Renegotiation-proof climate agreements with full participation — conditions for Pareto-efficiency *

Geir B. Asheim†
Bjart Holtsmark‡

October 15, 2008

Abstract

Recent contributions show that climate agreements with broad participation can be implemented as weakly renegotiation-proof equilibria in simple models of greenhouse gas abatement where each country has a binary choice between cooperating (i.e., abate emissions) or defecting (no abatement). Here we show that this result carries over to a model where countries have a continuum of emission choices. Indeed, a Pareto-efficient climate agreement can always be implemented as a weakly renegotiation-proof equilibrium, for a sufficiently high discount factor. This means that one need not trade-off a "narrow but deep" treaty with a "broad but shallow" treaty.

1 Introduction

In simple dynamic models of international environmental public good provision, such as mitigation of climate change, Barrett (1999, 2002) has argued that there is a trade-off between "narrow but deep" and "broad but shallow" treaties: either only a few countries participate

*We thank Reyer Gerlagh, Michael Hoel, Jon Hovi, two anonymous referees, and participants at the 2007 Oslo Workshop on Environmental Economics and the 2008 EAERE Conference for many helpful comments. Holtsmark gratefully acknowledges financial support from the Research Council of Norway through RENERGI.

†Department of Economics, University of Oslo, P.O. Box 1095 Blindern, NO-0317 Oslo, Norway. *E-mail:* g.b.asheim@econ.uio.no

‡Statistics Norway, P.O. Box 8131 Dep., NO-0033 Oslo, Norway. *E-mail:* Bjart.Holtsmark@ssb.no

each with a large abatement, or many countries participate each with a small abatement.

By applying the Barrett (1999) model, where each country has a binary choice between *cooperating* (i.e., abate emissions) or *defecting* (no abatement), Asheim et al. (2006) show that extended participation is feasible. They show that participation can essentially be doubled in a two-region world.

The analysis of Asheim et al. (2006) exploits the fact that Barrett (1999) considers only strategy profiles with a special structure, namely where there is a subset of participating countries ("signatories") in a treaty, and where a defecting signatory is punished by having *all* other signatories defect in the next period (only). Then if there are too many signatories, these will gain by renegotiating back to cooperation without imposing the punishment, thereby undermining the credibility of the equilibrium. Asheim et al. (2006) limit the number of punishing countries by letting a defection be punished only by the other signatories in the same region, while the signatories in the other region continue to cooperate.

The possibility that only a subset of the signatories within a global treaty punishes a deviant is investigated to its logical conclusion by Froyn and Hovi (2008) within the binary choice model of Barrett (1999). They show that full participation can indeed by implemented as a weakly renegotiation-proof equilibrium.

It is still an open question whether these insights carry over to a continuum choice model like the one considered by Barrett (2002). To reach a Pareto-efficient agreement in such a setting, one need not only agree on a *broad* treaty with full participation, but also a *deep* treaty where each country's abatement is at an efficient level.

By considering a model where the public benefits of emission abatement are linear and private costs of emission abatement are quadratic, we show as our main result (Proposition 1) that such an efficient broad and deep treaty can always be implemented as a weakly renegotiation-proof equilibrium, provided that the discount rate is sufficiently low and the number of countries is sufficiently small. In the same context we also show as an additional result (Proposition 2) how depth, but not broadness, must be compromised for high discount rates and a large number of countries.

Since low time discounting and a short detection lag contribute to a

low discount rate, and high time discounting and a long detection lag contribute to a high discount rate,[1] these results mean that

- low time discounting and a short detection lag combined with a small number of countries contribute to the feasibility of a Pareto-efficient agreement, with full participation and efficient depth,
- high time discounting and a long detection lag combined with a large number of countries undermine the feasibility of Pareto-efficient depth of cooperation. However, such a shallower agreement still allows full participation.

Both results follow from a technical result (Theorem 1), in which we characterize the set of weakly renegotiation-proof equilibria for a class of repeated game strategy profiles, where punishments last for one period only (Abreu, 1986; van Damme, 1989), but where participation in the treaty, participation in the punishment, the depth of the treaty and the severity of the punishment are parameters which are allowed to vary.

Linear benefits of abatement represent a simplification. Asheim et al. (2006) show that their result holds also when abatement yields non-linear benefits. It would be of value to check whether our findings carry over to a less restrictive model with non-linear benefits of abatement and asymmetric countries, as considered in the two-country model in Finus and Rundshagen (1998). However, we remain within the context of linear benefits in the present paper, as it is an analytically tractable setting which allows us to characterize weak renegotation-proofness.

Sections 2 and 3 present our results, while Section 4 contains a discussion of their relevance. All proofs are relegated to an appendix.

2 Main result

Consider a world with $n \geq 2$ countries, where $N = \{1, \ldots, n\}$ denotes the set of all countries. The countries interact in periods (or "stages") $0, 1, 2, \ldots$. The countries are identical in all relevant characteristics. In every period, each country i must choose a non-negative level of abatement q_i of greenhouse gas emissions. Each country i's periodic payoff,

[1] If r is the positive rate of time discounting, and Δ is the detection lag (= period length), then the per-period discount factor, δ, is given by $\delta = \int_0^\Delta e^{-rt} dt$ and the per-period discount rate is $1 - \delta$.

relative to the situation where no country abates, is given by

$$\pi_i = b \sum_{j=1}^{n} q_j - \frac{c}{2}(q_i)^2 \,, \tag{1}$$

where b is the marginal benefit from abatement (which is a pure public good that benefits each and every country), and $(c/2)(q_i)^2$ represents the total abatement costs of country i. We assume that b, $c > 0$.

Following Barrett (1999, 2002) and Asheim et al. (2006), we abstract from the future benefits of abatement (which of course are important in the climate change setting; cf. Dutta and Radner, 2007), meaning that the situation can be modeled as an infinitely repeated game, with a stage game where the countries simultaneously and independently choose abatement levels, and receive payoffs according to (1).

