## Title
Participation Games and international environmental agreements: a nonparametric model



**Authors**
Karp, Larry
Simon, Leo

**Publication Date**
2012-02-13

University of California, Berkeley
Department of Agricultural &
Resource Economics

*CUDARE Working Papers*

*Year 2012* *Paper 1127*

# Participation Games and International Environmental Agreements: a nonparametric model

Larry Karp and Leo Simon

# Participation Games and international environmental agreements: a nonparametric model

LARRY KARP[†] AND LEO SIMON[††]

FEBRUARY 13, 2012

Abstract. We examine the size of stable coalitions in a participation game that has been used to model international environmental agreements, cartel formation, R&D spillovers, and monetary policy. The literature to date has relied on parametric examples; based on these examples, a consensus has emerged that in this kind of game, the equilibrium coalition size is small, except possibly when the potential benefits of cooperation are also small. In this paper, we develop a non-parametric approach to the problem, and demonstrate that the conventional wisdom is not robust. In a general setting, we identify conditions under which the equilibrium coalition size can be large even when potential gains are large. Contrary to previously examined leading special cases, we show that reductions in marginal abatement costs in an international environmental game can increase equilibrium membership, and we provide a measure of the smallest reduction in costs needed to support a coalition of arbitrary size.

Keywords: Stable coalitions, participation game, International Environmental Agreement, climate agreement, trans-boundary pollution, investment spillovers.

JEL classification numbers: C72, H4, Q54

[†]Department of Agricultural and Resource Economics, University of California, Berkeley, and the Ragnar Frisch Center for Economic Research, email: karp@berkeley.edu

[††]Department of Agricultural and Resource Economics, University of California, Berkeley, and the Giannini Foundation of Agricultural Economics, email: leosimon@berkeley.edu .

## 1. Introduction

There is a rapidly expanding literature on the prospects for international environmental agreements (IEAs) to combat trans-boundary pollution problems such as the depletion of the ozone layer, the proliferation of greenhouse gases and, more generally, climate change. Much of this literature is based on a particular participation game. Due to the difficulty of analyzing this game, virtually all research on the topic specifies particular functional forms for the costs and benefits of abatement. Based on this research, the unchallenged consensus opinion among researchers in this field is that IEAs are ineffective precisely when the potential gains from cooperation are large (Ioannidis et al. (2000), Finus (2001),Finus (2003), Barrett (2003)). Moreover, in leading special cases, an innovation—for example, a reduction in abatement costs—that increases the potential gains from cooperation will leave unchanged or reduce equilibrium participation. In these cases, a cost reduction that one would expect to be welfare enhancing has no effect on, or even reduces the fraction of the welfare gains from cooperation that could potentially be realized through the formation of an IEA. The existing theory on IEAs is thus pessimistic: sovereign nations cannot easily be induced to provide public goods; voluntary, self-enforcing agreements are not effective in providing public goods, and plausible remedies, such as investments that reduce abatement costs, will not help and might even be counterproductive. This pessimism spills into policy advice: it has been suggested that efforts devoted to promoting a climate agreement requiring signatories to reduce greenhouse gas emissions are bound to be unsuccessful. A corollary, discussed by Barrett (2003, Ch. 15) and Stiglitz (2006), is that an IEA cannot be successful in the absence of some kind of external punishment.

We show that the basis for this pessimism is fragile, in that it relies on specific functional forms. All of the forms considered in the IEA literature share a common property: the marginal abatement cost function is convex at positive levels of abatement. This property is not required by classical micro-theory, which assumes only that the cost function itself is convex. Indeed, as we illlustrate below in §3.2, when the marginal cost encompass more than one abatement technology, it is typically not convex. When this excessively stringent condition is relaxed, the pessimistic results in the literature can easily be overturned. In this paper, we avoid the straightjacket of specific functional forms, focusing instead on a nonparametric expression for the function—we call it the *joiner's gain function*—that determines the equilibrium level of participation. When the marginal benefits from abatement are linear, there is a simple decomposition of this function that leads to a quite general theory of IEAs, and provides intuition that helps in analyzing the case where marginal benefits of abatement are concave. We begin by showing that if marginal abatement costs are weakly convex and benefits are linear, then the equilibrium size of the IEA never exceeds three countries. To this extent, the basis for the literature's pessimism is more general than previously thought. However, when the restriction that marginal abatement costs be convex is relaxed (maintaining the classical assumption that total costs are convex) the equilibrium IEA can be arbitrarily large. Thus, for some classically admissible cost functions—although not the ones previously analyzed in the literature—the pessimistic conclusion is overturned.

When marginal benefits are linear, our decomposition of the joiner's gain function is a powerful analytic tool. We use it to determine, for any integer up to the total number of potential IEA signatories, the type and minimal size of the cost reduction needed to support an equilibrium with at least this many members, in circumstances where under the original cost function an IEA of this size would not be stable. It is thus possible to capture up to 100% of the potential gains from cooperation that are available under the appropriately specified, reduced cost function. We then show that arbitrarily large coalitions can also be sustained when the benefit of abatement is concave.

In most of this paper—except for §3.2—we restrict attention to cost reductions that globally weakly decrease marginal abatement costs. Such reductions necessarily increase the equilibrium level of abatement by an IEA with a given number of members. This restriction limits the number of special cases that we need to address, allowing us to focus on the most familiar scenario. By contrast, some recent papers have considered technological innovations that reduce total abatement costs, while increasing marginal abatement costs over some range of abatement levels. (See for example Enderes & Ohl (2003), Bauman, Lee & Seeley (2008), Baker & Adu-Bonnah (2008), Baker, Clarke & Shittu (2008).) Under these conditions, a cost reduction may reduce the equilibrium level of abatement for an IEA of a given size.

Our results are policy-relevant for two reasons. First, as noted above, our results show that the theoretical basis for pessimism about self-enforcing IEAs may have been exaggerated. In particular, our work challenges those who would advise policymakers that a successful IEA requires a punishment device. Second, using the techniques that we develop, it is relatively straightforward to determine the effect on participation incentives of a cost reducing technological innovation.

The literature on IEAs to which we contribute is based on a two-stage game developed in D'Aspremont et al. (1983).[1] In the first stage agents have a binary decision, whether to join or stay out of the coalition; in the second stage, agents take an action (abatement,

[1] An alternative to the participation game that we use is based on games of coalition formation, such as Chwe (1994), Ecchia & Mariotti (1998), Xue (1998) and Ray & Vohra (2001). These models invoke a more sophisticated interpretation of rationality, in which agents understand how their provisional decision to join or leave a coalition would affect other agents' participation decisions; nations are "farsighted". Diamantoudi & Sartzetakis (2002), Eyckmans (2001), de Zeeuw (2008) and Osmani & Tol (2009) apply this notion to IEA models. An entirely distinct line of research models climate policy using a cooperative game.

sales, etc.). Members of the coalition internalize the effect of their actions on other coalition members. The outcome at the participation stage is a noncooperative Nash equilibrium, and the outcome at the second ("abatement") stage is either a Nash or Stackelberg equilibrium.

While our research was motivated by the difficulty of sustaining IEAs, this two-stage game has been used to study a variety of problems. Because the essential features of the model arise in several fields, the relevance of our results extends well beyond the setting of this paper. In an IO setting, Donsimoni et al. (1986) provide conditions under which the size of the stable coalition is uniquely determined. Katz (1986) and De Bondt (1997) use the D'Aspremont et al. (1983) definition of coalition stability to study collusion in R&D games with spillovers. Thoron (1998) shows that stable coalitions and Nash equilibrium outcomes to the binary action participation game are equivalent. Kohler (2002) uses the definition of stability to study cooperation in monetary policy. Escrihuela-Villar (2009) and Bos & Harrington Jr (2010) use the same definition of stability in a setting where the post-participation outcome is a non-cooperative equilibrium to a repeated game, rather than the result of optimization by the coalition. Bloch & Dutta (2008) explain the relation between the particular definition of coalition stability used in these papers, and other prominent definitions.

Early applications of this model to IEA formation include Hoel (1992), Carraro & Siniscalco (1993), and Barrett (1994). Barrett (2002) and Finus & Maus (2008) show that coalitions with modest ambitions may be more successful than those that try to fully internalize damages. Using functional forms that are standard for this literature, Barrett (2006) presents a model in which cost-reducing investments can improve the outcome under a stable coalition if there are increasing returns to scale (IRTS) to the adoption decision. With IRTS, the participation game becomes a coordination game, and there may exist a non-cooperative equilibrium in which all nations participate. As Heal & Kunreuther (2011) suggest, a coordination game can also be induced by technological spillovers or trade policy. They provide conditions unders which there is a tipping point in the participation game: if a sufficient number of countries join, it is individually rational for all countries to join. Hoel & De Zeeuw (2010) show that cost-reducing investments (without IRTS) can improve the performance of an IEA if investment can reduce abatement costs sufficiently that a dominant strategy for countries is to abate at the socially optimal level. These examples of coordination games or dominant strategies constrast sharply with the examples cited above in which cost reductions weaken incentives to participate. Taken together, the two sets of examples may suggest that only special kinds of cost reductions, e.g. those associated with IRTS or dominant strategies, promote participation. We show that unless marginal costs are required to be convex, this conjecture is false. Dixit & Olson (2000) and Hong & Karp (2010) study the game when agents use mixed strategies at the participation stage. Ulph (2004), Kolstad (2007), Kolstad & Ulph (2008) and Karp (2012) examine the effect on participation and welfare when agents anticipate learning about costs or damages after the participation stage, before they choose abatement. Kosfeld, Okada & Riedl (2009), Burger & Kolstad (2009) and Dannenberg, Lange & Sturm (2009) find moderate support in laboratory settings for the prediction of low participation in the simplest game; they also examine extensions of that game.

Section 2 introduces some terminology relating to our joiner's gain function and its decomposition. Section 3 focuses on three parametric examples designed to introduce the main themes of the paper. In Section 4, we consider arbitrary convex abatement costs, but require that marginal benefits from abatement are linear. In Section 5, we assume that the abatement benefit function is strictly concave. Section 6 concludes. The Appendix contains the proofs of all Remarks and Propositions.

## 2. Preliminaries

We use “s” (for signatories) subscripts to denote signatories to the agreement and “f” (for freeloaders) subscripts to denote non-signatories. Superscripts indicate the number of signatories in the agreement. Signatories delegate their abatement decisions to the coalition, which chooses abatement to maximize coalition welfare. Each non-signatory ignores the public good nature of abatement and maximizes its individual welfare. Each of the $N$ countries is identical, so the model predicts equilibrium participation and abatement, but not the identity of participants. There are two variants of the second stage, a Cournot variant in which signatories and non-signatories choose abatement levels simultaneously, and a Stackelberg variant in which the signatories choose first. This paper considers the Cournot equilibrium. An online addendum studies the Stackelberg case.

Abatement is a public good. Each country derives a benefit $B(Q)$ from the global level of abatement, $Q$, and incurs a cost based on its individual abatement level, $q$. The individual abatement cost function is convex, denoted by $C(q)$. Because $B'(0) = 1$, non-signatories choose an abatement level of zero whenever $C'(0) > 1$; we examine simple examples of this case in 3.1 and 3.2. In the remainder of this paper, we assume that $C'(0) \leq 1$.

Although the number of signatories, $n$, has a natural interpretation only when $n$ is an integer, all of the variables in the paper are mathematically well-defined for any nonnegative real value of $n$. Accordingly, we define all functions below on the non-negative reals. For $r \in \mathbb{R}_+$, we denote by $\lceil r \rceil$ the smallest integer weakly greater than $r$ and by $\lfloor r \rfloor$ the greatest integer weakly less than $r$.

