-1)\alpha_t/\beta_t = w(n-1)\frac{\delta a/c}{1-\delta},$$

which equals y_j^*. Hence, the restriction to time-independent investment levels is nonbinding: the equilibrium is the same in both cases.

The choice of T is also irrelevant in both cases, since the equilibrium continuation value is:

$$U_i(\mathbf{y}^*) = \delta a \frac{1}{(1-\delta)^2} (n-1)^2 w \frac{\delta a/c}{1-\delta} - \frac{c/2}{1-\delta} \left[(n-1) w \frac{\delta a/c}{1-\delta} \right]^2 = \frac{[\delta a (n-1)]^2}{c(1-\delta)^3} w \left(1 - \frac{w}{2}\right).$$

(ii) Contracts on carbon tax: I will first permit $z_t = (z_{1,t}, ..., z_{n,t})$ to be time-dependent, so that $z = (\mathbf{z}_1, ..., \mathbf{z}_T)$ is a matrix.

With an emission tax equal to $z_{i,t}$, collected by the government in country i, the equilibrium ensures that the marginal benefit when consuming fossil fuel (or the marginal abatement cost) equals $z_{i,t}$. This implies:

$$x_{i,t} - Y_{i,t} = z_{i,t}/b,$$

and, therefore, i's continuation value can be written as the function

$$U_i(\mathbf{z}) = \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} (z_{j,t}/b + Y_{j,t}) - \frac{z_{i,t}^2}{2b} - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1} + \delta^T U_i(\mathbf{z}^*)$$

so, there is no value for i to invest beyond the BAU-levels, and $y_{i,t} = 0$, so:

$$\begin{aligned} U_i(\mathbf{z}) - \delta^T U_i(\mathbf{z}^*) &\equiv \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} z_{j,t}/b - \frac{z_{i,t}^2}{2b} \right] = \sum_{t=1}^{T} \left[\alpha_t \sum_{j \neq i} z_{j,t} - \frac{\beta_t}{2} z_{i,t}^2 \right], \text{ where} \\ \alpha_t &= a\delta^{t-1}/b \text{ and } \beta_t = \delta^{t-1}/b. \end{aligned}$$

If the emission tax is time-independent, we can write:

$$U_i(\mathbf{z}) - \delta^T U_i(\mathbf{z}^*) = \alpha \sum_{j \neq i} z_j - \frac{\beta}{2} z_i^2, \text{ where } \alpha = \frac{a}{b}\frac{1-\delta^T}{1-\delta} \text{ and } \beta = \frac{1}{b}\frac{1-\delta^T}{1-\delta}.$$

In this case, Corollary 1 implies:

$$z_i^* = w(n-1)\alpha/\beta = w(n-1)a.$$

Time-dependent tax: Since i's payoff is separable in the $z_{j,t}$'s, we can apply Corollary 1 for each z_t, if we fix the emission tax levels for the other periods, in order to get:

$$z_{j,t}^* = w(n-1)\alpha_t/\beta_t = w(n-1)a,$$

which equals z_i^*. Hence, the restriction to time-independent emission tax levels is nonbinding: the equilibrium is the same in both cases.

The choice of T is also irrelevant in both cases, since the equilibrium continuation value is:

$$U_i(\mathbf{z}^*) = \frac{a}{b}\frac{1}{1-\delta}(n-1)^2 wa - \frac{1}{2}\frac{1}{b}\frac{1}{1-\delta}[(n-1)wa]^2 = \frac{[a(n-1)]^2}{b(1-\delta)} w\left(1-\frac{w}{2}\right).$$

By comparison, a tax gives higher payoff than an investment agreements if:

$$\frac{[a(n-1)]^2}{b(1-\delta)} > \frac{[\delta a(n-1)]^2}{c(1-\delta)^3} \Longleftrightarrow c(1-\delta)^2 > b\delta^2 \Longleftrightarrow \frac{1}{\delta} > 1 + \sqrt{\frac{b}{c}}.$$

Clearly, the investment agreement is better if investments are inexpensive and the tax ineffective (because b is large). If δ is large, investments are, in effect, less expensive, and thus the investment agreement is more attractive.