The stage game has a unique Nash equilibrium where each country abates

$$q^1 = \frac{b}{c} \,.$$

Actually, for each country, q^1 strictly dominates any other action of the stage game and is thus its unique best response independently of what the other countries' abatement levels are. However, the unique symmetric Pareto-efficient abatement profile entails that each country abates

$$q^n = \frac{nb}{c} \,. \tag{2}$$

Hence, the Pareto-efficient abatement is n times the abatement level in the Nash equilibrium of the stage game, cf. Barrett (2002, p. 540). In particular, the n countries would want to agree on implementing

$$\mathbf{a} \;=\; (\underbrace{q^n, \ldots, q^n}_{j \in N}), (\underbrace{q^n, \ldots, q^n}_{j \in N}), \ldots \,,$$

where each country contributes to the Pareto-efficient total abatement in a cost-efficient manner in every period.

In the absence of third-party enforcement, such a Pareto-efficient agreement needs to be self-enforcing, where deviations from this agreement—leading to a short-run benefit for the deviating country—is deterred through the threat of future punishment, which must also be self-enforcing. Here, "self-enforcing" refers to the play of a non-cooperative equilibrium of the infinitely repeated game; the analysis of such equilibria requires the introduction of some game-theoretic formalism.

A history at the beginning of stage t describes the countries' abatement levels in periods $0, \ldots, t-1$:

$$(q_1(0), \ldots, q_n(0)), (q_1(1), \ldots, q_n(1)), \ldots, (q_1(t-1), \ldots, q_n(t-1)) .$$

A strategy σ_i for country i is a function which for every history, including the "empty" history at the beginning of stage 0, determines an abatement level for player i. Country i's average discounted payoff in the repeated game is given by

$$\big(1-\delta\big) \sum_{t=0}^{\infty} \delta^t \pi_i(t) , \tag{3}$$

where $\delta \in (0,1)$ is the discount *factor*, and $\pi_i(t)$ is country i's periodic payoff according to (1) in stage t when the abatement profile is $(q_1(t), \ldots, q_n(t))$. A strategy profile $(\sigma_1, \ldots, \sigma_n)$ is a subgame-perfect equilibrium if, for every history, there is no country that can increase its discounted payoff by deviating from its strategy, provided that all other players follow their strategies in the continuation of the game. A subgame-perfect equilibrium is weakly renegotiation-proof (Farrell and Maskin, 1989) if there do not exist two histories such that all players strictly prefer the continuation equilibrium in the one to the continuation equilibrium in the other.

We can now state our main result.

Proposition 1 *For any positive integer $n \geq 2$ and positive real numbers b and c, there exists a weakly renegotiation-proof equilibrium with* **a** *as the equilibrium path if the countries' repeated game payoffs are discounted by discount factor δ in the interval $\big[(n-1)/n,\, 1\big)$.*

In the remainder of this section, we describe a strategy profile with an uncomplicated structure, leading to the Pareto-efficient agreement **a**. According to a general theorem stated in the next section, this strategy profile is a weakly renegotiation-proof equilibria for $\delta \geq (n-1)/n$, thereby proving Proposition 1. Since $\delta \geq (n-1)/n$ is equivalent to the discount *rate*, $1-\delta$, not exceeding $1/n$, this shows that **a** can be implemented in a self-enforcing manner, provided that the discount rate is sufficiently low and the number of countries is sufficiently small. In the next section we also consider slightly more complicated renegotiation-proof equilibria that implement **a** in a self-enforcing manner even if $\delta < (n-1)/n$, provided that (10) and (11) are satisfied.

Following Abreu (1988) we consider simple strategy profiles, consisting of an equilibrium path to be implemented, and n punishment paths, one for each player. The equilibrium path is followed until a single country deviates, an occurrence that leads to this player's punishment path being initiated in the next period. Also any unilateral deviation from a punishment path leads to the initiation of the (new) deviating country's punishment path. Through these rules, the $n+1$ paths specify a strategy for each player. Hence, with $\mathbf{a}$ as the equilibrium path, we need only construct the n punishment paths and show that the resulting simple strategy profile is indeed a weakly renegotiation-proof equilibrium.

To construct the path, $\mathbf{p}_i$, used to punish country i, consider a function which for any n determines a subset $P_i(n) \subset N$ of punishing countries. Let $P_i(n)$ have the properties that (1) $i \notin P_i(n)$ and (2) the number of countries in $P_i(n)$ equals $n/2$ if n is even and $(n+1)/2$ if n is odd. The interpretation is that each country in $P_i(n)$ punishes a unilateral deviation by country i by choosing the abatement level q^1 in the period immediately following country i's deviation, while countries in $N \backslash P_i(n)$ (including country i) abates at the Pareto-efficient level q^n. In the subsequent periods all countries return to the Pareto-effcient abatement level. Hence, the punishment path of country i is:

$$\mathbf{p}_i = (\underbrace{q^1, \ldots, q^1}_{j \in P_i(n)}, \underbrace{q^n, \ldots, q^n}_{j \in N \backslash P_i(n)}), (\underbrace{q^n, \ldots, q^n}_{j \in N}), (\underbrace{q^n, \ldots, q^n}_{j \in N}), \ldots .$$

In the next section we show that, for any positive integer $n \geq 2$ and positive real numbers b and c, the simple strategy profile described by the $n+1$ paths $(\mathbf{a}, \mathbf{p}_1, \ldots, \mathbf{p}_n)$ is a weakly renegotiation-proof equilibrium if the discount factor, δ, satisfies $\delta \geq (n-1)/n$. Having $n/2$ (or $(n+1)/2$ if n is odd) countries punishing for one period by choosing their best response q^1 of the stage game is sufficiently many to discipline a potential deviator, while being sufficiently few to ensure that each punisher in $P_i(n)$ gains at least as much by reducing its abatement as it loses by the fact that the other countries in $P_i(n)$ abate less.

