Supplemental Information:

Designing the Optimal International Climate Agreement with Variability in Commitments

Jordan H. McAllister and Keith E. Schnakenberg

June 18, 2021

1 Full information benchmark

Proof of Lemma 1. The IO solves

$$\max_{(\tilde{c}(\theta_1),\ldots\tilde{c}(\theta_i),\ldots,\tilde{c}(\theta_n))} \sum_{i=1}^{n} \left[\theta_i \log(\tilde{c}(\theta_i)) + (1-\theta_i)\log(\omega - C)\right]. \tag{1}$$

Let $\tilde{C}(\theta) = \sum_{i=1}^{n} \tilde{c}(\theta_i)$ denote the total emissions under a mechanism $\tilde{c}$ and let $T = \sum_{i=1}^{n} \theta_i$ denote the sum of all types. This problem generates n first-order conditions:

$$\begin{aligned}
\frac{\theta_i}{\tilde{c}(\theta_i)} =& \frac{1-\theta_i}{\omega - \tilde{C}(\theta)} + \sum_{k=1}^{n} \frac{1-\theta_k}{\omega - \tilde{C}(\theta)} && (2)\\
\frac{\theta_i}{n-T} =& \frac{\tilde{c}(\theta_i)}{\omega - \tilde{C}(\theta)} && (3)\\
\frac{\theta_i}{n-T}\left[\omega - \tilde{C}(\theta)\right] =& \tilde{c}(\theta_i). && (4)
\end{aligned}$$

Solving for $\tilde{C}(\theta)$ gives:

$$\tilde{C}(\theta) =\sum_{i=1}^{n} \frac{\theta_i}{n-T}\left[\omega - \tilde{C}(\theta)\right] \tag{5}$$
$$=\frac{T}{n-\overline{\theta}}\left[\omega - \tilde{C}(\theta)\right] \tag{6}$$
$$\tilde{C}(\theta) =\frac{T\omega}{n}. \tag{7}$$

Plugging this solution into (4) gives $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$ as claimed. □

2 Proof of Proposition 1

To prove this result, we follow the general steps in **?**. First, we consider the IO's "relaxed problem" of maximizing welfare subject only to (**??**). We show that, in a solution to this problem, no type is allowed emissions more than $\frac{\overline{\theta}\omega}{n}$, the amount of emissions allowed for the highest type in the full information solution in Lemma 1. Second, we show that all types would choose emissions higher than $\frac{\overline{\theta}\omega}{n}$ if they freely chose their own emissions. This implies that the only solutions to this relaxed problem are fully compressed because all types for all countries i would be incentivized to misrepresent their true type by reporting $\hat{\theta} > \theta_i$.

Lemma 1. *Let c^0 be a solution to the IO's relaxed problem of maximizing* (**??**) *subject to* (**??**)*. Then $c^0(\hat{\theta}_i, \theta_{-i}) \leq \frac{\overline{\theta}\omega}{n}$ for all i with probability 1.*

Proof. Our proof follows the steps in Harrison and Lagunoff (2017). In line with their proof, we establish similar notation. We write the utility of a type θ_i of player i of a consumption plan c as

$$u_i(\theta_i, \theta_{-i}; c) = r(c) - \theta_i q_i(c)$$

where $r(c) = \log(\omega - C)$ and $q_i(c) = \log(\omega - C) - \log(c_i)$. This is simply a rewriting of the payoff where r represents the (common) rewards to conservation and q_i represents the individual costs. Consider a solution c^0 and assume that $c_i^0(\theta) > \overline{c} \equiv \frac{\overline{\theta}\omega}{n}$ for some individual i and type realizations θ. Denote the interim values of the cost and reward functions defined above by

$$\begin{aligned} R_i^0(\theta_i) &= \int_{\theta_{-i}} r(c^0(\theta)) dF_{-i}(\theta_{-i}) \text{ and} \\ Q_i^0 &= \int_{\theta_{-i}} q_i(c^0(\theta)) dF_{-i}(\theta_{-i}). \end{aligned}$$