(iii) Combining (i) and (ii): When the parties face both a matrix of emission taxes and a matrix of investment levels, i's payoff can be written as:

$$U_i(\mathbf{x}) - \delta^T U_i(\mathbf{x}^*) \equiv \sum_{t=1}^{T} \delta^{t-1}\left[a\sum_{j\neq i}\left(\frac{z_{j,t}}{b} + Y_{j,t}\right) - \frac{z_{j,t}^2}{2b} - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i} Y_{j,T+1}$$

$$= \left[\sum_{t=1}^{T}\delta^{t-1}\left(a\sum_{j\neq i} Y_{j,t} - \frac{c}{2}y_{i,t}^2\right) + a\frac{\delta^T}{1-\delta}\sum_{j\neq i} Y_{j,T+1}\right] + \left[\sum_{t=1}^{T}\delta^{t-1}\left(a\sum_{j\neq i}\frac{z_{j,t}}{b} - \frac{z_{j,t}^2}{2b}\right)\right],$$

where the first (second) bracket can be recognized as i's payoff in the situation when only the investment levels (the emission taxes) were negotiated. The two problems are thus separable, and the results above continue to hold when the parties can negotiate both policy instruments. In this case, the additional payoff, relative to BAU, is also the sum of the two additional payoffs, derived above:

$$U_i(\mathbf{y}^*) + U_i(\mathbf{z}^*) = \left[\frac{1}{c(1/\delta - 1)^2} + \frac{1}{b}\right]\frac{[a(n-1)]^2}{(1-\delta)} w\left(1-\frac{w}{2}\right).$$

(iv) Complete contracts: When the parties negotiate the investment levels, the $z_{j,t}$'s pin down the $x_{j,t}$'s, given the $y_{j,t}$'s, so negotiating the $z_{j,t}$'s is then equivalent to negotiating the $x_{j,t}$'s: Also when the $y_{j,t}$'s and the $x_{j,t}$'s are negotiated, the contract is complete and the choice of T is irrelevant. One optimal T is thus $T = \infty$.

(v) Time-path for x: When the $y_{j,t}$'s and the $x_{j,t}$'s are negotiated, one optimal T is $T = \infty$. In this situation, pinning down the $x_{j,t}$'s is equivalent to pinning down both the $y_{j,t}$'s and the $x_{j,t}$'s, because there is no externality when it comes to the $y_{j,t}$'s (given every future $x_{j,t}$) and, hence, every party will invest optimally, without any need to specify the investment levels.

This reasoning completes the proof but, to illustrate, consider the time profile for the contribution levels when the parties negotiate both the emission taxes and the investment levels:

$$x_{i,t} = (n-1)\, wa/b + t\,(n-1)\, w\frac{\delta a/c}{1-\delta}.$$

Given this path, optimal investments, from i's point of view, are:

$$\begin{aligned} cy_{i,t-1} - \delta c y_{i,t} &= \delta b\,(x_{i,t} - Y_{i,t}) = \delta b\left((n-1)\, wa/b + t\,(n-1)\, w\frac{\delta a/c}{1-\delta} - t\,(n-1)\, w\frac{\delta a/c}{1-\delta}\right) \\ &= \delta b\,((n-1)\, wa/b) \Rightarrow y_{i,t-1} = \frac{\delta\,(n-1)\, wa}{c\,(1-\delta)}, \end{aligned}$$

just as in the optimal contract. So, the combination of negotiating investment levels and emission taxes is indeed equivalent to pinning down the path of $x_{i,t}$.

(vi) Firms: It suffices to prove that when $T = 1$ and the parties negotiate $x_{i,t}$ at the start of every period t, and the firms invest to maximize profit, then the outcome coincides with the outcome when all the $y_{i,t}$'s and the $x_{i,t}$'s are negotiated at the very beginning.