3 Participation and punishment

In this section we consider a class of strategy profiles in the repeated games described in Section 2, and establish as Theorem 1 under what parameter values members of this class are weakly renegotiation-proof equilibria. Since the strategy profiles used to establish existence in

Proposition 1 are members of this class, this result follows as a corollary to Theorem 1. We also present Proposition 2, our result on the maximal treaty depth for low discount factors and a large number of countries.

Fix the set of countries $N = \{1, \ldots, n\}$. Let $M = \{i_i, \ldots, i_m\}$ $(\subseteq N)$ be the signatories to a treaty, with m members (where $0 < m \leq n$). The treaty specifies that the agreement $\mathbf{a}^s$ be implemented, where

$$\mathbf{a}^s = (\underbrace{q^s, \ldots, q^s}_{j \in M}, \underbrace{q^1, \ldots, q^1}_{j \in N \backslash M}), (\underbrace{q^s, \ldots, q^s}_{j \in M}, \underbrace{q^1, \ldots, q^1}_{j \in N \backslash M}), \ldots,$$

and $q^s = sb/c$ with $s > 1$. Hence, the signatories of the treaty abate s times the level that constitutes the individual country's best response, while the non-signatories choose the best response level.

Since each signatory is not playing a best response of the stage game, a deviation from the agreement by a signatory must be prevented by the threat of future punishment. To construct the path, $\mathbf{p}_i^s$, used to punish country $i \in M$, consider a set $P_i \subset M$ of punishing countries, satisfying $i \notin P_i$. We assume that $|P_i|$, the number of countries in P_i, is the same for all $i \in M$, while of course the identities of the countries may not (and can not) be the same. Write $k = |P_i|$. Each country in P_i punishes a unilateral deviation by country i by choosing the abatement level $q^p = pb/c$ in the period immediately following country i's deviation, where $p \geq 0$. The other signatories including country i (i.e., the countries in $M \backslash P_i$) abate at the agreed upon level q^s. In the subsequent periods all signatories return to the agreed upon level q^s. All non-signatories (i.e, $j \in N \backslash M$) continue to play their best response q^1 throughout. Hence, the punishment path of country $i \in M$ is:

$$\begin{aligned}\mathbf{p}_i^s = (&\underbrace{q^p, \ldots, q^p}_{j \in P_i}, \underbrace{q^s, \ldots, q^s}_{j \in M \backslash P_i}, \underbrace{q^1, \ldots, q^1}_{j \in N \backslash M}), \\ &(\underbrace{q^s, \ldots, q^s}_{j \in M}, \underbrace{q^1, \ldots, q^1}_{j \in N \backslash M}), (\underbrace{q^s, \ldots, q^s}_{j \in M}, \underbrace{q^1, \ldots, q^1}_{j \in N \backslash M}), \ldots.\end{aligned}$$

Since each non-signatory is playing a best response of the stage game, a deviation from the agreement by a non-signatory requires no punishment. Hence, even if a non-signatory unilaterally deviates from $\mathbf{a}^s$ or $\mathbf{p}_i^s$ for some $i \in M$, the path in question is simply continued, meaning that any such unilateral deviation is followed by $\mathbf{a}^s$. Hence, formally, the punishment path of country $i \in N \backslash M$ equals $\mathbf{a}^s$.

The simple strategy profile determined by the $n+1$ paths

$$(\mathbf{a}^s, \mathbf{p}^s_{i_1}, \ldots, \mathbf{p}^s_{i_m}, \underbrace{\mathbf{a}^s, \ldots, \mathbf{a}^s}_{j \in N \setminus M}) \tag{4}$$

corresponds to what Froyn and Hovi (2008) refer to as "Penance k" . In Theorem 1 we establish under what conditions this strategy profile is a weakly renegotiation-proof equilibrium.

Since the identities of the signatories and the punishing countries do not matter, as all countries are identical, the conditions of Theorem 1 depend on only the parameters δ (the discount factor), n (the total number of countries), m (the broadness of the treaty; i.e., the number of signatories), k (the number of punishing countries), s (the depth of the treaty), and p (the severity of the punishment). In fact, δ, k, s and p are sufficient to decide whether the simple strategy profile determined by the paths in (4) is weakly renegotiation-proof, while n and m do not matter as long as they satisfy $n \geq m > k$.

Theorem 1 *The simple strategy profile determined by (4) with $s > 1$ and $p \geq 0$ is a weakly renegotiation-proof equilibrium for $\delta \in (0,1)$ if and only if k, s and p satisfy $s > p$ and*

$$\frac{1}{2\delta} \cdot \frac{(\max\{s-1, |p-1|\})^2}{s-p} \;\leq\; k \;\leq\; \frac{1}{2}(s+p)\,. \tag{5}$$

In the kind of weakly renegotiation-proof equilibria used to establish existence in Proposition 1, we have that

$$s = n \qquad \text{and} \qquad p = 1\,.$$

Then expression (5) simplifies to

$$\frac{1}{\delta}(n-1) \;\leq\; 2k \;\leq\; n+1\,. \tag{6}$$

If n is even and $k = n/2$, then the right inequality is satisfied, and the left inequality is satisfied if

$$\delta \geq \frac{n-1}{n}\,. \tag{7}$$

If n is odd and $k = (n+1)/2$, then the right inequality is satisfied, and the left inequality is satisfied if

$$\delta \geq \frac{n-1}{n+1}\,,$$

which is implied by (7). In either cases, $\delta \in \big[(n-1)/n,\, 1\big)$ is sufficient, thus showing that Proposition 1 follows as a corollary to Theorem 1.