Suppose that in the first stage of the participation game, $r \in \mathbb{R}_+$ countries choose to participate in an IEA. Let $q_{\mathrm{s}}^r$ and $q_{\mathrm{f}}^r$ denote the abatement levels of signatories and non-signatories respectively, and let $Q^r = rq_s^r + (N - r)q_f^r$ denote aggregate abatement. The objective functions for signatories and non-signatories are, respectively,

$$U^s = rB(Q^r) - C(q_s^r) \tag{1a}$$
$$U^f = B(Q^r) - C(q_f^r) \tag{1b}$$

In an interior equilibrium of the Cournot variant of the abatement stage, the abatement levels $q_s^r$ and $q_f^r$ for signatories and non-signatories respectively are simultaneously determined as the solutions to the first order conditions (2a) and (2b) below:

$$0 = r\frac{\partial B(Q^r)}{\partial q_s^r} - C'(q_s^r) \tag{2a}$$
$$0 = \frac{\partial B(Q^r)}{\partial q_f^r} - C'(q_f^r). \tag{2b}$$

For $r \in \mathbb{R}_+$, we let $g(r)$ denote the "joiner's gain" function, representing the increment to a non-signatory's payoff if it becomes the $r$'th member of an IEA:

$$g(r) = \begin{cases} [B(Q^r) - C(q_s^r)] - \left[B(Q^{r-1}) - C\left(q_f^{r-1}\right)\right] & \text{if } r \leq N \\ -1 & \text{if } r > N \end{cases}. \tag{3}$$

For $r \leq N$, the first term in square brackets is the net (private) benefit obtained by a signatory to an $r$-member IEA; the second term is the net benefit to not joining: if it did not join, the non-signatory would benefit from aggregate abatement $Q^{r-1}$ and incur a cost $C(q_f^{r-1})$. Therefore, the difference between the two terms in square brackets is indeed the net gain to a potential signatory of becoming the $r$'th member of an IEA. We assume that a country indifferent between joining an IEA or not chooses to join. Breaking ties in this way, an IEA with an integer $n \geq 2$ members[2] is a *stable equilibrium* iff

$$g(n) \geq 0 > g(n+1). \tag{4}$$

[2]For many specifications of the model, (4) can be satisfied for $n = 1$. Because in this case there is no distinction between signatories and non-signatories, such solutions are clearly uninteresting; we ignore them throughout the paper.

The first inequality is known as the "internal stability condition"—a member of an IEA with $n$ members has no incentive to leave—while the second is called the "external stability condition"—a non-member strictly prefers not to join an IEA that has $n$ members. Note that the external stability condition is vacuously satisfied if $n = N$ because in this case there are no non-members; for this reason we have set $g(n) = -1$ for $n > N$.

To solve the model, the natural approach is to compute the real-valued roots of the equation $g(\cdot) = 0$, and then check each root to see if the integers on either side of it satisfy (4). For a *non-integer* value[3] of $r \in \mathbb{R}_+$ s.t. $g(r) = 0$, we say that $r$ is a *stable root of $g$* if

$$g(\lfloor r \rfloor) \quad \geq \quad 0 \quad > \quad g(\lceil r \rceil). \tag{4'}$$

Having made this point, we shall for the remainder of the paper focus on integer-valued solutions to the model, and index solutions with $n$ rather than $r$.

In sections 2-4, we assume that benefits are linear in abatement, and choose units so that each country's benefit $B(Q)$ is equal to $Q$. Linearity simplifies the analysis considerably, because each non-signatory has a dominant abatement strategy: in particular, its abatement choice is independent of both the size of, and the collective abatement decision made by the IEA; hence under this assumption the Cournot and Stackelberg variants are equivalent. In §5, we generalize the model to allow for benefits that are concave in $Q$.

When there are $n$ signatories and $B(Q) = Q$, each signatory in an interior equilibrium abates at the level $q_{\mathrm{s}}^n > 0$, where $q_{\mathrm{s}}^n$ solves $n = C'(q)$, while each non-signatory abates at $q_{\mathrm{f}} > 0$, where $q_{\mathrm{f}}$ solves $1 = C'(q)$, i.e., $q_{\mathrm{f}}$ is independent of $n$. In this case, we can rewrite the joiner's

[3]Because $\lfloor r \rfloor = \lceil r \rceil$ for any integer $n$, (4′) will be violated whenever the root of $g$ is an integer. A necessary and sufficient condition for *any* real-valued root $r$ to be stable is that $g(\lfloor r \rfloor) \geq 0 > g(\lceil r + \varepsilon \rceil)$ for sufficiently small $\varepsilon > 0$.

gain function (3) as:

$$\begin{aligned} g(n) &= nq_s^n + (N-n)\,q_f - C\left(q_s^n\right) - \left[(n-1)\,q_s^{n-1} + (N-n+1)\,q_f - C\left(q_f\right)\right] \\ &= n\left(q_s^n - q_s^{n-1}\right) + \left(q_s^{n-1} - q_f\right) - \left[C\left(q_s^n\right) - C\left(q_s^{n-1}\right) + C\left(q_s^{n-1}\right) - C\left(q_f\right)\right] \end{aligned} \tag{5}$$

For heuristic reasons, it is helpful to write

$$g(n) = CR\,(n) - UA\,(n)\,, \qquad \text{where in the linear case} \tag{6a}$$

$$UA\,(n) = C\left(q_s^{n-1}\right) - C\left(q_f\right) \; - \; \left(q_s^{n-1} - q_f\right) \; > \; 0; \tag{6b}$$

$$CR\,(n) = n\left(q_s^n - q_s^{n-1}\right) \; - \; \left[C\left(q_s^n\right) - C\left(q_s^{n-1}\right)\right] \; > \; 0. \tag{6c}$$

*UA* (unilateral action) is the utility loss to a non-signatory if it increases its abatement from $q_f$ to the level $q_s^{n-1}$ produced by the signatories to an $(n-1)$-member IEA, while other countries maintain the same abatement levels; because the non-signatory's abatement level $q_f$ is individually optimal, the cost increment of this unilateral step is greater than the benefit increment. *CR* (collective response) is the utility gain to the signatory when all signatories to the augmented coalition respond to the additional member by increasing abatement from $q_s^{n-1}$ to $q_s^n$; because $q_s^n$ is collectively optimal for the augmented coalition, this second increment in aggregate abatement is greater than the second increment in cost.

## 3. Some special cases

To motivate the research orientation of this paper, we present in this section three parametric special cases. These serve as a tutorial for readers unfamiliar with the topic, and illustrate some of the broad conclusions of the literature reviewed above, as well as showing that these conclusions are not robust.

Our first example, drawn from Barrett (2003), has been widely discussed in the literature cited above. It illustrates the pessimistic conventional wisdom that technical innovations that reduce abatement costs tend to reduce, rather than increase equilibrium participation. The second example, based on one in Baker & Adu-Bonnah (2008), generalizes the first but yields quite different conclusions. In particular, it illustrates how non-convexities in the marginal cost function can arise in a natural way through technological innovation, and how these non-convexities can induce stable IEAs with arbitrarily many members. The themes of our third special case, based on Barrett (1994), mimic those of our second, but in a piecewise linear setting. In both the second and third cases, we illustrate the usefulness of our decomposition (6) of the joiner's gain function.

3.1. **Constant marginal costs with capacity constraints.** In our first example, benefits are linear in aggregrate abatement, while marginal costs are constant up to a capacity constraint, normalized to 1. A country's cost of abatement level $q$ is

$$C(q) = \begin{cases} cq & \text{for } 0 \leq q \leq 1 \\ \infty & \text{for } q > 1 \end{cases}. \tag{7}$$

We choose units so that each country's benefit equal to the level of aggregate abatement, $Q$. As the marginal benefit of abatement is constant at 1, this model is interesting only when $c > 1$, so that no country, acting individually, has an incentive to abate.

In this model, signatories to an IEA of size $n$ will abate to the capacity level of 1 iff $n \geq \lceil c \rceil$; if $n < \lceil c \rceil$, signatories abate zero. Non-signatories always abate zero. To obtain the joiner's gain function if this model, we substitute these properties into (5):

$$g(n) = \begin{cases} 0 & \text{if } n < \lceil c \rceil \\ n - c & \text{if } n = \lceil c \rceil \\ 1 - c & \text{if } n > \lceil c \rceil \end{cases}. \tag{8}$$

An IEA with $n > \lceil c \rceil$ is not internally stable, because if one member were to leave, the remaining IEA would continue to abate at capacity. Our tie-breaking assumption (see p. 7

and eq. (4)) implies that IEAs wtih $n < \lceil c \rceil$ are not externally stable. The unique stable equilibrium, $n = \lceil c \rceil$, results in a positive level of abatement.

A striking property of this specification is that equilibrium coalition size weakly increases with abatement costs. Moreover, equilibrium global welfare equals[4] $(N - c) \lceil c \rceil$, which is a "saw-toothed" function of $c$: when $c$ reaches any integer $n$, participation increases discontinuously while costs increase continuously, leading to a jump in equilibrium welfare; as $c$ increases through the interval $(n, n+1)$, participation remains unchanged but costs increase, so that global welfare declines.

One might expect technological innovations that reduce the cost of abatement to increase global welfare, but the model with cost specification (7) delivers the opposite result: it implies that IEAs with large numbers of signatories are sustainable only if abatement is very costly, and that efforts to increase abatement efficiency may be counter-productive. We focus on cost-reducing technological enhancements that *can* increase both equilibrium participation and aggregate welfare.

3.2. **Piecewise Constant marginal costs.** Our second example, based on Baker & Adu-Bonnah (2008), generalizes our first and illustrates that a cost-reducing innovation might either reduce equilibrium membership (as in §3.1) or increase it. We assume that the only source of emissions is electricity generation, and that countries can produce electricity using any combination of coal, gas or solar technologies. By choice of units, a country that uses all coal produces one unit of emissions. A country that uses all solar produces no emissions; one that uses all gas produces $1 - x$ units, with $x < 1$. The marginal cost of switching from coal to solar is $c > 1$ and the marginal cost of switching from coal to gas is $a > x$. These

[4] In equilibrium, $\lceil c \rceil$ countries each abate one unit at a cost of $c$, so that each of $N$ countries benefit from $Q = \lceil c \rceil$ units of abatement. Hence, global welfare is $N \lceil c \rceil - \lceil c \rceil c = (N - c) \lceil c \rceil$.

inequalities imply that it is not individually rational for a country to use either solar or gas. Gas will be an efficient means of reducing emissions iff ${}^{a}/_{x} < c$; if ${}^{a}/_{x} \geq c$ the model below is in effect no different from the one in §3.1, in which the unique equilibrium is $\lceil c \rceil$.

Now consider a technological innovation which reduces ${}^{a}/_{x}$ to below $c$, moving gas onto the efficiency frontier. To avoid dealing with some distracting special cases, we assume that

$$a < 1 \qquad \text{and} \qquad \lceil {}^{c-a}/_{1-x} \rceil \quad > \quad \lceil {}^{a}/_{x} \rceil + 1. \tag{9}$$

Figure 1 graphs the marginal cost of abatement both when $a/x \geq c$ (the heavy solid line) and when ${}^{a}/_{x} < c$ (the dashed step function). To verify its shape, note that $\alpha < x$ units of abatement can be achieved by converting a fraction ${}^{\alpha}/_{x}$ of the coal generators to gas, at a constant marginal cost of ${}^{a}/_{x}$; to abate $\alpha \geq x$, some fraction $y(\alpha)$ of the coal plants must be converted to solar; the minimum fraction required solves $y(\alpha) + x(1 - y(\alpha)) = \alpha$, implying that[5] the marginal cost of increasing $\alpha$ beyond $x$ is ${}^{c-a}/_{1-x}$.

An $n$-member IEA uses whichever generation technology is most profitable; it is straightforward to verify that this method is $\begin{cases} \text{coal} & \text{if } n < \lceil {}^{a}/_{x} \rceil \\ \text{gas} & \text{if } \lceil {}^{a}/_{x} \rceil \leq n < \lceil {}^{c-a}/_{1-x} \rceil \\ \text{solar} & \text{if } \lceil {}^{c-a}/_{1-x} \rceil \leq n \end{cases}$. Taking as given the IEA's technology decision rule, we can now define $g : \mathbb{N} \rightarrow \mathbb{R}$, the joiner's gain function:

$$g(n) \quad = \quad \begin{cases} 0 & \text{if } n < \lceil {}^{a}/_{x} \rceil \\ nx - a & \text{if } n = \lceil {}^{a}/_{x} \rceil \\ x - a & \text{if } n \in \left( \lceil {}^{a}/_{x} \rceil + 1, \lceil {}^{c-a}/_{1-x} \rceil - 1 \right] \\ n - (n-1)x - c & \text{if } n = \lceil {}^{c-a}/_{1-x} \rceil \\ 1 - c & \text{if } n > \lceil {}^{c-a}/_{1-x} \rceil \end{cases}. \tag{10}$$

The only two candidates for stable equilibrium are $\lceil {}^{a}/_{x} \rceil$ and $\lceil {}^{c-a}/_{1-x} \rceil$, the two sizes at which switching to a new technology first becomes profitable. At all other values of $n$ exceeding $\lceil {}^{a}/_{x} \rceil$, $g(\cdot) < 0$; values of $n < \lceil {}^{a}/_{x} \rceil$ are not externally stable because of our tie-breaking assumption. For $n \in \left( \lceil {}^{a}/_{x} \rceil + 1, \lceil {}^{c-a}/_{1-x} \rceil - 1 \right]$, gas technology is profitable even with $n - 1$

[5]From the equation defining $y(\alpha$, $\frac{dy(\alpha)}{d\alpha} = {}^{1}/_{1-x}$. Letting $C(\alpha) = a(1 - y(\alpha)) + by(\alpha)$ denote the total cost of abating $\alpha$, we have $\frac{dC(\alpha)}{d\alpha} = (c - a)\frac{dy(\alpha)}{d\alpha} = {}^{c-a}/_{1-x}$.