Following Harrison and Lagunoff (2017) we can rewrite the relaxed problem as

$$\max_c \sum_i \left[R_i^0(\overline{\theta}) - \overline{\theta} Q_i(\overline{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i(\theta_i) d\theta_i \right] \tag{8}$$

subject to Q_i weakly decreasing. We construct an alternative consumption plan c^{**} as follows:

$$c_i^{**}(\theta) = \begin{cases} \overline{c} & \text{if } \theta_i = \overline{\theta} \\ c_i^0(\theta) & \text{otherwise.} \end{cases}$$

$$c_j^{**}(\theta) = \begin{cases} \overline{c} & \text{if } \theta_k = \overline{\theta} \text{ for any k} \\ c_i^0(\theta) & \text{otherwise.} \end{cases}$$

We can write the difference in the objective functions for these two consumption plans as

$$\begin{aligned} &\sum_i \left[R_i^{**}(\overline{\theta}) - \overline{\theta} Q_i^{**}(\overline{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i^{**}(\theta_i) d\theta_i \right] - \\ &\qquad \sum_i \left[R_i^0(\overline{\theta}) - \overline{\theta} Q_i^0(\overline{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i^0(\theta_i) d\theta_i \right] \qquad (9) \\ &= \sum_i \left[R_i^{**}(\overline{\theta}) - \overline{\theta} Q_i^{**}(\overline{\theta}) - \left(R_i^0(\overline{\theta}) - \overline{\theta} Q_i^0(\overline{\theta}) \right) \right] + \\ &\sum_i \left[\int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i^{**}(\theta_i) d\theta_i - \int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i^0(\theta_i) d\theta_i \right]. \qquad (10) \end{aligned}$$

The first sum in (10) must be positive because $\bar{c}$ is the unconstrained optimal consumption for a player of type $\bar{\theta}$ regardless of the type profile of the other players. The second sum is positive because $c^{**}(\theta) \leq c^0(\theta)$ and Q_i is monotone in c. Therefore, c^{**} has an overall higher value of the objective function in (8), contradicting the statement that c^0 is optimal. □

Lemma 2. *If $\underline{\theta} > \frac{\bar{\theta}}{n+\bar{\theta}-n\bar{\theta}}$ then for all types of all countries we have*

$$\arg\max_{c} \int_{\Theta_{-i}} \theta_i \log(c) + (1-\theta_i)\log(\omega - c - \sum_{j\neq i} c_j^0(\theta_{-i}))dF_{-i}(\theta_{-i}) > \frac{\bar{\theta}\omega}{n}.$$

That is, if each country could freely choose its consumption, it would prefer to choose a level greater than that allowed to the highest type in the unconstrained optimum.

Proof. Country i's first-order condition (using Leibniz's rule) is

$$\int_{\Theta_{-i}} \left[\frac{\theta_i}{c} - \frac{1-\theta_i}{\omega - c - \sum_{j\neq i} c_j^0(\theta_{-i})}\right] dF_{-i}(\theta_{-i}) = 0. \tag{11}$$

Let c^* denote a solution to this first-order condition. Note that country i's optimal choice is always decreasing in the total amount consumed by the other players. Therefore c^* must be greater than the amount i would consume if all other players consumed $\bar{c} = \frac{\bar{\theta}\omega}{n}$ (the maximum amount possible given Lemma 1) given any θ. We let $\hat{c}$ denote this amount and solve for it as follows:

$$\begin{aligned}
\int_{\Theta_{-i}} \left[\frac{\theta_i}{\hat{c}} - \frac{1-\theta_i}{\omega - \hat{c} - \frac{n-1}{n}\bar{\theta}\omega}\right] dF_{-i}(\theta_{-i}) =& 0 && (12)\\
\frac{\theta_i}{\hat{c}} =& \frac{1-\theta_i}{\omega - \hat{c} - \frac{n-1}{n}\bar{\theta}\omega} && (13)\\
\hat{c} =& \frac{\theta_i\omega}{n}(n+\bar{\theta}-n\bar{\theta}). && (14)
\end{aligned}$$