Proposition 1 shows that a weakly renegotiation-proof Pareto-efficient agreement can be implemented if $\delta \geq (n-1)/n$. Hence, few countries and a high δ, reflecting low time discounting and a short detection lag, contribute to the feasibility of a Pareto-efficient treaty. However, it is of interest to investigate what can be achieved with many countries and a low δ, reflecting high time discounting and a long detection lag. Therefore, in Proposition 2 we analyze the complement case where $\delta < (n-1)/n$.

Proposition 2 *Assume $\delta \in \big(0, (n-1)/n\big)$,[2] and consider the simple strategy profile determined by (4) with $s > 1$ and $p \geq 0$. Maximal treaty depth in a weakly renegotiation-proof equilibrium is given by*

$$s(\delta) = 1 + 2k\delta + 2\sqrt{k\delta(1-k(1-\delta))}\,, \tag{8}$$

with the severity of punishment given by $p(\delta) = 2k - s(\delta)$ $(\in (0,1])$, if there exists $k \in \mathbb{N}$ s.t. $(k-1)/k \leq \delta < \big((2k-1)/2k\big)^2$, and by

$$s(\delta) = 1 + k\delta + \sqrt{k\delta(2+k\delta)}\,, \tag{9}$$

with the severity of punishment given by $p(\delta) = 0$, if there exists $k \in \mathbb{N}$ s.t. $\big((2k-1)/2k\big)^2 \leq \delta < k/(k+1)$. In both cases, the number of punishing countries equals k $(\in \{1, \ldots, n-1\})$, and the number of participating countries can be any m satisfying $k < m \leq n$.

Proposition 2, which is illustrated by Figure 1, means that Pareto-efficient treaty depth, $s = n$, is feasible if and only if maximal treaty depth, $s(\delta)$, satisfies $s(\delta) \geq n$. This holds under the following conditions:

$$n \text{ odd} \quad \text{and} \quad \delta \geq \frac{n-1}{n+1}\,, \tag{10}$$

$$n \text{ even} \quad \text{and} \quad \delta \geq \left(\frac{n-1}{n}\right)^2. \tag{11}$$

In case (10), $k = (n+1)/2$ and $p = 1$, and Pareto-efficiency for the lowest discount factor is implemented by the weakly renogotiation-proof equilibrium considered in Proposition 1. In case (11), however, $k = n/2$ is

[2]The upper bound $(n-1)/n$ on the discount factor δ ensures that the number of punishing countries k determined by the proposition is smaller than n.

combined with $p = 0$, implying that Pareto-efficiency for the lowest discount factor is implemented by a weakly renogotiation-proof equilibrium with a harsher punishment than the one considered in Proposition 1.

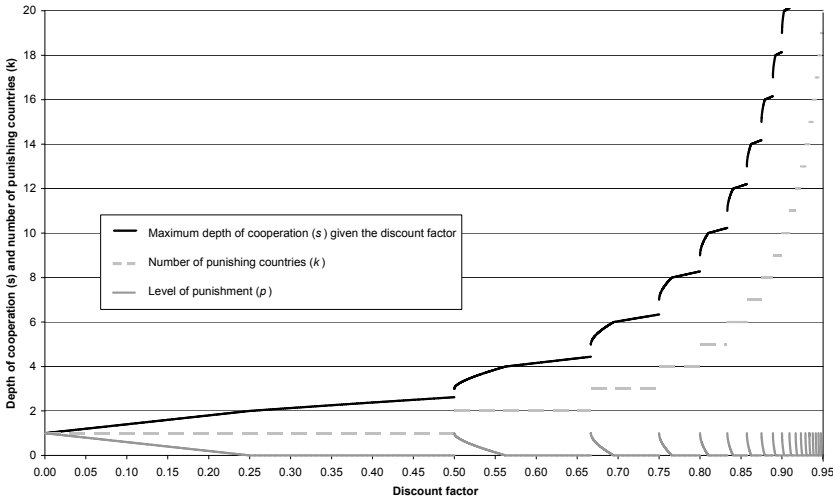

Figure 1: Maximal treaty depth as a function of the discount factor
The black curve depicts maximal depth of cooperation as a function of the discount factor. The dotted grey, horizontal lines depict the number of punishing countries, while the thin grey curves depict the severity of punishment.

If neither condition (10) nor condition (11) is satisfied, due to a high n and a low δ, then Proposition 2 shows that there is no trade-off between treaty depth and treaty broadness: treaty depth has to give, while full participation is feasible even in a many country world with high time discounting and long detection lags.

In the appendix we prove Theorem 1 and show how Proposition 2 follows from the theorem.

4 Discussion

In this section we first explore the equilibrium concept that underlies the analysis, before discussing our results in the context of climate change.

4.1 Weakly renegotiation-proof equilibrium

In this paper we have applied the game-theoretic concept of "weakly renegotiation-proof equilibrium" (Farrell and Maskin, 1989) to study self-enforcing climate agreements. In this section we first discuss what this equilibrium concept means for the significance of our results, before providing numerical illustrations. We also indicate how our findings may have relevance for the ongoing negotiations on a follow-up agreement to the Kyoto Protocol.

A weakly renegotiation-proof equilibrium is a subgame-perfect equilibrium. This presumes that the countries are coordinated in the sense that they all play according to a particular strategy profile described by (4). In other words, the analysis assumes that all countries have agreed upon who will participate in the treaty and how unilateral deviations will be punished. Hence, the framework is designed to analyze how non-compliance can be avoided: it shows how signatories can be induced to fulfill their treaty obligations under the threat of future punishment.

The framework is not suitable for analyzing how coordination is achieved: it is incapable of answering how countries manage to agree on a particular treaty, involving strategies specifying behavior under both compliance and non-compliance. Even though it is of interest to consider a situation where coordination has not yet occurred and where countries seeking a Pareto-efficient climate agreement attempt to punish a would-be-free-rider into joining the effort, this is not an equilibrium of a repeated game, and thus, outside the scope of the present paper.