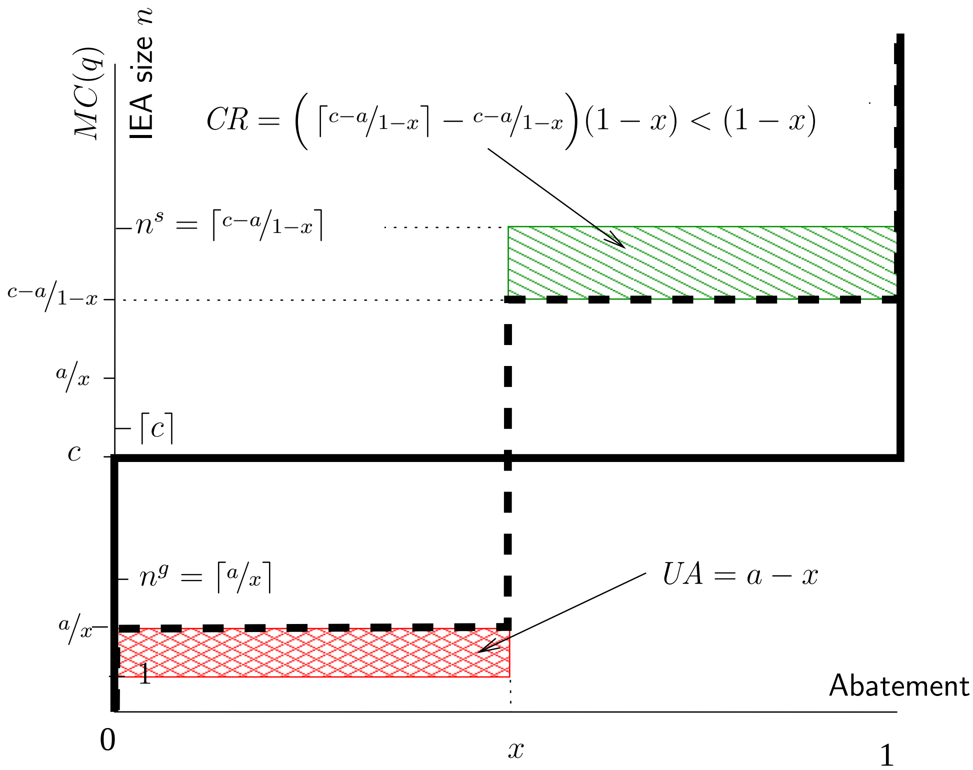


FIGURE 1. An example with piecewise constant marginal cost

members; hence the addition of an additional member would increase abatement by only $x$ units, while costing the joiner $a > x$ units. (Inequality (9) ensures that the interval $\left( \lceil a/x \rceil + 1, \lceil c-a/1-x \rceil - 1 \right]$ is non-empty.) Moreover, $g(\lceil a/x \rceil)$ is unambiguously positive, so an IEA with $\lceil a/x \rceil$ members is always a stable equilibrium. The sign of $g(\lceil c-a/1-x \rceil)$ is ambiguous; $n^s := \lceil c-a/1-x \rceil$ satisfies (4) iff

$$g\big(n^s\big) \quad = \quad n^s - \big(n^s - 1\big)x - c \quad \geq \quad 0 \tag{11}$$

An important characteristic of the post-innovation marginal cost function is that it is concave over the interval $[x, 1)$. Without this property, it would not be the case that for $n \geq 3$, $CR\,(n)$ could be arbitrarily large relative to $UA\,(n)$. The role that locally concavity plays in supporting large agreements is discussed heuristically in §3.3 and proved formally in Prop. 1.

We now decompose $g(n^s)$. Using the notation introduced on p. 6

$$q_{\rm f} = 0,\, q_{\rm s}^{n^s-1} = x \text{ and } q_{\rm s}^{n^s} = 1, \text{ while } C(q_{\rm s}^{n^s-1}) = a \text{ and } C(q_{\rm s}^{n^s}) = c \tag{12a}$$

From (6), the internal stability condition (11) can be written as $CR\,(n^s) - UA\,(n^s)$, where

$$UA\,(n^s) \;=\; a - x \tag{12b}$$

$$CR\,(n^s) \;=\; \left\lceil \frac{c-a}{1-x} \right\rceil (1-x) - (c-a) \;=\; \left( \left\lceil \frac{c-a}{1-x} \right\rceil - \frac{c-a}{1-x} \right) (1-x) \;<\; (1-x) \tag{12c}$$

Equations (12b) and (12c) imply that $g(n^s) < 1-a$, so a necessary condition for $g(\lceil {}^{c-a}/_{1-x} \rceil) \geq 0$ (stability of $n^s$) is that $a < 1$. If this propery holds—we assume it in (9)—there exists $x < 1 < {}^{a}/_{x} < c$ such that $g(\lceil {}^{c-a}/_{1-x} \rceil) \geq 0$ so that a solar IEA is stable. Whether or not this inequality is satisfied, however, depends in a fragile way on parameter values. To analyze this dependence, we fix $a \in (0,1)$ and define $\mathrm{CX}(a)$ to be the set of $(c, x)$ values which, paired with $a$, support a solar IEA.

$$\mathrm{CX}(a) = \{(c,x) \in \mathbb{R}^2 : x < 1 < {}^{a}/_{x} < c \text{ and } g(\lceil {}^{c-a}/_{1-x} \rceil) \geq 0\} \tag{13}$$

Define $\mathrm{AX}(c)$ and $\mathrm{AC}(x)$ analogously to $\mathrm{CX}(a)$. Heuristically, if $\mathrm{CX}(a) \subsetneq \mathrm{CX}(a')$, for all $0 < a' < a < 1$, then the likelihood of a stable solar IEA is strictly monotone decreasing in the average cost of gas generation. On the other hand, if there exists $c \neq c'$ such that $\mathrm{AX}(c)$ and $\mathrm{AX}(c')$ are *not nested*[6] then the likelihood of a stable solar IEA is not monotonically related to $c$.

**Remark 1.**

(a) *For all* $x < a' < a < 1$, *$CX(a) \subsetneq CX(a')$;*
(b) *For all* $1 < c'$, *there exists* $c \approx c'$ *s.t.* $AX(c')$ *and* $AX(c)$ *are not nested.*
(c) *For all* $0 < x' < 1$, *there exists* $x \approx x'$ *s.t.* $AC(x')$ *and* $AC(x)$ *are not nested.*

Part (a) is at first sight paradoxical. Its implication is that when the gas technology is less efficient it is *more* rather than less difficult to sustain a solar IEA. Upon reflection, however, the paradox is easily resolved: the difficulty of sustaining a solar IEA depends

[6] Sets $A$ and $B$ are said to be not nested if $A \not\subseteq B$ and $B \not\subseteq A$.

on the incremental benefit that a non-signatory obtains from participating in it, relative to free-riding on the gas IEA that would prevail if she did not participate. This incremental benefit is weakly *lower*, the *less* efficient is the gas technology. The reason is that the non-signatory does not care about the cost of abatement under the gas IEA, while the benefit she obtains from free-riding on the gas IEA weakly increases with its cost. To see this, recall first our observation on p. 10 that in the example presented in §3.1, global abatement increases weakly with abatement costs; the same is true for the gas IEA: the more expensive is gas, the more members are required in order to sustain a gas IEA, and hence the more abatement results from the agreement; true, the cost per IEA member is greater also, but as noted above, the non-signatory contemplating joining the solar IEA does not take this cost into account. Finally, the greater is the level of abatement under the gas IEA, the smaller is the *incremental* benefit of switching to a solar IEA. To summarize, the higher is the cost of abatement under the gas technology, the smaller is the incremental benefit to a non-signatory—the difference between total abatement under solar and total abatement under gas—of joining the solar IEA.

Why does not the same monotonicity argument above apply also to a decrease in $x$? Clearly, a decrease in $x$ also weakly increases the number of members—$n^g := \lceil a/x \rceil$—needed in order to sustain a gas IEA. Unlike an increase in $a$, however, a decrease in $x$ affects in a nonmonotonic way the benefit to a free-rider of the gas IEA. Indeed, as $x$ decreases holding other parameters constant, the average benefit to a non-signatory of a gas IEA declines so long as $\lceil a/x \rceil$ remains constant, but jumps up discontinuously once the integer threshold is crossed. To summarize, the impact of a small change in $x$ on the incremental benefit to a non-signatory of joining a solar IEA may be positive or negative, depending on initial conditions.

The parameter $c$ plays has a quite different impact than either $a$ or $x$ on the sustainability of a solar IEA. Just as the cost to signatories of abatement in the gas IEA increases monotonically in $a$, so the cost of a solar IEA increases monotonically in $c$. But a non-signatory considering joining such an agreement weighs the *net* benefit of solar against the *gross* benefit of gas. This tradeoff depend on $c$ in a non-monotonic way: from (11), the difference is $(\lceil {}^{c-a}/_{1-x} \rceil - {}^{c-x}/_{1-x})$, which cycles with $c$ in a "saw-tooth" pattern between ${}^{x-a}/_{1-x} < 0$ and ${}^{1-a}/_{1-x} > 0$.

The above example illustrates a number of key aspects of this paper. First, it complements the point we made on p. 2: when an abatement cost function encompasses multiple abatement technologies, there are good reasons to expect that marginal abatement costs will not be convex. Second, it demonstrates how easy it is, once this widely adopted but not well-justified convexity assumption is relaxed, to construct models in which technological innovations both globally reduce abatement costs and also increase equilibrium participation. Third, it highlights in a particularly dramatic way a point reiterated in our conclusion: whether or not high-participation IEAs are stable may depend sensitively on particular parameter values.

3.3. **Linear marginal costs.** We begin this subsection with a review of the model with linear marginal costs and benefits. Using a simple diagramatic argument, we confirm Barrett (1994)'s result that for this specification, the unique stable equilibrium is $n = 3$, regardless of the slope of the marginal cost function. A change in the slope of the marginal cost function changes the equilibrium level of abatement, leaving unchanged both the equilibrium number of signatories and the fraction of potential global welfare gains realized in the game. We then show that by introducing a kink in the marginal cost function, flattening it north-east of the kink so that marginal costs are now concave, we can support any integer $3 < n \leq N$ as a second stable equilibrium; $n = 3$ remains as an equilibrium. Thus we show in this second context that technological innovations that reduce costs *are* compatible with increased, indeed even 100%, participation. In Fig. 2, we again decompose the joiner's gain function $g(n)$ into two components, $CR(n)$ and $UA(n)$ (see (5)-(6c)). The cost function is $C(q) = \frac{q^2}{2M}$, for some $M \in \mathbb{R}$, so that $C'(q) = q/M$. The upward sloping curve represents both the marginal cost curve $MC'(q)$ and its inverse, $S(p)$. Recall that $q_s^n$ solves $n = C'(q)$, while $q_f$ solves $1 = C'(q)$. From (6b), $UA(n)$ is the area under the marginal cost curve between $q_f$ and $q_s^{n-1}$, minus the area of the rectangle with boundaries $q_f$ and $q_s^{n-1}$ and height 1. The difference is the area of the cross-hatched triangle labeled $UA(n)$. From (6c), $CR(n)$ is the area of the rectangle with boundaries $q_f$ and $q_s^{n-1}$ and height $n$ minus the area under the marginal cost curve between $q_s^{n-1}$ and $q_s^n$. The difference is the area of the hatched triangle labeled $CR(n)$. This second triangle has unit height, regardless of $n$; the height of the first triangle is $n-2$. Because the two triangles are similar, the area $CR(n) \geq UA(n)$ iff $n \leq 3$. The equilibrium condition (4) is that $CR(n) \geq UA(n)$ while $CR(n+1) < UA(n+1)$. We have thus established

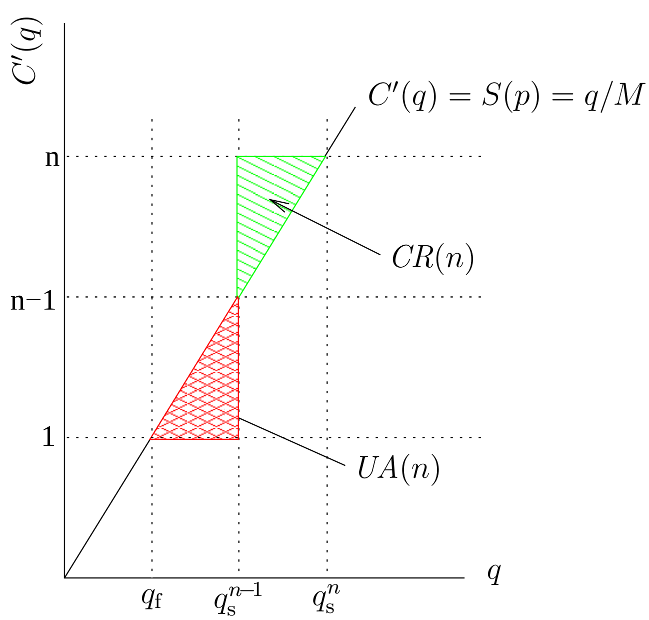


FIGURE 2. Linear benefits and linear marginal costs.