Note that this is increasing in θ_i. Using that our assumption that $\underline{\theta} > \frac{\bar{\theta}}{n+\bar{\theta}-n\bar{\theta}}$ we have

$$\begin{aligned}\hat{c} =& \frac{\theta_i \omega}{n}(n + \overline{\theta} - n\overline{\theta}) & (15)\\ \geq& \frac{\bar{\theta}}{n + \bar{\theta} - n\bar{\theta}} \frac{\omega}{n}(n(1 + \overline{\theta}) - \overline{\theta}) & (16)\\ =& \frac{\overline{\theta}\omega}{n}. & (17)\end{aligned}$$

Thus, we have $c^* > \hat{c} > \frac{\overline{\theta}\omega}{n}$ as claimed. □

We are now ready to prove the main result.

Proof of Proposition 1. Suppose c^0 is a solution to the planner's problem and is not fully compressed. Then for some player i and type θ_i we have $c_i^0(\theta_i, \theta_{-i}) < c_i^0(\theta', \theta_{-i})$ for $\theta' \neq \theta_i$. By Lemma 1, we have $c_i^0(\theta_i, \theta_{-i}) < c_i^0(\theta', \theta_{-i}) < \frac{\overline{\theta}\omega}{n}$. But the interim expected utility for type θ_i of player i is strictly concave and, by Lemma 2, maximized at some value $c > \frac{\overline{\theta}\omega}{n}$. This implies that type θ_i of player i prefers any consumption level in $(c_i^0(\theta_i, \theta_{-i}), \frac{\overline{\theta}\omega}{n}]$ to $c_i^0(\theta_i, \theta_{-i})$. In particular, type θ_i of player i strictly prefers $c_i^0(\theta', \theta_{-i})$ to $c_i^0(\theta_i, \theta_{-i})$, contradicting the truth-telling constraint. Thus, any solution to the planner's problem must be fully compressed. □

3 Limited investigations

Proof of Proposition 3. Recall that the optimal quota is $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$. Type interim expected utility of type θ_i of player i for participating in an agreement with the optimal quota given θ_{-i} is therefore

$$\theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1 - \theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right). \tag{18}$$

The utility to type θ_i of player i if i does not participate is

$$\max_{c \geq 0} \theta_i \log(c) + (1-\theta_i) \log\left(\omega - c - \frac{\omega}{n}\sum_{j \neq i} \theta_j\right) - K. \tag{19}$$

Taking first-order conditions and solving for the optimal c at each $\theta_{-i} \in \Theta_{-i}$ gives:

$$\begin{aligned} \frac{\theta_i}{c} &= \frac{1-\theta_i}{\omega - c - \frac{\omega}{n}\sum_{j \neq i} \theta_j} && (20) \\ \Rightarrow c &= \theta_i \left(\omega - \frac{\omega}{n}\sum_{j \neq i} \theta_j\right). && (21) \end{aligned}$$

Define

$$U_i^O(\theta_i) := \theta_i \log\left(\theta_i\left(\omega - \frac{\omega}{n}\sum_{j \neq i}\theta_j\right)\right) + (1-\theta_i)\log\left(\omega - \theta_i\left(\omega - \frac{\omega}{n}\sum_{j \neq i}\theta_j\right) - \frac{\omega}{n}\sum_{j \neq i}\theta_j\right). \tag{22}$$

The utility to type θ_i of player i for opting out of the agreement is therefore

$$U_i^O(\theta_i) - K. \tag{23}$$

Clearly, for $K > U_i^O(\theta_i) - \theta_i \log\left(\frac{\omega\theta_i}{n}\right) - (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j \neq i}\theta_j\right)$, the participation constraint is met.