Weak renegotiation-proofness considers the possibility of a coordinated deviation by *all* countries, but abstracts from the possibility that also a coordinated deviation by a subset of countries can be profitable. Coordinated deviations by a subset of players in a game have been considered by Bernheim et al. (1987) through their concepts "coalition-proof Nash equilibrium" in static games and "perfectly coalition-proof Nash equilibrium" in finite horizon dynamic games. The latter concept has been generalized to infinite horizon games by Asheim (1997, Definition 2). Perfectly coalition-proof equilibrium in infinite horizon games is not, however, a refinement of weak renegotiation-proofness. To our knowledge, there exists no refinement of the concept of weakly renegotiation-proof equilibrium that takes into account that also a subset of players can gain by implementing a coordinated deviation.

4.2 Related literature on climate change

Proposition 1 is a "folk-theorem" result for weakly renegotiation-proof equilibria, showing that an efficient outcome can be disciplined through the threat of punishment if δ is high enough (van Damme, 1989, is an early contribution of this kind in the case of weak renegotiation-proofness). It shows that a Pareto-efficient climate agreement can be implemented if $\delta \geq (n-1)/n$.

We can illustrate how a weakly renegotiation-proof equilibrium implementing a Pareto-efficient climate agreement may look like, by applying the result of Proposition 1 to a two-region world, each with $n/2$ countries. Refer to the regions as A and B. Since the total number of countries is even with two equally sized regions, we may satisfy the requirements on the subset $P_i(n)$ of countries punishing a deviating country i (namely (1) $i \notin P_i(n)$ and (2) the number of countries in $P(n)$ equals $n/2$) by having a unilateral deviation by a country in region A be punished by all countries in region B and vice versa.

This means that a unilateral deviation by a country in region A triggers a one-period reduction in abatement by all countries in region B. This inflicts an equally hard punishment on all countries in A. On the other hand, since $k = n/2 < (n+1)/2 = (s+p)/2$, it follows from Proposition 4 of the appendix that all countries in region B strictly benefit by carrying out the punishment.

This arrangement can be contrasted with that of Asheim et al. (2006), where a global weakly renegotiation-proof equilibrium is cast in terms of two regional agreements. In this equilibrium, a unilateral deviation by a country in region A triggers a one-period reduction in abatement by all the other countries in region A, and likewise in region B. Hence, the weakly renegotiation-proof equilibrium entails that the countries that benefit during the one-period punishment phase are in the same region as the deviating country, while the countries in region B are harmed twice: *both* by the initial unilateral deviation of a country in region A *and* by the subsequent punishment by the other countries in region A.

The alternative proposed in our paper has the appealing feature of inflicting the punishments within the region which is to blame for the temporary break-down of cooperation, and rewarding the innocent countries of the other region. With this set-up the countries in the same region as the deviator are harmed twice, an arrangement that might have

a more disciplining effect on a potential deviator than the equilibrium proposed by Asheim et al. (2006).

In the present model with a continuum choice of abatement levels, there is full participation independently of the number of countries, provided that the discount rate is sufficiently high. So in a world with 200 countries, all 200 countries abate at the Pareto-efficient level $q^{200} = 200b/c$, *with 100 countries punishing a unilateral deviation* by reducing their abatement to $q^1 = b/c$.

In contrast, the model of Asheim et al. (2006), having a binary choice of abatement levels, leads to a fixed absolute number of participating countries. E.g., if the cost of abatement in the binary choice model is 8 times each country's benefit of abatement, then the analysis of Asheim et al. (2006) yields 18 participating countries, 9 countries in each region, *with 8 countries punishing a deviator*, even in a world with 200 countries. With these parameter values, the result that 8 countries punish carries over to the analysis of Froyn and Hovi (2008). However, by relaxing a restrictive assumption made by Asheim et al. (2006), namely that of a two-region world where all other countries in the deviator's region must punish, they are able to construct a weakly renegotiation-proof equilibrium where there is full participation.

What is the reason for this striking divergence between the binary and continuum abatement choice models? Under the parameter values of the previous paragraph, with the cost of abatement being 8 times each country's benefit of abatement, punishing is at least as good as renegotiating back to cooperation only if there is at most 8 punishing countries. Hence, in the binary choice model, the requirement for weak renegotiation-proofness precludes more than 8 punishing countries. On the other hand, since the binary choice is fixed, each country's short-term gain from non-compliance (i.e., by not abating when specified to do so) is independent of the total number of countries. Hence, also the requirement for subgame-perfectness is unrelated to the total number of countries, leading to a fixed *absolute* number of punishing countries.

In comparison, in the continuum abatement choice model, the Pareto-efficient abatement level of each country is a linear function of the total number of countries: $q^n = nb/c$ (cf. equation (2)). In the equilibrium of Proposition 1, this relaxes the requirement for weak renegotiation-proofness (r.h.s. of (6)), but tightens the requirement for subgame-perfectness (l.h.s. of (6)). As we have shown, with a fixed *fraction* ($\approx 1/2$)

of punishing countries, both these requirements are satisfied.

This difference between the binary and continuum abatement choice models leads also to different requirements for the discount factor δ. In the binary choice model with the cost of abatement being 8 times each country's benefit of abatement, Froyn and Hovi (2008) find that a Pareto-efficient agreement with 200 countries can be implemented if the discount factor exceeds 0.95; in fact, a discount factor equal to 0.95 is sufficient independently of how many countries the world consists of.

In comparison, in the continuum choice model it follows from (11) that a Pareto-efficient agreement between 200 countries can be implemented if and only if the discount factor exceeds 0.99. Moreover, by applying Proposition 2, it follows that only a shallow treaty is feasible if $\delta = 0.95$, with all 200 countries abating $q^{39} = 39b/c$ and 20 countries punishing a deviating country by reducing their abatement to $q^1 = b/c$. Hence, even though this less ambitious agreement has full participation, the resulting total abatement is less than 20% of the Pareto-efficient level.