**Remark 2.** *When marginal benefit and marginal cost are both linear, the unique stable equilibium size is $n = 3$.*

Moreover:

**Remark 3.** *When three countries form an IEA, the equilibrium global gain is $6M(N-2)$; this is a fraction $\frac{12(N-2)}{N(N^2-2N+1)}$ of the total potential gain that could be achieved if all $N$ countries joined the agreement. This fraction is independent of the cost parameter $M$ and decreases with $N$ for $N > 3$.*

Although an increase in $M$ lowers total and marginal costs and increases global welfare, technology enhancements that reduce $M$ have no effect on the fraction of the potential welfare gain that is achieved in equilibrium. In §4, Prop. 1 establishes that an equilibrium size exceeding $n = 3$ cannot be sustained if the *marginal* cost function is convex, i.e., if $C'''(\cdot) \geq 0$. Membership can, however, exceed $n = 3$ if this restriction is relaxed, allowing the marginal cost function to be locally concave beyond $q = q_s^{\bar{n}-1}$. (We reiterate that conventional economic theory requires only that the *cost* function is convex; no restrictions are imposed on marginal costs except that they be non-decreasing.) Indeed, we have

**Remark 4.** *Assume that initially, marginal benefit and marginal cost are both linear. For any $3 < \bar{n} \leq N$, a stable equilibrium of size $\bar{n}$ can be implemented by a cost-reducing innovation resulting in a piece-wise linear marginal cost curve.*

Fig. 3 illustrates Remark 4. In the figure, the modified cost function is $\tilde{C}(\cdot)$, whose derivative is piecewise linear and agrees with $C'(\cdot)$ for $q \leq q_s^{\bar{n}-1} = (\bar{n}-1)M$, but is flatter than $C'(\cdot)$ for larger values of $q$. Specifically, let

$$\tilde{C}'(q) \quad = \quad \begin{cases} q/M & \text{if } q \leq (\bar{n}-1)M \\ c_0 + q/K & \text{if } q > (\bar{n}-1) \end{cases},$$

where $K > M$ and $c_0 = (\bar{n}-1)(1-M/K)$. ($c_0$ is chosen so that $\tilde{C}'(\cdot)$ is continuous at $q = q_s^{\bar{n}-1} = (\bar{n}-1)M$.) The new marginal cost curve $\tilde{C}'(\cdot)$ is depicted in Fig. 3 by a heavy solid line; the original one by a dashed line. Note that the area of $CR(\bar{n})$ just exceeds the area of $UA(\bar{n})$: the shorter height of the first triangle is now more than offset by its longer base. The cost reduction from $C(\cdot)$ to $\tilde{C}(\cdot)$ leaves individual abatement unchanged at $M$ if no IEA is implemented; in this case the reduction has no effect on global welfare. The cost

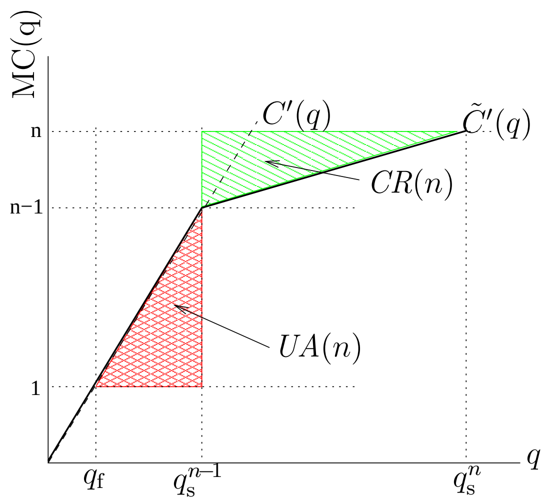


FIGURE 3. Linear benefits and kinked marginal costs.

reduction can however increase participation in an IEA. Indeed, if the reduction is sufficiently great—specifically, if $K \geq M(N-2)^2$ (see the proof of Remark 4)—an agreement to which all $N$ countries are signatories is stable, and all of the potential gains from cooperation under this cost structure are realized.

## 4. Convex costs and linear benefits

In this section we generalize the model to incorporate arbitrary convex cost functions, while maintaining the restriction that benefits are linear in abatement. We begin by showing that if the *marginal* cost function is initially strictly convex in abatement, then an IEA with more than two members is unsustainable. Once this restriction is relaxed (maintaining convexity of the *cost* function), an IEA of any size can be sustained. Our approach is to fix $n \leq N$ and a given strictly convex cost function for which $n$ is not a stable equilibrium, then lower this function to the point at which a stable equilibrium of size *at least* $n$ exists. When benefits are linear in abatement, there are countless ways of attaining this goal: we identify the "smallest possible" reduction in the cost function that can accomplish it. This result is formalized in Prop. 2. Our motivation for this exercise is that cost-reducing technological progress are more difficult to achieve, the greater is the extent to which costs must be reduced. Our "smallest possible" reduction can be interpreted as the least resource-intensive way to implement a coalition with at least $n$ members.

The cost reduction we construct on p. 28 below (see (28) and (29)) implements a coalition with exactly $n$ members if and only if $g(n+1) \leq g(n) \leq 0$ under the initial cost function. If instead $g(n+1) \geq g(n)$, this reduced cost function will satisfy only the internal, but not the external stability condition at $n$ (see (4)). In this second case, however, both internal and external stability will "eventually" be satisfied:

$$\text{if } n \text{ is internally stable, there exists } n' \geq n \text{ such that } n' \text{ is a stable equilibrium.} \quad (14)$$

To confirm this claim we consider the two cases, first where $n$ is externally stable and then when it is not. If $n$ is externally stable, then by definition $n$ is stable. If $n$ is not externally stable, let $\underline{\mathcal{N}} = \{n' \in (n+1, N] \cap \mathbb{N} : \tilde{g}(n') < 0\}$, where $\tilde{g}(\cdot)$ is the joiner's gain function corresponding to the reduced cost function. If $\underline{\mathcal{N}}$ is non-empty, let $\underline{n}$ be its smallest element. In this case, because $\tilde{g}(\underline{n} - 1) > 0 > \tilde{g}(\underline{n})$, $n' = \underline{n} - 1$ is both internally and externally stable. On the other hand, if $\underline{\mathcal{N}}$ is empty, then $\tilde{g}(N) \geq 0$ and $N$ will be both internally and externally stable, because the external stability condition is satisfied vacuously at $N$. A straightforward extension of the methodology we develop in this section shows that it is possible to implement a coalition with exactly $n$ members even when $g(n+1) > g(n)$.

For the rest of this section, we fix a differentiable, strictly convex cost function $C$, and let $S(\cdot)$ denote the inverse of $\mathrm{MC}(\cdot)$. When the benefit function is linear, the functions $UA(n)$ and $CR(n)$ defined above ((6b) and (6c)) can be rewritten in a particularly convenient form:

$$UA\,(n) \quad = \quad C\left(q_{\rm s}^{n-1}\right) - C\left(q_{\rm f}\right) - \left(q_{\rm s}^{n-1} - q_{\rm f}\right) \quad = \quad (n-2)\,q_{\rm s}^{n-1} \quad - \quad \int_1^{n-1} S(p)dp \qquad (15a)$$

$$CR\,(n) \quad = \quad n\left(q_{\rm s}^{n} - q_{\rm s}^{n-1}\right) - \left[C\left(q_{\rm s}^{n}\right) - C\left(q_{\rm s}^{n-1}\right)\right] \quad = \quad \int_{n-1}^{n} S(p)dp \quad - \quad q_{\rm s}^{n-1}. \qquad (15b)$$

4.1. **A pessimistic result when marginal cost is convex.** Figure 2 provides geometric intuition for why the formulations in (6) and (15) are equivalent provided benefits are linear. In our discussion of the figure on p. 17, the cross-hatched area $UA\,(n)$ was identified as the area under the marginal cost curve between $q_{\rm f}$ and $q_{\rm s}^{n-1}$, minus the area of the rectangle with vertical boundaries $q_{\rm f}$ and $q_{\rm s}^{n-1}$ and height 1. As Fig 2 illustrates, this area is also equal to the area of the rectangle bounded vertically by 0 and $q_{\rm s}^{n-1}$, and horizontally by 1 and $n-1$, minus the trapezoid to the left of $S(p)$ between 1 and $n-1$. The area of the difference between the trapezoid and the rectangle is given by the right-hand side of (15a). The hatched area $CR\,(n)$ was identified on p. 17 as equal to the area of the rectangle with vertical boundaries $q_{\rm s}^{n-1}$ and $q_{\rm s}^{n}$ and height $n$ minus the area under the marginal cost curve between $q_{\rm s}^{n-1}$ and $q_{\rm s}^{n}$. It is also equal to the trapezoid to the left of the curve $S(p)$ between $n-1$ and $n$, minus the area of the rectangle between 0 and $q_{\rm s}^{n-1}$ with height one. The area of the difference between this second rectangle and smaller trapezoid is equal to the right-hand side of (15b).

We use (15) to simplify expression (5) for the joiner's gain function $g(n) = CR\,(n) - UA\,(n)$:

$$\begin{aligned} g(n) \quad &= \quad \int_{n-1}^{n} S(p)dp - q_{in}^{n-1} \quad - \quad \left[(n-1-1)\,q_{in}^{n-1} - \int_1^{n-1} S(p)dp\right] \\ &= \quad \int_1^{n} S(p)dp \quad - \quad (n-1)\,q_{in}^{n-1} \quad = \quad \int_1^{n} S(p)dp \quad - \quad (n-1)\,S\,(n-1)\,. \qquad (16) \end{aligned}$$

Using expression (16), Prop. 1 establishes that when marginal costs are strictly convex, there cannot exist a stable equilibrium with more than two signatories.[7]

**Proposition 1.** *If the marginal cost function is strictly convex, so that $S(p)$ is strictly concave, then the largest stable equilibrium is less than or equal to 2.*

[7]Using a convex marginal cost function which had previously been used to estimate the costs of greenhouse gas abatement, Barrett (1994)'s Proposition 4 shows that the equilibrium IEA consists of 2 members.

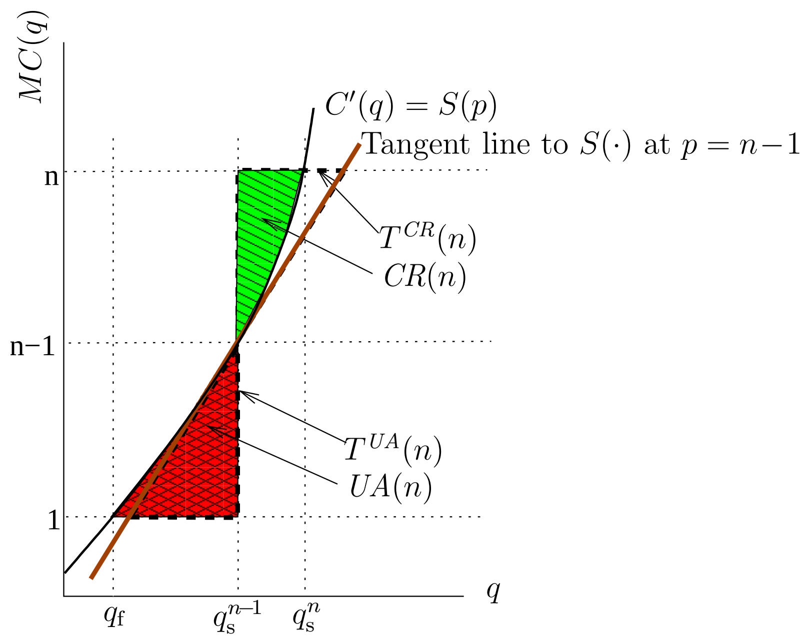


FIGURE 4. Intuition for Prop. 1

Fig. 4 provides intuition for Prop. 1. Because $MC'(\cdot)$ is strictly convex (i.e., $S(\cdot)$ is strictly concave), its graph lies everywhere except at $p = n{-}1$ strictly to the left of the line that is tangent to $S(\cdot)$ at $p = n{-}1$. Therefore the hatched region $CR(n)$ is strictly contained in the triangle $T^{CR}(n)$, indicated by dashed lines, which is bounded below by $n - 1$ and above by $n$, and so has height 1. Because $T^{CR}(n)$ is similar to the triangle $T^{UA}(n)$, marked by dotted lines, with height $n - 2$, the area of $T^{UA}(n)$ weakly exceeds that of $T^{CR}(n)$ for all $n \geq 3$. Moreover, the cross-hatched region $UA(n)$ strictly contains $T^{UA}(n)$. It follows that for all $n \geq 3$, $UA(n) > CR(n)$, verifying that $g(n) < 0$, for every integer $n > 2$. Prop. 1 now follows from (4).