Next, consider the payoff to type θ_i of player i to submitting a report $\hat{\theta} \neq \theta_i$ given the investigation mechanism $r_i(\hat{\theta}_i, \hat{\theta}_{-i})$. In this case, player i gets its reservation payoff from (23) with probability $r_i(\hat{\theta}_i, \hat{\theta}_{-i})$ and, with probability $1 - r_i(\hat{\theta}_i, \hat{\theta}_{-i})$, gets its payoff from successfully imitating type $\hat{\theta}$ in the optimal mechanism. This expected payoff is

$$\int_{\Theta_{-i}} \Bigg[(1 - r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] + \\ r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[U_i^O(\theta_i) - K\right] \Bigg] dF_{-i}(\theta_{-i}). \tag{24}$$

The truth-telling constraint is

$$\int_{\Theta_{-i}} \left[\theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] dF_{-i}(\theta_{-i}) \geq \\ \int_{\Theta_{-i}} \Bigg[(1 - r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] + \\ r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[U_i^O(\theta_i) - K\right] \Bigg] dF_{-i}(\theta_{-i}). \tag{25}$$

A sufficient condition is to satisfy this constraint for every θ_{-i}, so we can write the constraint as

$$\theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right) - \\ (1 - r_i(\hat{\theta}, \hat{\theta}_{-i})) \left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] - \\ r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[U_i^O(\theta_i) - K\right] \geq 0. \tag{26}$$

To save on notation, let

$$\begin{aligned} \tilde{U}_i(\theta_i) =& \theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right) \\ \tilde{U}_i(\hat{\theta}) =& \theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right). \end{aligned}$$

For $\hat{\theta} > \theta_i$, (26) holds with equality if

$$r_i(\hat{\theta}, \hat{\theta}_{-i}) = \frac{\tilde{U}_i(\hat{\theta}) - \tilde{U}_i(\theta_i)}{K - U_i^O(\theta_i) + \left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right]}. \qquad (27)$$

By Lemma 2, we have

$$\left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] -$$
$$\left[\theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right] >0$$

for $\hat{\theta} > \theta_i > \frac{\overline{\theta}}{n(1+\overline{\theta})+\overline{\theta}}$. Thus, the numerator of (27) is positive and the denominator goes to ∞ as $K \to \infty$.

Finally, define $r^*(K)$ as follows:

$$r^*(K) = \sup_{\theta_i\in\Theta_i}\sup_{\theta_{-i}\in\Theta_{-i}}\sup_{\hat{\theta}\in\Theta_i} \frac{\tilde{U}_i(\hat{\theta}) - \tilde{U}_i(\theta_i)}{K - U_i^O(\theta_i) + \left[\theta_i \log\left(\frac{\omega\hat{\theta}}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\hat{\theta}}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right]}$$

denote the smallest amount of investigative resources that deters all types from submitting a false report given any distribution of other players' types as long as the participation constraint is met. For a given K large enough to satisfy the pa44rticipation constraint, the constant investigation plan setting $r_i(\hat{\theta}_i\hat{\theta}_{-i}) = r^*(K)$ for all $\hat{\theta}$ and all i implements the full information optimum. The total investigative budget is therefore $nr*(K)$. Since $\lim_{K\to\infty} nr^*(K) = n\lim_{K\to\infty} r^*(K) = 0$, this shows that for any $R > 0$ there exists a value of K large enough to implement the full information optimal quota. □

4 Extensions and additional results

4.1 An example of a non-inclusive optimal compressed mechanism

In the main text, we impose the constraint that all countries must participate and show that the optimal mechanism subject to that constraint is fully compressed. Here, we show by example that the optimal mechanism may not be fully inclusive. That is, the optimal mechanism may have some set of types opt out. By way of example, let $\omega = 2$, $n = 3$, and assume that the set of types is $[1/5, 4/5]$. We will consider the limiting case as $K \to 0$. Recall from our arguments in the main text that the optimal fully inclusive mechanism is

$$c^* = \frac{\overline{\theta}\omega}{1 + \overline{\theta}(n-1)}. \tag{28}$$

The optimal fully inclusive mechanism from applying (28) is therefore a compressed mechanism in which $c^* = 8/13$. Our task is to show that this mechanism is not always ex-post optimal among all incentive compatible mechanisms. To show this, let's consider a realization of types such that $\theta_1 = \theta_2 = \frac{1}{5}$ and $\theta_3 = \frac{3}{5}$. Total utility under the optimal mechanism is

$$2\left[\frac{1}{5}\log\left(\frac{8}{13}\right) + \frac{4}{5}\log\left(\frac{2}{13}\right)\right] + \frac{4}{5}\log\left(\frac{8}{13}\right) + \frac{1}{5}\log\left(\frac{2}{13}\right) \approx -3.95. \tag{29}$$