Thus, Proposition 2 considers what can be implemented if δ is not sufficiently high, echoing the kind of analysis done by Abreu (1986, 1988), only that we here consider weakly renegotiation-proof equilibria. This problem has been mostly ignored in the literature on self-enforcing climate agreement (with Finus and Rundshagen, 1998, Section 4, as a notable exception). In our view, the real possibility of high time discounting and long detection lags makes it a subject worthy of analysis.

To support this claim, note that the first commitment period of the Kyoto Protocol is 5 years, and the Protocol's rules for emissions accounting and reporting entail that deviations will be detected no earlier than 2–3 years after the end of the commitment period. With such considerable time lags between deviations and punishments, the relevant discount rate will be high, and under such circumstances, our analysis shows that a shallow agreement might result. More generally, our findings highlight the importance of designing a climate agreement where non-compliance is detected early and punishments are carried out promptly. The choice between agreements with quantitative restrictions vs. agreements where the parties commit to use particular policy instruments, like emission taxes, illustrate this. Our findings might serve as an argument in the favor of the latter type of agreements if these can be designed with shorter detections lags than the former.

Appendix: Proof of Theorem 1 and Proposition 2

In this appendix we characterize weak renegotiation-proofness for the simple strategy profile determined by (4) when $s > 1$ and $p \geq 0$.

Proof of Theorem 1

We first find through Proposition 3 the condition that ensures that this strategy profile is a subgame-perfect equilibrium and then proceed to provide through Proposition 4 the condition that ensures that such a subgame-perfect equilibrium is weakly renegotiation-proof. Theorem 1 is a direct consequence of Propositions 3 and 4.

Subgame-perfectness

Let $\boldsymbol{\alpha}^s$ denote the average discounted payoff of each signatory when $\mathbf{a}^s$ is followed. Likewise, let $\boldsymbol{\pi}_i^s$ denote the average discounted payoff of a signatory i when $\mathbf{p}_i^s$ is followed.

Lemma 1 *The punishment inflicted on country i through $\mathbf{p}_i^s$ relative to following the agreement $\mathbf{a}^s$ equals*

$$\boldsymbol{\alpha}^s - \boldsymbol{\pi}_i^s \;=\; (1-\delta)(s-p)k\frac{b^2}{c}\,.$$

Proof. By inserting $\mathbf{a}^s$ into (1) and (3), we obtain

$$\boldsymbol{\alpha}^s \;=\; bms\frac{b}{c} \;+\; b(n-m)\frac{b}{c} \;-\; \frac{c}{2}\Big(s\frac{b}{c}\Big)^2\,. \tag{A1}$$

By inserting $\mathbf{p}_i^s$ into (1) and (3), we obtain

$$\boldsymbol{\pi}_i^s \;=\; (1-\delta)\left(b(m-k)s\frac{b}{c} \;+\; bkp\frac{b}{c} \;+\; b(n-m)\frac{b}{c} \;-\; \frac{c}{2}\Big(s\frac{b}{c}\Big)^2\right) \;+\; \delta\boldsymbol{\alpha}^s\,.$$

The lemma is obtained by subtracting $\boldsymbol{\pi}_i^s$ from $\boldsymbol{\alpha}^s$. ■

Lemma 1 gives the size of the future punishment inflicted on a signatory when it deviates from the simple strategy profile determined by (4). This must be compared to the short-term gain that a signatory can reap by deviating from the abatement prescribed by this strategy profile. The size of this short-term gain is provided by the following lemma.

Lemma 2 *Assume that the simple strategy profile determined by (4) prescribes the abatement rb/c, with $r \geq 0$, for country i. Then the maximal short-term gain that country i can reap through a unilateral deviation equals*

$$\frac{(r-1)^2}{2}\cdot\frac{b^2}{c}\,.$$

Proof. The short-term gain of a unilateral deviation by country i does not depend on the fixed behavior of the other countries. Furthermore, independently of r and the behavior of the other countries, country i maximizes its short-term payoff by choosing $q_i = b/c$. Hence,

$$\left[b\frac{b}{c} - \frac{c}{2}\Big(\frac{b}{c}\Big)^2\right] - \left[br\frac{b}{c} - \frac{c}{2}\Big(r\frac{b}{c}\Big)^2\right] = \frac{(r-1)^2}{2}\cdot\frac{b^2}{c}$$

is the maximal short-term gain that country i can reap through a unilateral deviation. ■

We can now characterize subgame-perfectness for the set of strategy profiles considered.

Proposition 3 *The simple strategy profile determined by (4) with $s > 1$ and $p \geq 0$ is a subgame-perfect equilibrium for $\delta \in (0,1)$ if and only if k, s and p satisfy $s > p$ and*

$$\frac{1}{2\delta}\cdot\frac{(\max\{s-1,|p-1|\})^2}{s-p} \leq k\,. \tag{A2}$$

Proof. *If part.* Let k, s and p satisfy $s > 1$, $p \geq 0$, $s > p$ and (A2). We only need to check for one-period deviations, since it follows from the theory of repeated games with discounting (Abreu, 1988, p. 390) that a player cannot gain by a multi-period deviation if he cannot gain by some one-period deviation.

Throughout, the strategy profile prescribes that non-signatories choose b/c as their abatement level. Hence, even though any one-period deviation by a non-signatory is not punished, it follows from Lemma 1 that they have no incentive to deviate.

Signatories are prescribed to choose sb/c along $\mathbf{a}^s$. It follows from Lemmas 1 and 2 that there is no profitable deviation if

$$(1-\delta)\frac{(s-1)^2}{2}\cdot\frac{b^2}{c} \leq \delta(1-\delta)(s-p)k\frac{b^2}{c}\,, \tag{A3}$$

which can be rewritten as

$$\frac{1}{2\delta}\cdot\frac{(s-1)^2}{s-p} \leq k\,. \tag{A4}$$

The signatories (including country i itself) not inflicting punishment on country i in the first stage of $\mathbf{p}_i^s$ and all signatories in later stages of $\mathbf{p}_i^s$ are also prescribed to choose sb/c, followed by $\mathbf{p}_j^s$ if there is a unilateral deviation by a signatory j and by $\mathbf{a}^s$ if there is no such deviation. Hence, also in these cases there is no profitable deviation if (A4) is satisfied.