The intuition underlying Prop. 1 can be extended to provide insight into the relationship between the internal and external stability conditions. Say that a function $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is *convex beyond* $x$ if $f(\cdot)$ is convex on the interval $[x, \infty]$. It will be clear by inspection of Fig. 4 that if for some $\bar{n}$, the marginal cost curve is convex beyond $\mathrm{MC}^{-1}(\bar{n})$, then as $n$ increases beyond $\bar{n}$, the breadth of $CR(n)$ at its base will either remain constant or decrease.

Its height, however, will remain constant at 1. On the other hand, both the height and breadth of $UA(n)$ will increase with $n$. It follows immediately that

$$\text{for any } \bar{n} \in \mathbb{N}, \text{ if marginal cost is convex beyond } \mathrm{MC}^{-1}(\bar{n}), \text{ then the joiner's gain function } g(\cdot) \text{ is strictly decreasing beyond } \bar{n}. \tag{17}$$

An immediate consequence of (17) is that if marginal cost is convex beyond $\mathrm{MC}^{-1}(\bar{n})$, then the external stability requirement is satisfied beyond $\bar{n}$ whenever the internal stability requirement is satisfied (i.e., for $n \geq \bar{n}$, $g(n) \leq 0$ implies $g(n+1) < 0$). Summarizing

$$\text{if MC is convex beyond } \mathrm{MC}^{-1}(\bar{n}), \text{ then } n \geq \bar{n} \text{ is stable if } g(n) = 0. \tag{18}$$

In Prop. 2 below, we reduce a given initial cost function in a way that leaves the original cost function unchanged beyond $\bar{n}$, while the joiner's modified gain function is zero at $\bar{n}$. It follows from (18) that if the marginal cost were originally convex, then our construction is sufficient to ensure that $\bar{n}$ is stable. This fact is significant because in the literature on participation games to date, costs are almost invariably parameterized so that marginal costs are convex. Indeed, we are unaware of *any* paper in the related literature in which this property is violated.

4.2. **Admissible cost reductions.** We now consider the effect on the equilibrium size of an IEA of a technological improvement that lowers the abatement cost function. We model this kind of improvement by adding to an initially specified inverse marginal cost function $S(\cdot)$ a non-negative function that is positive on an open subset of its domain. Observe that a (weak) increase in $S(\cdot)$, for all $p$, implies a (weak) decrease in marginal costs, for all $q$.

Let $\varepsilon(p) \geq 0$ denote a *cost reduction function.* The new (lower) inverse marginal cost function is denoted by $\tilde{S}(\cdot) = S(\cdot) + \varepsilon(\cdot)$. In an $n$-member IEA, the per-member equilibrium level of abatement is $S(n)$ before the cost reduction, and $\tilde{S}(n)$ after it. Thus, $\varepsilon(n)$ is the amount by which the cost reduction increases per-member abatement for an $n$-member IEA. Let $\tilde{g}(\cdot)$ denote the joiner's gain function under the modified cost structure $\tilde{S}(\cdot)$.

We say that a cost reduction function $\varepsilon(\cdot)$ is *admissible* if it satisfies certain properties. To streamline the exposition, we allow $\varepsilon(\cdot)$ to be discontinuous, in which case there will be $p$-values at which the optimal abatement choice is not uniquely defined.[8] To resolve indeterminacies, we assume:

If multiple abatement levels satisfy the first order condition (2)

each non-signatory chooses the *lowest* optimal abatement level (19a)

each signatory chooses the *highest* optimal abatement level. (19b)

From the exposition which follows, it will be clear that we could have restricted $\varepsilon(\cdot)$ to be continuous, and thus eliminated all indeterminacies. This would have required more notation, increased the complexity of proofs, and made our heuristic explanations more cumbersome, without changing any of our formal results. We now define the conditions for admissibility:

**Definition 1.** *For given $S(\cdot)$, $\varepsilon(p)$ is an admissible cost reduction function if*

$$\varepsilon(p) \geq 0, \text{ with strict inequality on some open subset of } \mathbb{R} \tag{20a}$$

$$S(\cdot) + \varepsilon(\cdot) \text{ is a non-decreasing function of } p. \tag{20b}$$

Inequality (20a) implies that for any given level of membership, the cost reduction never decreases abatement, and strictly increases it for some levels; inequality (20b) states that marginal cost after the reduction is nondecreasing, so that costs remain convex after the cost reduction. For convenience, we add a third restriction:

$$\varepsilon(p) = 0 \text{ for all } p \in (N+1, \infty). \tag{20c}$$

Because the largest possible IEA has $N$ members, $q_s^n$ never exceeds $S(N)$. Hence marginal

[8] A function $\varepsilon(p)$ such that for some $p$, $\lim_{\delta\searrow 0}\varepsilon(p-\delta) = \underline{q} < \bar{q} = \lim_{\delta\searrow 0}\varepsilon(p+\delta)$ is consistent with a cost function $C(q)$ which is affine on the interval $[\underline{q}, \bar{q}]$, i.e., marginal cost is flat on this interval. In this case, all of the $q$'s in $[\underline{q}, \bar{q}]$ will be equi-profitable. Strictly, speaking, $S(p) + \varepsilon(p)$ is not the inverse of any marginal cost function, since viewed as a function of $q$ rather than $p$, $S(q) + \varepsilon(q)$ is not one-to-one.

cost reductions at levels of $q > S(N)$ can have no impact on equilibrium outcomes, so restriction (20c) is without loss of generality. Replacing $S(\cdot)$ with $\tilde{S}(\cdot)$ in (16), we see that the change in the joiner's gain function at $n$ due to the cost reduction $\varepsilon(\cdot)$ is

$$\Delta^{\varepsilon} g\,(n) \quad = \quad \tilde{g}(n) - g(n) \quad = \quad \int_1^n \varepsilon(p)dp - (n-1)\,\varepsilon\,(n-1)\,. \tag{21a}$$

Similarly, the change in $g(\cdot)$ at $n{+}1$ is

$$\Delta^{\varepsilon} g\,(n{+}1) \quad = \quad \int_1^{n+1} \varepsilon(p)dp - n\varepsilon\,(n) \tag{21b}$$

$$= \quad \Delta^{\varepsilon} g(n) \quad - \quad n\varepsilon(n) \quad + \quad (n-1)\varepsilon(n-1) \quad + \quad \int_n^{n+1} \varepsilon\,(p)\,dp. \tag{21c}$$

For any $3 \leq n \leq N$, we identify in §4.3 the "smallest possible" or "most efficient" reduction in abatement costs such that an IEA with at least $n$ members is stable. Because we consider cost reductions over an interval, there are many alternative definitions of "smallest possible." We adopt a natural specification, evaluating $\varepsilon(\cdot)$'s that satisfy conditions (20) according to the following norm:[9]

$$||\varepsilon|| \quad = \quad \int_0^{N+1} \varepsilon\,(p)\,dp; \tag{22}$$

that is, we take the integral over the change in inverse marginal costs from 0 to $N+1$. For $\varepsilon(\cdot)$'s satisfying restriction (20c), this norm is equivalent to one that computes the integral of marginal costs (cf. inverse marginal costs) over an interval of $q$'s containing the set on which the new and the old marginal cost curves (cf. the inverse marginal cost curves) are permitted to differ.

[9]Our conditions (20) for admissibility confine us to considering only cost reductions that weakly reduce marginal costs globally. This is a very strong restriction. For example, the cost reduction we consider in §3.2 involves an increase in marginal costs over some region. If we had imposed a weaker admissibility criterion, which admitted the possibility that cost reduction could locally increase marginal costs, thus relaxing a self-imposed constraint, we could have supported a given membership level with a cost reduction that was much smaller under our metric.

4.3. **Cost decreases that increase equilibrium membership.** Fix $\bar{n} > 2$ and a marginal cost function whose inverse is $S$. The integer $\bar{n}$ may be unstable under $S$ because it fails the test either for internal stability—i.e., $g(\bar{n}) < 0$—or for external stability—i.e., $g(\bar{n}+1) \geq 0$. The pessimistic results in the literature to date are all consequences of the fact that for the functional forms which that literature considers, it is the internal rather than the external condition that is violated for large $n$. Accordingly, to limit the number of cases we need to consider we assume that under the original cost structure, the internal stability condition fails at $\bar{n}$, i.e., $g(\bar{n})$ is negative.

Our goal, then, is to shift $g(\cdot)$ up at $\bar{n}$. From (5), $g(\cdot)|_{\bar{n}}$ decreases with $q_s^{\bar{n}-1}$; moreover, $q_s^{\bar{n}-1}$ increases as $\mathrm{MC}(\cdot)|_{q_s^{\bar{n}-1}}$ shifts down, so that a cost modification which lowers $\mathrm{MC}(\cdot)$ on a neighborhood of $q_s^{\bar{n}-1}$ will, holding all else constant, shift $g(\cdot)|_{\bar{n}}$ even further down. On the other hand, equation (5) also reveals that $g(\cdot)$ shifts up at $\bar{n}$ as the cost of producing $q_s^{\bar{n}}$ decreases. Accordingly, the cost modification we construct below leaves $\mathrm{MC}(\cdot)$ unchanged at $q_s^{\bar{n}-1}$ itself, while lowering it on an interval to the left of $q_s^{\bar{n}-1}$, thus lowering the cost of producing $q_s^{\bar{n}}$, while $q_s^{\bar{n}-1}$ remains constant. As Fig 4 illustrates and Prop. 1 formalizes, this reduction must be sufficiently large that the modified marginal cost curve must on some interval to the left of $q_s^{\bar{n}-1}$ lie below the tangent line to $\mathrm{MC}(\cdot)$ at $q_s^{\bar{n}-1}$—in other words the modified marginal cost curve must be locally concave on some interval—otherwise $g(\cdot)$ will necessarily be negative for all $n \geq 3$. (As we note on p. 29, we could also increase $g(\cdot)$ at $\bar{n}$ to some extent by lowering marginal cost between $q_s^{\bar{n}-1}$ and $q_s^{\bar{n}}$; there is, however, no efficiency gain to this alternative approach.)

Before proceeding with our construction, we note some key inequalities and introduce some terminology. Given a cost reduction function $\varepsilon(\cdot)$, a necessary and sufficient condition for $\bar{n}$

to be stable given the reduced cost structure $\tilde{S}(\cdot) = S(\cdot) + \varepsilon(\cdot)$ is that

$$\tilde{g}(\bar{n}) \quad = \quad g\,(\bar{n}) + \Delta^{\varepsilon} g\,(\bar{n}) \quad \geq \quad 0 \quad > \quad g(\bar{n}{+}1) + \Delta^{\varepsilon} g\,(\bar{n}{+}1) \quad = \quad \tilde{g}(\bar{n}{+}1). \quad (23)$$

It follows from (23) that $\bar{n}$ will be stable given a cost reduction $\varepsilon(\cdot)$ iff

$$\Delta^{\varepsilon} g\,(\bar{n}) \quad \geq \quad -g\,(\bar{n}) \quad \geq \quad 0 \qquad (24a)$$

$$g(\bar{n}{+}1) + \Delta^{\varepsilon} g\,(\bar{n}{+}1) \quad < \quad 0. \qquad (24b)$$

We established at the beginning of §4 that if (24a) is satisfied at $\bar{n}$, then there exists $n' \geq \bar{n}$ such that $n'$ is stable. Accordingly we say that

$\varepsilon(\cdot)$ *implements at least* $\bar{n}$ if $\varepsilon(\cdot)$ satisfies (20) and if $\Delta^{\varepsilon} g(\bar{n})$ satisfies (24a). (25a)

$\varepsilon(\cdot)$ *implements exactly* $\bar{n}$ if $\varepsilon(\cdot)$ satisfies both (24a) and (24b) (25b)

That is, to implement at least $\bar{n}$, it is necessary only to satisfy internal stability at $\bar{n}$; to implement exactly $\bar{n}$, external stability must be satisfied as well. The cost reduction we construct below satisfies internal stability with equality, i.e., sets $\Delta^{\varepsilon} g\,(\bar{n}) = -g\,(\bar{n})$. As noted on p. 22 (see (18)), if the initial cost function belongs to a class that has been considered in the related literature to date and internal stability is satisfied with equality, then external stability is satisfied as well. For such cost functions, the distinction between "at least" and "exactly" is moot.