Consider an alternative mechanism for which only the lowest types would participate. That is, the mechanism is a fully compressed quota such that the lowest types are indifferent between participating and not participating and all other types opt-out. The quota that makes the lowest types indifferent (as $K \to 0$) is $c'' = \frac{2}{7}$ (again applying (28) but using the lowest type rather than the highest type). All higher types opt out of this mechanism and choose $c_i = \theta_i(\omega - \sum_{j \neq i} c_j)$. In this case, with the type realizations above, the ex-post payoff from this mechanism is therefore

$$2\left[\frac{1}{5}\log\left(\frac{2}{7}\right)+\frac{4}{5}\log\left(\frac{2}{7}\right)\right]+\frac{4}{5}\log\left(\frac{8}{7}\right)+\frac{1}{5}\log\left(\frac{2}{7}\right)\approx -2.65. \tag{30}$$

The additional calculations here are that $\frac{4}{5}(2-2\frac{2}{7})=\frac{8}{7}$ and $2-2\frac{2}{7}-\frac{8}{7}=\frac{2}{7}$. That is, a non-inclusive compressed mechanism gives higher total welfare than the inclusive compressed mechanism.

This example may seem extreme because it relies on (a) a very strange realization of types at the endpoints of the support of the distribution and (b) a very extreme mechanism that only includes a measure zero set of types (i.e., only the very lowest type opts into the mechanism). However, this is only for convenience. Consider for instance a mechanism designed to make all types lower than $\frac{1}{2}$ accept. Applying (28) using $\theta=\frac{1}{2}$ rather than the highest type yields $c^*=\frac{1}{2}$. Under this mechanism (but under the same type realizations), the low types agree and choose $\frac{1}{2}$ while the high type opts out and chooses $c_i=\frac{4}{5}(2-2*\frac{1}{2})=\frac{4}{5}$. Therefore, the total utility is

$$2\left[\frac{1}{5}\log\left(\frac{1}{2}\right)+\frac{4}{5}\log\left(\frac{1}{5}\right)\right]+\frac{4}{5}\log\left(\frac{4}{5}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)\approx -3.35. \tag{31}$$

Thus, this less extreme mechanism still outperforms the fully inclusive mechanism. Furthermore, because both expected welfare equations are continuous with respect to the type realizations except near the participation cutoff, this mechanism beats the fully inclusive mechanism for a positive measure of types.

This construction does depend on having a small value of K. In fact, for large values of K it must be the case that the optimal compressed mechanism is fully inclusive, because all of the participation constraints are easily satisfied. This explains why there is no difference between the optimal mechanism and the optimal fully inclusive mechanism in **?**, where dynamic strategies resemble an unbounded value of K.

4.2 Endogenous K

The model in the text assumes an exogenous punishment K for noncompliance. Of course, international organizations do not typically have significant enforcement power, so here we consider how K may arise endogenously without the need for intrinsically powerful organizations. We consider two possibilities.

One way to think about K is as a short form for forward-looking strategies under dynamic mechanisms. We do not fully analyze this case here but instead point out the relationship to **?**. That paper considers a dynamic model with the same basic elements that we present here but with a carbon stock that evolves over time as a function of the choices of the players. In the dynamic model, shirking is responded to with punishments that deplete the total carbon stock in future periods. Since these potential punishments are unbounded, the results of the dynamic model most closely resemble our limiting case of $K \to \infty$. In fact, under complete information any feasible payoff is implementable. In the case of complete information, the optimal compressed mechanism is implementable and in fact is very similar to our compressed mechanism in the limiting case of $K \to \infty$. Thus, if we interpret our model as a simplification of a richer dynamic model, we should take the punishments to be arbitrarily large. Note also that these punishments are self-enforcing and do not require state-like powers for the IO.