When only this period's $x_{i,t}$ are negotiated at the start of period t, then, when applying Corollary 1:

$$b\,(x_{i,t} - Y_{i,t}) = aw\,(n-1)\,.$$

Firms invest such as to equalize the marginal investment cost to the present-discounted value of their investment, where the willingness to pay for more $R_{i,t}$ equals $b\,(q_{i,t} - R_{i,t})$ at time t. Thus:

$$\begin{aligned} cr_{i,t} &= \sum_{t=1}^{\infty} \delta^t b\,(q_{i,t} - R_{i,t}) = \sum_{t=1}^{\infty} \delta^t b\left(q_{i,t}^{BAU} - R_{i,t}^{BAU} + x_{i,t} - Y_{i,t}\right) \\ &= \sum_{t=1}^{\infty} \delta^t b\left(\frac{a}{b} + x_{i,t} - Y_{i,t}\right) = \sum_{t=1}^{\infty} \delta^t b\left(\frac{a}{b} + \frac{a}{b} w\,(n-1)\right) = \frac{\delta}{1-\delta} b\left(\frac{a}{b} + \frac{a}{b} w\,(n-1)\right). \end{aligned}$$

With $r_{i,t} = r_{i,t}^{BAU} + y_{i,t}$ and $r_{i,t}^{BAU} = \frac{\delta}{1-\delta}\frac{a}{c}$, we get $cy_{i,t} = \frac{\delta}{1-\delta} aw\,(n-1)$, as with complete contracts.

(vii)-(ix) are trivial and thus omitted.

(x) Compliance: In all the above situations, and also in the basic model if $c \to \infty$, we have that U_i^* is independent of T and it can, when the policy instrument is given by the matrix $\psi = (\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_K)$, where $\psi_k = \left(\psi_{1,k}, ...\psi_{n,k}\right)$ for each $k \in \{1, ..., K\}$, be written as the following (for some constants α_k and β_k):

$$U_i^* = \sum_{k\in\{1,...,K\}} \frac{1}{1-\delta}\left[\alpha_k' \sum_{j\neq i} \psi_{j,k} - \frac{\beta_k'}{2}\psi_{j,k}^2\right], \text{ so } \psi_{j,k} = w\,(n-1)\,\alpha_k'/\beta_k',$$

from Corollary 1. If defection is punished by reverting to BAU for l periods with probability ϕ, then the incentive constraint is:

$$\sum_{k\in\{1,...,K\}} \alpha_k' \sum_{j\neq i} \psi_{j,k} + \delta\left(1 - \phi + \phi\delta^l\right) U_i^* \leq U_i^* \Longleftrightarrow$$

$$\sum_{k\in\{1,...,K\}} \frac{(n-1)^2\,(\alpha_k')^2}{\beta_k'} w \leq \sum_{k\in\{1,...,K\}} \frac{1-\delta\left(1-\phi+\phi\delta^l\right)}{1-\delta}\left[\frac{(n-1)^2\,(\alpha_k')^2}{\beta_k'} w - \frac{\beta_k'}{2}\left[\frac{(n-1)\,\alpha_k'}{\beta_k'} w\right]^2\right] \Longleftrightarrow$$

$$1 \leq \frac{1-\delta\left(1-\phi+\phi\delta^l\right)}{1-\delta}\left[1 - \frac{1}{2}w\right] \Longleftrightarrow$$

$$w \leq 2 - 2\frac{1-\delta}{1-\delta\left(1-\phi+\phi\delta^l\right)} = 2\frac{1-\delta\left(1-\phi+\phi\delta^l\right) - 1 + \delta}{1-\delta\left(1-\phi+\phi\delta^l\right)} = 2\delta\frac{\phi\left(1-\delta^l\right)}{1-\delta\left(1-\phi+\phi\delta^l\right)},$$

which simplifies to $w \leq 2\delta$ if $\phi = 1$ and $l = \infty$. *Q.E.D.*

REFERENCES FOR ONLINE APPENDIX

Leonard, D., and N. Van Long (1992): *Optimal Control Theory and Static Optimization in Economics*, Cambridge University Press.

Sydsaeter, K., and P. J. Hammond (1995): *Mathematics for Economic Analysis*, Prentice Hall.