There are countless admissible cost reductions that implement at least $\bar{n}$. We identify the minimal value of $||\varepsilon||$, taken over the set of all such reductions. When satisfying internal stability with equality does *not* imply external stability, there will not exist a minimal value of $||\varepsilon||$ that implements exactly $\bar{n}$. Accordingly, we define the:

minimal cost reduction for at least $\bar{n}$ as $\epsilon^{\bar{n}}_{\text{min}} = \min \left\{ ||\varepsilon|| : \varepsilon(\cdot) \text{ implements at least } \bar{n} \right\}$. (26a)

infimal cost reduction for exactly $\bar{n}$ as $\epsilon^{\bar{n}}_{\text{inf}} = \inf \left\{ ||\varepsilon|| : \varepsilon(\cdot) \text{ implements exactly } \bar{n} \right\}$. (26b)

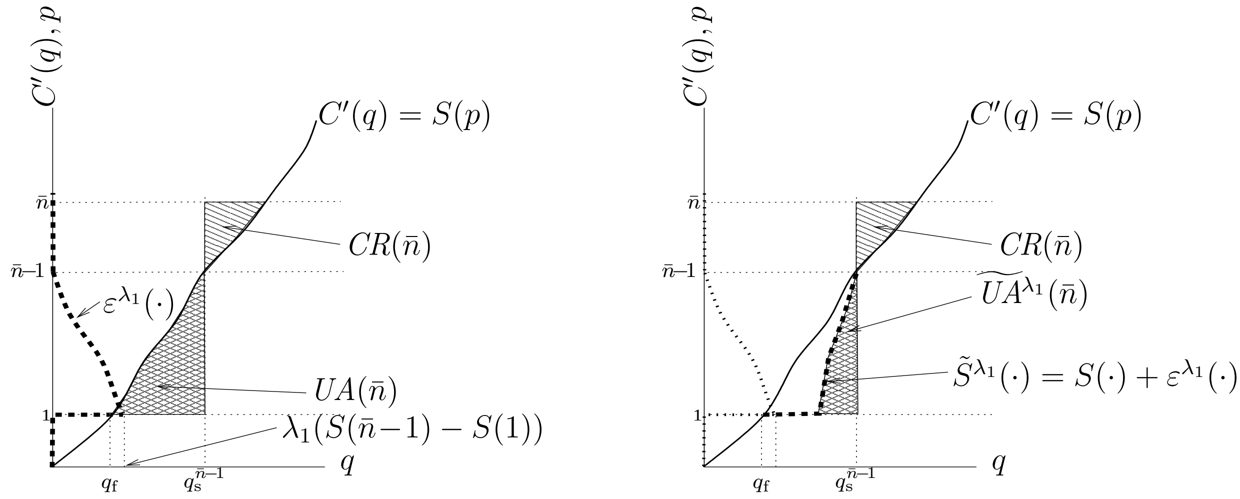


FIGURE 5. Shrinking $UA(\bar{n})$ increases $\tilde{g}(\bar{n})$.

Finally, say that

$$\varepsilon(\cdot) \text{ is } \textit{efficient for at least } \bar{n} \text{ if } \varepsilon(\cdot) \text{ implements at least } \bar{n} \text{ and } ||\varepsilon|| = \epsilon_{\text{min}}^{\bar{n}}. \tag{27a}$$

$$\varepsilon(\cdot) \text{ is } \textit{efficient for exactly } \bar{n} \text{ if } \varepsilon(\cdot) \text{ implements exactly } \bar{n} \text{ and } ||\varepsilon|| = \epsilon_{\text{inf}}^{\bar{n}}. \tag{27b}$$

A cost reduction that is efficient for exactly $\bar{n}$ exists only if satisfying internal stability with equality implies external stability.

The construction below identifies an $\varepsilon(\cdot)$ that is efficient for at least $\bar{n}$. First, for each $\lambda_1 \in [0, 1)$, define

$$\varepsilon^{\lambda_1}(p) \quad = \quad \begin{cases} \lambda_1(S(\bar{n}-1) - S(p)) & \text{if } p \in [1, \bar{n}-1) \\ 0 & \text{otherwise} \end{cases}. \tag{28}$$

For $\lambda_1 \approx 2/3$, the graph of $\varepsilon^{\lambda_1}(\cdot)$ is depicted in the left panel of Fig. 5 by heavy dashed lines. It is vertical until $p = 1$, at which point it jumps horizontally, then slopes negatively until $p = (\bar{n}-1)$, beyond which it is again vertical. In the right panel of the figure, $\varepsilon^{\lambda_1}(\cdot)$ is added to $S(\cdot)$ to form $\tilde{S}^{\lambda_1}(\cdot)$, which by construction agrees with $S(\cdot)$ on $[0, 1) \cup [\bar{n}-1, \infty)$, is discontinuous at $p = 1$, then rises more steeply than $S(\cdot)$ over the interval $(1, \bar{n}-1]$. Note that as $\lambda_1$ approaches 1, $\lim_{\delta \searrow 0} \tilde{S}^{\lambda_1}(1+\delta)$ approaches $q_s^{\bar{n}-1}$, while $S^{\lambda_1}(\cdot)$ rises increasingly steeply on $(1, \bar{n}-1]$. In the left panel, $UA(\bar{n})$ is larger than $CR(\bar{n})$, so that $g(\bar{n}) \equiv CR(\bar{n}) - UA(\bar{n}) < 0$.

In the right panel, the modified region $\widetilde{UA}^{\lambda_1}(\bar{n})$ (the region above $p = 1$, below $p = \bar{n}-1$, right of $\tilde{S}^{\lambda_1}(\cdot)$ and left of $q_s^{\bar{n}-1}$) is considerably smaller. The location of $q_f$ remains unchanged, in accordance with our assumption (19a) about how indeterminacies are resolved. It will be clear from the figure that when $\lambda_1 = 0$, $\tilde{S}(\cdot)$ is identical to $S(\cdot)$, so that the area of $\widetilde{UA}^{\lambda_1}(\bar{n})$ exceeds that of $CR(\bar{n})$, while as $\lambda_1$ approaches 1, the area of $\widetilde{UA}^{\lambda_1}(\bar{n})$ shrinks to zero and is hence eventually smaller than that of $CR(\bar{n})$. Since the area $\widetilde{UA}^{\lambda_1}(\bar{n})$ changes continuously with $\lambda_1$ it follows from the intermediate value theorem that

$$\text{there exists } \lambda_1^* \in (0,1) \text{ s.t. the areas of } \widetilde{UA}^{\lambda_1^*}(\bar{n}) \text{ and } CR(\bar{n}) \text{ are equal.} \tag{29}$$

Hence for the function $\tilde{S}^{\lambda_1^*}$, $\tilde{g}^{\lambda_1^*}(\bar{n}) = 0$. From (21a), $\Delta^{\lambda_1^*} g\,(\bar{n}) = \tilde{g}^{\lambda_1^*}(\bar{n}) - g(\bar{n})$, so that $\Delta^{\lambda_1^*} g\,(\bar{n}) = -g(\bar{n})$. Moreover, by construction, $\varepsilon^{\lambda_1^*}(\bar{n}-1) = 0$. Hence from (21a) and (22), $||\varepsilon^{\lambda_1^*}|| = -g(\bar{n})$, so that $\epsilon_{\text{min}}^{\bar{n}}$, the minimal cost reduction for at least $\bar{n}$ is at most $-g(\bar{n})$. Part 1 of Prop. 2 establishes that in fact $\epsilon_{\text{min}}^{\bar{n}} = -g(\bar{n})$.

A property of the construction (28) is that $\varepsilon^{\lambda_1}(p) = 0$, for all $p \geq \bar{n}-1$. Hence from (21c), $\Delta^{\lambda_1^*} g(\bar{n}) = \Delta^{\lambda_1^*} g(\bar{n}+1)$. Since from (23), $\tilde{g}^{\lambda_1^*}(\bar{n}+1) = g(\bar{n}+1) + \Delta^{\lambda_1^*} g(\bar{n}+1)$ while $\tilde{g}^{\lambda_1^*}(\bar{n}) = 0 = g(\bar{n}) + \Delta^{\lambda_1^*} g(\bar{n})$, we have that

$$\tilde{g}^{\lambda_1^*}(\bar{n}+1) \quad = \quad g(\bar{n}+1) \quad - \quad g(\bar{n}). \tag{30}$$

It now follows from (24b) and (30) that if $g(\bar{n}+1) < g(\bar{n})$, then $\bar{n}$ is stable under $\varepsilon^{\lambda_1^*}$.

The construction above is by no means unique: if the heavy dashed line that bounds $\widetilde{UA}^{\lambda_1}(\bar{n})$ on the left were replaced by *any* other non-decreasing line such that the area of the corresponding region $UA(\bar{n})$ were equal to that of $\widetilde{UA}^{\lambda_1}(\bar{n})$, then the cost reduction function associated with that line would also be efficient for at least $\bar{n}$. There is also a second class of $\varepsilon(\cdot)$'s that have this property: functions in this class are positive on $(\bar{n}-1, \bar{n})$ and thus increase the area of $CR(\bar{n})$ rather than simply reducing the area of $UA(\bar{n})$. There is, however, a limit in general on how much can be accomplished by increasing $CR(\bar{n})$ without also reducing $UA(\bar{n})$; for example, if $S(\cdot)$ were nearly affine on $(\bar{n}-1, \bar{n})$, then one could not do much better than double the area of $CR(\bar{n})$, which, if the initial difference between $UA(\bar{n})$ and $CR(\bar{n})$ were sufficiently large, would be alone insufficient to equalize the two areas. By contrast, regardless of the initial relative sizes of $UA(\bar{n})$ and $CR(\bar{n})$, the area of $UA(\bar{n})$ can always be reduce sufficiently to equalize them.

Prop 2 below summarizes and extends the preceding graphical analysis. The extensions are proved in the appendix. The result focuses exclusively on conditions under which the *internal* stability condition (24a) is satisfied. As we observed on p. 19 (statement (14)), if the internal stability condition is satisfied at $\bar{n}$, then both the internal and external stability conditions will be satisfied for some $n' \geq \bar{n}$. For this reason, the first three parts of Prop 2 relate to properties that hold for "at least $\bar{n}$." The fourth part notes that when the external stability constraint is slack—i.e., when $g(\bar{n}+1) < g(\bar{n})$—then a property will hold for at least $\bar{n}$ if and only if it holds for exactly $\bar{n}$.

**Proposition 2.** *Let $C$ be a twice differentiable, strictly convex cost function and let $B$ be a linear benefit function. For $1 < \bar{n} \leq N$, assume that $g(\bar{n}) < 0$ so that $\bar{n}$ is not stable. Then*

(1) *the minimal cost reduction for at least $\bar{n}$ is $-g(\bar{n})$.*
(2) $\exists \lambda_1^* \in (0,1)$ *s.t. the cost reduction $\varepsilon^{\lambda_1^*}$ defined by (28) is efficient for at least $\bar{n}$.*
(3) $\varepsilon(\cdot)$ *is efficient for at least $\bar{n}$ iff $\int_1^{\bar{n}} \varepsilon(p)dp = -g(\bar{n})$ and $\varepsilon\Big([0,1) \cup \{\bar{n}-1\} \cup [\bar{n}, \infty]\Big) = 0$.*
(4) *if $g(\bar{n}+1) < g(\bar{n})$ then $\varepsilon(\cdot)$ is efficient for at least $\bar{n}$ iff $\varepsilon(\cdot)$ is efficient for exactly $\bar{n}$.*

An immediate corollary of part (1) of Prop. 2 is the intuitive result that if $g(\cdot)$ is monotone decreasing in $n$ over an interval, then larger cost reductions are required in order to implement larger coalitions in that interval. Note from part (3) that efficient cost reductions necessarily leave unchanged the original cost function beyond $\bar{n}$. Thus, if the initial *marginal* cost function is convex, then from (17) and part (4), $\varepsilon(\cdot)$ is efficient for exactly $\bar{n}$ iff it is efficient for at least $\bar{n}$. As we have emphasized above, marginal costs are indeed convex for the functions that have been considered in the literature on participation games to date, so that part (4) applies to these functions.