An alternative way to think of K is as something arising from other issues that states care about and that may be affected by climate negotiations. A benefit of this perspective is that we can still think about the effects of varying K, but this becomes a measure of the value of the non-environmental issue. Therefore, we consider an extension in which the punishment for noncompliance (or reward for compliance) is determined by issue bundling and is self-enforcing and requires no formal enforcement power by the IO. That is, the role of the IO is merely to help countries coordinate on a particular self-enforcing punishment strategy. Though climate action could in principle be bundled with any issue, we use trade as our leading example. This allows us to focus on a concrete and simple model and also places our model firmly in line with the climate club proposals of **?** and others (**??**). Following **?**, we take the repeated prisoner's dilemma (PD) as the useful metaphor for trade cooperation.

Consider the following game. There is an infinite horizon with time indexed by $t \in \{0, 1, 2, \dots\}$. At time $t = 0$, the countries play the climate game exactly as described in the main text of the paper but without the exogenous penalty for noncooperation. In each period $t \geq 1$, the countries engage in a trade game which we represent as follows. In each period, every country sets trade barriers with every other country. Following **?**, we can represent the trade barrier game between each pair of countries as a prisoner's dilemma.[1] Thus, for each pair of countries $i, j \in N$, the stage game payoffs can be represented by the following:

		Country j	
		C	D
Country i	C	(b, b)	$(-d, a)$
	D	$(a, -d)$	$(0, 0)$

where $a > b > 0$ and $d > 0$. Since these payoffs apply to a particular pair, each country's total stage game payoff is its payoff from each of these interactions summed over all players. Formally, let $s_{ijt} \in \{C, D\}$ denote player i's action toward player j in time t and let s_t denote the profile of all actions in time t. Let $v_{ijt}(s_{ijt}, s_{jit})$ be the pairwise stage game payoff defined in the table above (i.e., $v_{ijt}(C, C) = b, v_{ijt}(D, C) = a, v_{ijt}(D, D) = 0, v_{ijt}(C, D) = -d$). Country i's stage game payoff is then $v_{it}(s_t) = \sum_{j \neq i} v_{ijt}(s_{ijt}, s_{jit})$. The discounted present value of a stream of payoffs is therefore $\sum_{t=1}^{\infty} \delta^{t-1} v_{it}(s_t)$ where $\delta \in (0, 1)$ is a common discount factor.

There are a great many equilibria to this game. Our purpose is merely to show that (a)

[1] The prisoner's dilemma setup is derived from a continuous game in Rosendorff and Milner (2001) by setting the payoff when both countries defect equal to the countries' payoffs from the static Nash equilibrium to the game, the payoff when both players cooperate equal to the payoffs from the Pareto optimal pair of trade barriers, and the payoff when one player defects and the other cooperates equal to the payoff obtained when one player acts according to the Pareto optimal trade barriers and the other chooses the optimal defection given that player's action.

punishing climate defectors can be self-enforcing in the trade stage of this game and (b) anticipation of the trade stage can endogenously generate the equivalent of our K without any formal enforcement power by the IO. We begin by analyzing the trade stage of the game. Suppose that a (possibly empty) subset of countries $M \subseteq N$ are considered non-compliant with the climate agreement and the rest are compliant. We set up a strategy profile in which compliant countries play grim trigger strategies with one another and always defect on countries in M, and where countries in M always defect regardless of which countries they are matched with. To be complete, the strategy is described as follows for all countries $i \in N$:

- If $i \in M$ then $s_{ijt} = D$ for all $j \in N \backslash \{i\}$ and all t.
- If $i \notin M$ then $s_{ijt} = D$ for all $j \in M$ and $s_{ij1} = C$ for all $j \notin M$, and for $t > 1$ $s_{ijt} = C$ if and only if $s_{ijt'} = s_{jit'} = C$ for all $t' < t$, otherwise $s_{ijt} = D$.