Finally, the signatories inflicting punishment on country i are prescribed to choose pb/c in the first stage of $\mathbf{p}_i^s$, followed by $\mathbf{p}_j^s$ if there is a unilateral deviation by a signatory j and by $\mathbf{a}^s$ if there is no such deviation. By Lemmas 1 and 2, there is no profitable deviation if

$$(1-\delta)\frac{(p-1)^2}{2}\cdot\frac{b^2}{c} \;\leq\; \delta(1-\delta)(s-p)k\frac{b^2}{c}\,, \tag{A5}$$

which can be rewritten as

$$\frac{1}{2\delta}\cdot\frac{(p-1)^2}{s-p} \;\leq\; k\,. \tag{A6}$$

Since $s>1$, inequalities (A4) and (A6) are equivalent to inequality (A2).

Only-if part. Suppose $s\leq p$. Since $s>1$, it follows from (A3) that there is a profitable deviation from $\mathbf{a}^s$.

Assume that $s>p$. Suppose that (A4) is not satisfied. Then it follows from (A3) that there is a profitable deviation from $\mathbf{a}^s$. Suppose that (A6) is not satisfied. Then it follows from (A5) that there is a profitable deviation from the first stage of each punishment path $\mathbf{p}_i^s$.

Since (A4) and (A6) are equivalent to (A2), we have that $s>p$ and (A2) are necessary conditions for the subgame-perfectness of the simple strategy profile determined by (4) with $s>1$ and $p\geq 0$. ■

Weak renegotiation-proofness

Let $\boldsymbol{\beta}_i^s$ denote the average discounted payoff of each of the signatories inflicting punishment on country i when $\mathbf{p}_i^s$ is implemented.

Proposition 4 *Assume that the simple strategy profile determined by (4) with $s>1$ and $p\geq 0$ is a subgame-perfect equilibrium for $\delta\in(0,1)$. Then this strategy profile is weakly renegotiation-proof if and only if k, s and p satisfy*

$$k \;\leq\; \frac{1}{2}(s+p)\,. \tag{A7}$$

Proof. By the definition of weak renegotiation-proofness, we must determine when there do not exist two continuation equilibria such that all players strictly prefer the one to the other. Given the structure of the simple strategy profile determined by (4), there exist $m+1$ different continuation equilibria, implementing the play of $\mathbf{a}^s$ and $\mathbf{p}_i^j$ for all $j\in M$.

Since the strategy profile is subgame-perfect, it follows from Proposition 3 that $s>p$, implying that $\boldsymbol{\alpha}^s>\boldsymbol{\pi}_i^s$. It follows that all non-signatories as well as all signatories not inflicting punishment strictly prefer $\mathbf{a}^s$ to any punishment path $\mathbf{p}_i^s$. If $\boldsymbol{\alpha}^s>\boldsymbol{\beta}_i^s$, then all countries, including the punishing signatories, strictly prefer the continuation equilibrium in the "empty" history

to the continuation equilibrium following a unilateral deviation by country i. If $\boldsymbol{\alpha}^s \leq \boldsymbol{\beta}_i^s$, then the continuation equilibrium following a unilateral deviation by country i is a best continuation equilibrium for each signatory inflicting punishment on country i and a worst continuation equilibrium for country i itself, implying that all players never strictly prefer one continuation equilibrium to another.

Hence, there do not exist two continuation equilibria such that all countries strictly prefer the one to the other if and only if

$$\boldsymbol{\beta}_i^s \; - \boldsymbol{\alpha}^s \; \geq \; 0\,. \tag{A8}$$

By inserting $\mathbf{p}_i^s$ into (1) and (3), it follows that

$$\boldsymbol{\beta}_i^s \; = \; (1-\delta)\left(b(m-k)s\frac{b}{c} \; + \; bkp\frac{b}{c} \; + \; b(n-m)\frac{b}{c} \; - \; \frac{c}{2}\Big(p\frac{b}{c}\Big)^2\right) \; + \; \delta\boldsymbol{\alpha}^s\,.$$

By comparing with (A1) we obtain

$$\boldsymbol{\beta}_i^s \; - \boldsymbol{\alpha}^s \; = \; (1-\delta)(s-p)\Big(\tfrac{1}{2}(s+p)-k\Big)\frac{b^2}{c}\,,$$

implying that (A8) is equivalent to (A7). ■

Proof of Proposition 2

Assume $\delta \in \big(0,(n-1)/n\big)$, and consider the simple strategy profile determined by (4) with $s>1$ and $p \geq 0$. We now apply Theorem 1 to find the maximum treaty depth for which this simple strategy profile is weakly renegotiation-proof, thereby proving Proposition 2. For the statement of Proposition 2 and the working of the proof below, it is helpful to note that

$$\left\{\left[\frac{k-1}{k},\left(\frac{2k-1}{2k}\right)^2\right),\;\left[\left(\frac{2k-1}{2k}\right)^2,\frac{k}{k+1}\right)\right\}_{k\in\{1,\,\dots\,,n-1\}}$$

is a partition of the interval $\big[0,(n-1)/n\big)$.