Now suppose that $g(\bar{n}+1) \geq g(\bar{n})$. As we noted on p. 19, cost reductions that are efficient for at least $\bar{n}$ will now violate the external stability constraint. It is possible to implement exactly $\bar{n}$ in this case, but costs must be reduced still further. The required construction is similar to our first reduction (28), although the analysis is less "clean" in several respects. First, we were able to satisfy the internal stability condition at $\bar{n}$ while holding constant the abatement levels $q_{\mathrm{f}}$, $q_{\mathrm{s}}^{\bar{n}-1}$ and $q_{\mathrm{s}}^{\bar{n}}$; to satisfy external stability, on the other hand, it is now necessary to increase $q_{\mathrm{s}}^{\bar{n}}$. (We need a cost reduction $\varepsilon(\cdot)$ s.t. $\Delta^{\varepsilon} g(\bar{n}+1) < 0$. From (21b), it is now necessary that $\varepsilon(\bar{n}) > 0$, since $\Delta^{\varepsilon} g(\bar{n}+1)$ increases with $\varepsilon(p)$ for all other values of $p$.) Second, because of the strict inequality in the external stability condition (24b), a minimal cost reduction no longer exists. Third, while it is possible to identify an infimal cost reduction, a closed-form expression for this infimum does not exist. For completeness, Prop 3 summarizes most of what can be said when $g(\bar{n}+1) \geq g(\bar{n})$ under the original cost function. The proof, which is rather technical, is included in the online addendum.

**Proposition 3.** *Let $C$ be a twice differentiable, strictly convex cost function and let $B$ be a linear benefit function. For $1 < \bar{n} \leq N$, assume that $g(\bar{n}) \leq g(\bar{n}+1) < 0$, so that $\bar{n}$ is not stable. Then*

(1) *there exists a cost reduction that implements exactly $\bar{n}$.*
(2) *if $g(\bar{n}+1) = g(\bar{n})$ (resp. $g(\bar{n}+1) > g(\bar{n})$), then $\epsilon_{inf}^{\bar{n}} = -g(\bar{n})$ (resp. $\epsilon_{inf}^{\bar{n}} > -g(\bar{n})$).*

Props 2 and 3 highlight the importance of the relative magnitudes of $g(\bar{n})$ and $g(\bar{n}+1)$, under the assumption that both are negative. If $g(\bar{n}) > g(\bar{n}+1)$ the external stability constraint is slack, in the sense that internal stability guarantees external stability. If $g(\bar{n}) < g(\bar{n}+1)$, the external stability condition is binding. In the latter case, some cost reduction on an open interval above $\bar{n}$ is required in order to satisfy the external stability condition.

## 5. General costs, concave abatement benefits

In this section we assume that the benefit function $B(\cdot)$ is increasing and strictly concave in aggregate benefits. We continue to assume, as in §4, that the abatement cost function $C$ is initially strictly convex. To avoid dealing with nonnegativity constraints (see for example, Diamantoudi & Sartzetakis (2006), Rubio & Ulph (2006)) we assume

$$C'(0) \quad < \quad B'(Nq^*) \text{ where } q^* \text{ is the socially optimal level of abatement.} \tag{31}$$

Since signatory abatement levels are bounded above by $q^*$ in any Cournot or Stackelberg equilibrium, this assumption guarantees that an interior solution to the non-signatory's first order condition (2b) exists.

When we relax the assumption of linear marginal benefits, our problem becomes more complex for two related reasons. First, we can no longer normalize a non-signatory's marginal benefit function to unity; nor can we normalize a signatory's marginal benefit function to the number of members of the IEA. A consequence is that our simplification (5) of expression (3) for the joiner's gain function is no longer valid; also invalid is our decomposition (6), derived from the simplification (5), of the joiner's gain function into *UA* and *CR*. Second, a non-signatory's optimization problem is no longer independent of the decisions made by signatories, so that for each $n$, the abatement choices made by signatories and non-signatories must be determined simultaneously. Moreover, if an additional member joins an IEA, so that total abatement by signatories increases, abatement levels by non-signatories will decrease in response. (Accordingly, we now denote non-signatory abatement when there are $n$ signatories by $q_{\mathrm{f}}^n$ rather than $q_{\mathrm{f}}$.)

Because of these complexities, the cost-reduction required to implement an IEA with at least $n$ members is less straightforward than the one we constructed to prove Prop. 2. Moreover, in this context, it would be difficult to characterize the class of efficient cost reductions.

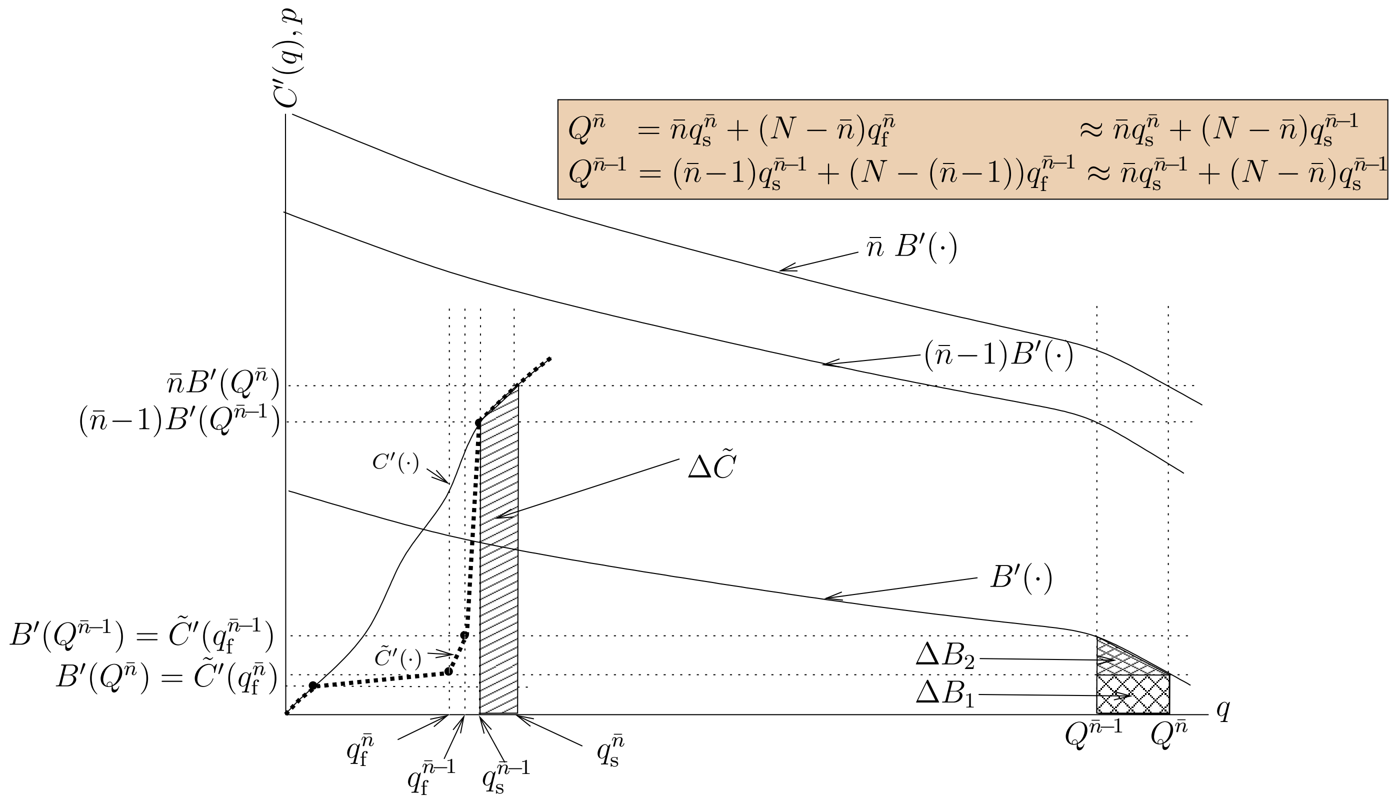


FIGURE 6. Implementing $\bar{n}$ with concave benefits

Accordingly the scope of Prop. 4, presented in §5 below, is narrower than its counterpart, Prop 2, for linear benefits. We have not attempted to identify the further cost-reduction that would be required to implement exactly $n$, so we have no counterpart to Prop 3. Another consequence of the additional complexity is that the Cournot and Stackelberg versions of our game no longer yield identical results. In the remainder of this section, we restrict our attention to the Cournot case. In an online appendix, we establish that given a benefit and initial cost function, a cost-reduction that implements at least $n$ as a Cournot equilibrium will also implement at least $n$ as a Stackelberg equilibrium.

Fix a strictly increasing marginal cost function $C'$, a strictly decreasing marginal benefit function $B'$, and $1 < \bar{n} \leq N$. Fig. 6 illustrates a reduced marginal cost function $\tilde{C}'$ which implements at least $\bar{n}$ as a stable equilibrium. The original marginal cost curve $C'$ is indicated by a thin solid line and the modified curve $\tilde{C}'$ by a heavy dotted line. The two curves overlap near the origin, and beyond $q_s^{\bar{n}-1}$. The region where they differ consists of a segment that is arbitrarily close to horizontal, followed by one that is arbitrarily close to vertical.

We show that under the modified marginal cost function $\tilde{C}'$, the joiner's gain function $\tilde{g}(\cdot)$ satisfies the internal stability condition (24a). From (14), both internal and external stability hold for some $n' \geq \bar{n}$. The key to our construction is to ensure that for both $\bar{n}-1$ and $\bar{n}$, the optimal levels of abatement for non-signatories, respectively $q_f^{\bar{n}-1}$ and $q_f^{\bar{n}}$, are virtually the same as, but slightly smaller than $q_s^{\bar{n}-1}$; we impose this by making modified marginal costs virtually vertical on the interval $(q_f^{\bar{n}}, q_s^{\bar{n}-1})$ which contains $q_f^{\bar{n}-1}$.

Because $q_s^{\bar{n}-1}$ and $q_f^{\bar{n}-1}$ are arbitrarily close together under this construction, aggregate abatement with $\bar{n}-1$ signatories is $Q^{\bar{n}-1} \approx Nq_s^{\bar{n}-1}$. Similarly, because $q_s^{\bar{n}-1}$ and $q_f^{\bar{n}}$ are arbitrarily close together, aggregate abatement with $\bar{n}$ signatories is $Q^{\bar{n}} \approx \bar{n}q_s^{\bar{n}} + (N-\bar{n})q_s^{\bar{n}-1}$. Thus, $Q^{\bar{n}} - Q^{\bar{n}-1} \approx \bar{n}(q_s^{\bar{n}} - q_s^{\bar{n}-1})$. This property simplifies our analysis because, by construction, we have eliminated the complication that non-signatory abatement depend on $\bar{n}$.

A closely related consequence of our construction is that when there are $\bar{n}-1$ signatories under the modified cost structure, it costs a signatory virtually the same to produce $q_s^{\bar{n}-1}$ as it costs a non-signatory to produce $q_f^{\bar{n}-1}$. This implies that in the joiner's gain function $\tilde{g}(\cdot)$ (see expression (3)), the term $\tilde{C}(q_f^{\bar{n}-1})$ can be replaced by $\tilde{C}(q_s^{\bar{n}-1})$, so that $\tilde{g}(\bar{n}) \approx \Delta B - \Delta\tilde{C}$, where $\Delta B$ is the per-signatory gain from increased abatements, and $\Delta\tilde{C}$ is the per-signatory incremental cost, when the IEA acquires an additional signatory. To establish that $\tilde{g}(\bar{n})$ is positive, we now only need to show that $\Delta B$ exceeds $\Delta\tilde{C}$.

The magnitude of $\Delta\tilde{C}$ (the tall thin hatched trapezoid in Fig. 6) is strictly less than the area of the rectangle that encloses it with height $\bar{n}B'(Q^{\bar{n}})$ and width $(q_s^{\bar{n}} - q_s^{\bar{n}-1})$. $\Delta B$ is the short wide cross-hatched trapezoid to the right of the figure, the union of the rectangle $\Delta B_1$ and triangular-shaped $\Delta B_2$. The height of $\Delta B_1$ is $B'(Q^{\bar{n}})$, i.e., $1/\bar{n}$ times the height of rectangle that contains $\Delta\tilde{C}$. The width of $\Delta B_1$ is $Q^{\bar{n}} - Q^{\bar{n}-1} \approx \bar{n}(q_s^{\bar{n}} - q_s^{\bar{n}-1})$ i.e., $\bar{n}$ times the width of the rectangle that contains $\Delta\tilde{C}$. Thus the area of $\Delta B_1$ strictly exceeds that of $\Delta\tilde{C}$. Moreover, the trapezoid $\Delta B$ also includes $\Delta B_2$ which has positive area. Thus $g(\bar{n}) = \Delta B - \Delta\tilde{C} > 0$, verifying that $\bar{n}$ is internally stable. This informal argument is made precise in Prop 4 below.