No $i \in M$ will deviate from this strategy: defection is a dominant strategy in the static game and cooperating in any period cannot change future payoffs, so defection remains a best response for $i \in M$. Similarly, because countries in M will never cooperate, no country $i \notin M$ will deviate from the part of its strategy that dictates $s_{ijt} = D$ for all $j \in M$. Finally, in pairs where neither country is in M, neither country has any incentive to deviate from the grim trigger punishment strategy for the same reasons as above. Thus, this strategy profile is an equilibrium as long as pairs of climate-compliant countries have enough incentive to cooperate along the path of play. Since actions are chosen independently across pairs (e.g., player i defecting on player j affects future rounds with player j but does not affect pairs involving player i but not player j), we can simply analyze the standard PD game within each pair. Player i's expected payoff in interactions with player j from cooperating is

$$b + \delta b + \delta^2 b + \cdots = \frac{b}{1-\delta}.$$

Player i's expected payoff in interactions with player j from defecting is

$$a + \delta 0 + \delta^2 0 + \cdots = a.$$

Therefore, player i cooperates along the path of play if $\frac{b}{1-\delta} \geq a$ which holds when $\delta \geq \frac{a-b}{a}$. Thus, this strategy profile is an equilibrium when $\delta \geq \frac{a-b}{a}$.

The above analysis makes a very simple point that is only slightly modified from standard analyses of the repeated PD. However, we now consider the implications for the climate agreement at time $t = 0$. We consider whether all countries will choose to participate in a particular agreement. Consider an agreement that prescribes a profile of carbon consumption levels $(\tilde{c}_1(\theta_1), \ldots, \tilde{c}_n(\theta_n))$. We consider a profile in which all countries comply and ask whether any particular country would unilaterally withdraw. We first consider the fully compressed mechanism, so we drop dependence of the quota on θ_i for the moment and ignore truth-telling constraints. The utility of a country of type θ for complying given the trade equilibrium above is:

$$u_i(\tilde{c}, \tilde{\mathbf{c}}_{-i}, \theta) + \delta \frac{b(n-1)}{1-\delta}.$$

That is, a country's expected utility for complying is now the static utility from carbon consumption plus the discounted trade benefits from being in compliance. The total trade benefits are the pairwise benefit from above multiplied by the number of other players, because the player contemplating a deviation expects all other players to remain in compliance.

The expected benefit of noncompliance is

$$\max_{c \geq 0} u_i(c, \tilde{\mathbf{c}}_{-i}, \theta).$$

That is, a country that does not comply gets the benefit of optimally choosing its carbon consumption given the levels of the other players, but it expects zero trade benefits in the future because every country will defect in every interaction. Therefore, the country complies if

$$u_i(\tilde{c}, \tilde{\mathbf{c}}_{-i}, \theta) + \delta\frac{b(n-1)}{1-\delta} \geq \max_{c\geq 0} u_i(c, \tilde{\mathbf{c}}_{-i}, \theta)$$

or

$$u_i(\tilde{c}, \tilde{\mathbf{c}}_{-i}, \theta) \geq \max_{c\geq 0} u_i(c, \tilde{\mathbf{c}}_{-i}, \theta) - \delta\frac{b(n-1)}{1-\delta}.$$

Setting $K = \delta\frac{b(n-1)}{1-\delta}$, this is exactly the participation constraint from (3). This analysis holds for the fully compressed mechanism but works the same for the optimal mechanism with investigations. If a country is labelled noncompliant when investigations reveal that it submitted a false report, we simply set $K = \delta\frac{b(n-1)}{1-\delta}$ to obtain the same truth-telling constraint as in (26).

This simple exercise is intended merely to demonstrate how we can think of K in the original model as coming from a mechanism that does not depend on having an IO with formal enforcement power. Taking the PD model a bit more seriously, however, we could characterize how K changes as a function of the underlying parameters. We see that K is larger when players are more patient, when the total number of countries negotiating increases, and when the value of preferential treatment on trade increases. As we mentioned, however, we think of trade as only one possible source of issue linkage. In addition, a more detailed model of the linked issue may yield further interesting insights about enforcement.