It can be checked that the functions

$$s:(0,(n-1)/n)\to\mathbb{R}_+$$
$$p:(0,(n-1)/n)\to\mathbb{R}_+$$

as given in Proposition 2 satisfy $s(\delta)\in(1,\infty)$ and $p(\delta)\in[0,1]$ for all $\delta \in \big(0,(n-1)/n\big)$. Furthermore, $s(\delta)-1 \geq |p(\delta)-1|$ for all $\delta\in(0,(n-1)/n)$, since $s(\delta)=1+4\delta$ and $p(\delta)=1-4\delta$ for $\delta\in(0,\frac{1}{4})$ and $s(\delta)\geq 2$ for $\delta\geq\frac{1}{4}$. Hence, Theorem 1 implies that if, for every $\delta\in\big(0,(n-1)/n\big)$, $s(\delta)$ is the maximum s for which there exist $p\in[0,s)$ and $k\in\{1,\dots,n-1\}$ satisfying (A4) and (A7), then $s(\delta)$ is the maximum s also under (5) of Theorem 1.

There are two cases to consider.

Case A: $p \in (0, s)$. In this case, we can assume that (A7) is satisfied with equality, because otherwise (A4) could have been relaxed by reducing p. Hence, $2k = s + p$, implying that (A4) and (A7) can be rewritten as:

$$f(s; k, \delta) := s^2 - 2(1 + 2k\delta)s + (1 + 4k^2\delta) \leq 0 \quad \text{and} \quad s < 2k\,. \tag{A9}$$

The equation $f(s; k, \delta) = 0$ has a solution if and only if $(k-1)/k \leq \delta$. If $(k-1)/k \leq \delta$, then the maximum s for which $f(s; k, \delta) \leq 0$ is given by

$$s^A(k, \delta) := 1 + 2k\delta + 2\sqrt{k\delta(1 - k(1-\delta))}\,.$$

Furthermore, $s^A(k, \delta) < 2k$ is equivalent to $\delta < \big((2k-1)/2k\big)^2$. Hence, (A9) can be satisfied for a maximized value of s if and only if $(k-1)/k \leq \delta < \big((2k-1)/2k\big)^2$.

Case B: $p = 0$. In this case, (A4) and (A7) can be rewritten as:

$$g(s; k, \delta) := s^2 - 2(1 + k\delta)s + 1 \leq 0 \quad \text{and} \quad s \geq 2k\,. \tag{A10}$$

The maximum s for which $g(s; k, \delta) \leq 0$ is given by

$$s^B(k, \delta) := 1 + k\delta + \sqrt{k\delta(2 + k\delta)}\,.$$

Furthermore, $s^B(k, \delta) \geq 2k$ is equivalent to $\big((2k-1)/2k\big)^2 \leq \delta$. Hence, (A10) can be satisfied if and only if $\big((2k-1)/2k\big)^2 \leq \delta$.

The analysis of cases A and B above has the following implications:

- If there exists $\bar{k} \in \mathbb{N}$ s.t. $\big((2\bar{k}-1)/2\bar{k}\big)^2 \leq \delta < \bar{k}/(\bar{k}+1)$, then only case B is possible. Since s^B is increasing in k, the treaty depth is maximized by choosing the largest k consistent with $\big((2k-1)/2k\big)^2 \leq \delta$, namely $k = \bar{k}$ ($\in \{1, \ldots, n-1\}$), so that $s = s^B(\bar{k}, \delta)$ and $p = 0$.
- If there exists $\bar{k} \in \mathbb{N}$ s.t. $(\bar{k}-1)/\bar{k} \leq \delta < \big((2\bar{k}-1)/2\bar{k}\big)^2$, then case A is possible with $k = \bar{k}$ and, provided $\bar{k} > 1$, case B is possible with $k < \bar{k}$. With $\bar{k} > 1$, it can be shown that, for all $\delta \in \big[(\bar{k}-1)/\bar{k},\ \big((2\bar{k}-1)/2\bar{k}\big)^2\big)$ and $k < \bar{k}$, $s^A(\bar{k}, \delta) > s^B(k, \delta)$, implying that treaty depth is maximized by choosing $k = \bar{k}$ ($\in \{1, \ldots, n-1\}$), $s = s^A(\bar{k}, \delta)$, and $p = 2\bar{k} - s^A(\bar{k}, \delta)$.

By writing $s(\delta) := s^A(k, \delta)$ and $p(\delta) := 2k - s(\delta)$ if there exists $k \in \mathbb{N}$ s.t. $(k-1)/k \leq \delta < \big((2k-1)/2k\big)^2$, and $s(\delta) := s^B(k, \delta)$ and $p(\delta) = 0$ if there exists $k \in \mathbb{N}$ if there exists $k \in \mathbb{N}$ s.t. $\big((2k-1)/2k\big)^2 \leq \delta < k/(k+1)$, Proposition 2 summarizes the results given in the bullet points above.

References

D. Abreu (1986), Extremal equilibria in oligopolistic supergames, Journal of Economic Theory **39**, 191–225.

D. Abreu (1988), On the theory of infinitely repeated games with discounting, Econometrica **56**, 383–396.

G.B. Asheim (1997), Individual and collective time consistency, Review of Economic Studies **64**, 427–443.

G.B. Asheim, C.B. Froyn, J. Hovi and F. Menz (2006), Regional versus global cooperation for climate control, Journal of Environmental Economics and Management **51**, 93–109

S. Barrett (1999), A theory of full international cooperation, Journal of Theoretical Politics **11**, 519–541.

S. Barrett (2002), Consensus treaties, Journal of Institutional and Theoretical Economics **158**, 529–547.

B.D. Bernheim, B. Peleg and M.D. Whinston (1987), Coalition-proof Nash equilibria I: Concepts, Journal of Economic Theory **42**, 1–12.

P.K. Dutta and R. Radner (2007), A strategic analysis of global warming: Theory and some numbers, Journal of Economic Behavior and Organization, forthcoming.

J. Farrell and E. Maskin (1989), Renegotiation in repeated games, Games and Economic Behavior **1**, 327–60.

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

C.B. Froyn and J. Hovi (2008), A climate agreement with full participation, Economics Letters **99**, 317–319.

E. van Damme (1989), Renegotiation-proof equilibria in repeated prisoners' dilemma, Journal of Economic Theory **47**, 206–217.