**Proposition 4.** *Assume that the benefit function $B$ and initial cost function $C$ are both twice continuously differentiable, that $B$ is strictly concave and $C$ is strictly convex. Then for any $1 < \bar{n} \leq N$, there exists an admissible cost reduction such that under the modified cost function $\tilde{C}$, an IEA with at least $\bar{n}$ members is stable.*

## 6. Conclusion

An important strand of the theory of IEAs focuses on the noncooperative Nash equilibria of the participation game introduced by D'Aspremont et al. (1983). The literature to date, which has focused exclusively on parametric examples, has been highly pessimistic about the prospects for international environmental cooperation. Its conclusion is that the equilibrium size of a stable IEA is small except when the potential gains from cooperation are also small. Moreover, although reductions in abatement costs necessarily increase the potential gains from cooperation, analysis of leading special cases has shown that such reductions leave unchanged or reduce equilibrium membership.

This paper approaches the problem from a non-parametric perspective. We introduce a novel decomposition of the gains to participation. For linear marginal benefits and general convex abatement costs, we use this decomposition to show that regardless of the potential gains from cooperation, if *marginal* abatement costs are convex then the equilibrium coalition size cannot exceed three. Maintaining convexity of marginal abatement costs, reductions in abatement costs cannot increase membership. In this respect, the results obtained in the previous literature are robust.

In an important respect, however, these previous results are fragile. When marginal abatement costs can be locally concave over a critical range of abatement levels, there are no restrictions on the size of the equilibrium coalition. Indeed, the noncooperative Nash equilibrium can realize up to 100% of the potential gains from cooperation, even when these potential gains are large. Moreover, we show that cost reductions *can* increase equilibrium membership and, when marginal benefits are linear, we derive a lower bound on the magnitude of the cost reduction needed to induce an arbitrary level of cooperation. Finally, we show that when benefits are strictly concave in abatement, it is still possible to support agreements among an arbitrarily large fraction of potential member countries.

Our analysis highlights the danger of drawing sweeping conclusions, however plausible, from parametric examples. In particular, our results undermine the apparently robust basis for pessimism that has become almost conventional wisdom in the field. On the other hand, one should not interpret our results as sufficient grounds for optimism about the prospects for real-world global agreements, since the kinds of cost functions we show to be consistent with such agreements might be difficult to design in practice.

## References


Baker, E. & Adu-Bonnah, K. (2008). Investment in risky r&d programs in the face of climate uncertainty, *Energy Economics* **30**(2): 465–486.

Baker, E., Clarke, L. & Shittu, E. (2008). Technical change and the marginal cost of abatement, *Energy Economics* **30**: 2799–2816.

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

Barrett, S. (2002). Consensus treaties, *Journal of Institutional and Theoretical Politics* **158**: 519–41.

Barrett, S. (2003). *Environment and Statecraft*, Oxford University Press.

Barrett, S. (2006). Climate treaties and breakthrough technologies, *American Economic Review* **96**: 22–25.

Bauman, Y., Lee, M. & Seeley, K. (2008). Does technological innovation really reduce marginal abatements costs?, *Environmental Resource Economics* **40**: 507–527.

Bloch, F. & Dutta, B. (2008). Formation of Networks and Coalitions, *Handbook of Social Economics, J. Benhabib, A. Bisin and M. Jackson, eds* .

Bos, I. & Harrington Jr, J. (2010). Endogenous cartel formation with heterogeneous firms, *The RAND Journal of Economics* **41**(1): 92–117.

Burger, N. & Kolstad, C. (2009). Voluntary public goods provision: coaltion formation and uncertainty. NBER Working Papers 15543.

Carraro, C. & Siniscalco, D. (1993). Strategies for the international protection of the environment, *Journal of Public Economics* **52**: 309–28.

Chwe, M. S. (1994). Farsighted coalitional stability, *Journal of Economic Theory* **63**: 299–325.

Dannenberg, A., Lange, A. & Sturm, B. (2009). On the formation of coalitions to provide public goods – experimental evidence from the lab. Working paper.

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

De Bondt, R. (1997). Spillovers and innovative activities* 1, *International Journal of Industrial Organization* **15**(1): 1–28.

de Zeeuw, A. (2008). Dynamic effects on the stability of international environmental agreements, *Journal of Environmental Economics and Management* **55**(2): 163–74.

Diamantoudi, E. & Sartzetakis, E. (2002). International environmental agreements - the role of foresight. University of Aarhus working paper 2002-10.

Diamantoudi, E. & Sartzetakis, E. (2006). Stable international environmental agreements: An analytical approach, *Journal of Public Economic Theory* **8**(2): 247–263.

Dixit, A. & Olson, M. (2000). Does voluntary participation undermine the coase theorem, *Journal of Public Economics* **76**: 309 – 335.

Donsimoni, M., Economides, N. & Polemarchakis, H. (1986). Stable cartels, *International Economic Review* **27**(2): 317–327.

Ecchia, G. & Mariotti, M. (1998). Coalition formation in international environmental agreements and the role of institutions, *European Economic Review* **42**(3-5): 573–582.

Enderes, A. & Ohl, C. (2003). International environmental cooperation with risk aversion, *Intrnational Journal of Sustainable Development* **6**: 378–392.

Escrihuela-Villar, M. (2009). A note on cartel stability and endogenous sequencing with tacit collusion, *Journal of Economics* **96**(2): 137–147.

Eyckmans, J. (2001). On the farsighted stability of the Kyoto Protocol. CLIMNEG Working Paper 40, CORE, Universite Catholique de Louvain.

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

Finus, M. (2003). 3. Stability and design of international environmental agreements: the case of transboundary pollution, *The international yearbook of environmental and resource economics 2003/2004: a survey of current issues* p. 82.

Finus, M. & Maus, S. (2008). Modesty may pay!, *Journal of Public Economic Theory* **10**: 801–826.

Heal, G. & Kunreuther, H. (2011). Tipping climate negotiations, NBER Working Paper 16954.

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

Hoel, M. & De Zeeuw, A. (2010). Can a Focus on Breakthrough Technologies Improve the Performance of International Environmental Agreements?, *Environmental and Resource Economics* pp. 1–12.

Hong, F. & Karp, L. (2010). International environmental agreements with mixed strategies and investment. unpublished Working Paper.

Ioannidis, A., Papandreou, A. & Sartzetakis, E. (2000). International environmental agreements: A literature review, *Cahiers de Recherche du GREEN 00-08. Québec: Université Laval* .

Karp, L. (2012). The effect of learning on membership and welfare in an International Environmental Agreement, *Climatic Change* **110**: 499–505.

Katz, M. (1986). An analysis of cooperative research and development, *The Rand Journal of Economics* **17**(4): 527–543.

Kohler, M. (2002). Coalition formation in international monetary policy games, *Journal of International Economics* **56**(2): 371–385.

Kolstad, C. D. (2007). Systematic uncertainty in self-enforcing international environmental agreements, *Journal of Environmental Economics and Management* **53**: 68–79.

Kolstad, C. D. & Ulph, A. (2008). Learning and international environmental agreements, *Climatic Change* **89**: 125–41.

Kosfeld, M., Okada, A. & Riedl, A. (2009). Institution formation in public goods games, *American Economic Review* **99**: 1335–1355.

Osmani, D. & Tol, R. (2009). Toward farsightedly stable international environmental agreements, *Journal of Public Economic Theory* **11**: 455–92.
Ray, D. & Vohra, R. (2001). Coalitional power and public goods, *J. Polit. Economy* **109**(6): 1355–1382.
Rubio, S. J. & Ulph, A. (2006). Self-enforcing international environmental agreements revisited, *Oxford Economic Papers* **58**: 233–263.
Stiglitz, J. E. (2006). A new agenda for global warming, *Economists' Voice* pp. 1–4.
Thoron, S. (1998). Formation of a coalition-proof stable cartel, *Canadian Journal of Economics* **31**(1): 63–76.
Ulph, A. (2004). Stable international environmental agreements with a stock pollutant, uncertainty and learning, *Journal of Risk and Uncertainty* **29**: 53–73.
Xue, L. (1998). Coalitional stability under perfect foresight, *Economic Theory* **11**: 603–627.

## Appendix: Proofs

**Proof of Remark 1:** From (6) and (12), we have, after rearranging

$$g(\lceil {}^{c-a}/_{1-x}\rceil) \;=\; CR(\lceil {}^{c-a}/_{1-x}\rceil) - UA(\lceil {}^{c-a}/_{1-x}\rceil) \;=\; (1-x)\Big(\underbrace{\lceil {}^{c-a}/_{1-x}\rceil}_{\text{Term L}} \;-\; \underbrace{{}^{c-x}/_{1-x}}_{\text{Term R}}\Big) \quad (32)$$

(a) Fix $1 < a' < a$ and $(c,x) \in \mathrm{CX}(a)$. Since Term L is weakly decreasing in $a$ on $(x,1)$, $\lceil {}^{c-a'}/_{1-x}\rceil \geq \lceil {}^{c-a}/_{1-a}\rceil \geq {}^{c-x}/_{1-x}$, so that $(c,x) \in \mathrm{CX}(a')$. To prove that $\mathrm{CX}(a)$ is a strict subset of $\mathrm{CX}(a')$, pick $(c,x)$ such that ${}^{c-a}/_{1-x} = \lfloor {}^{c-x}/_{1-x}\rfloor < {}^{c-x}/_{1-x} < \lceil {}^{c-x}/_{1-x}\rceil$. Clearly $\lceil {}^{c-a}/_{1-x}\rceil = \lfloor {}^{c-x}/_{1-x}\rfloor$ so that $(c,x) \notin \mathrm{CX}(a)$. But since $a' < a$, $\lceil {}^{c-a'}/_{1-x}\rceil \geq \lfloor {}^{c-x}/_{1-x}\rfloor + 1 > {}^{c-x}/_{1-x}$, so that $(c,x) \in \mathrm{CX}(a')$. This proves that $\mathrm{CX}(a) \subsetneq \mathrm{CX}(a')$.

(b) Fix $1 < c' < c$. Note that Term L is weakly, while Term R is strictly increasing in $c$.
   (i) First pick $(a,x)$ such that ${}^{c'-a}/_{1-x} = \lfloor {}^{c'-x}/_{1-x}\rfloor < {}^{c'-x}/_{1-x} < \lceil {}^{c'-x}/_{1-x}\rceil$. Clearly $\lceil {}^{c'-a}/_{1-x}\rceil = \lfloor {}^{c'-x}/_{1-x}\rfloor$ so that $(a,x) \notin \mathrm{AX}(c')$. Now pick $c$ larger than $c'$ but sufficiently close that ${}^{c-x}/_{1-x} < \lceil {}^{c-x}/_{1-x}\rceil = \lceil {}^{c'-x}/_{1-x}\rceil$ and hence $\lceil {}^{c-a}/_{1-x}\rceil \geq \lfloor {}^{c-x}/_{1-x}\rfloor + 1 > {}^{c-x}/_{1-x}$, so that $(a,x) \in \mathrm{AX}(c)\backslash\mathrm{AX}(c')$.
   (ii) Now pick $(a,x)$ such that $\lceil {}^{c'-a}/_{1-x}\rceil = {}^{c'-x}/_{1-x}$ so that $(a,x) \in \mathrm{AX}(c')$. Since $a > x$, we can pick $c$ larger than $c'$ but sufficiently close that $\lceil {}^{c-a}/_{1-x}\rceil = \lceil {}^{c'-a}/_{1-x}\rceil < {}^{c-x}/_{1-x}$, so that $(a,x) \in \mathrm{AX}(c')\backslash\mathrm{AX}(c)$. Hence $\mathrm{AX}(c') \ominus \mathrm{AX}(c) \neq \emptyset$.

(c) Since, once again, Term L is weakly, while Term R is strictly increasing in $x$, we can prove this part with an argument analogous to that which proved part (b). ■

**Proof of Remark 2:** Let $C'(q) = \frac{q}{M}$. Because marginal costs are linear, $q_{\mathrm{s}}^{\bar{n}} - q_{\mathrm{s}}^{\bar{n}-1} = M$ which implies the area of $CR(n) = \frac{M}{2}$ and $\frac{n-1-1}{q_{\mathrm{s}}^{\bar{n}-1} - q_{\mathrm{f}}^{\bar{n}}} = \frac{1}{M}$, so $q_{\mathrm{s}}^{\bar{n}-1} - q_{\mathrm{f}}^{\bar{n}} = M(n-2)$ which implies that the area of $UA(n) = \frac{M}{2}(n-2)^2$. The root of $g(n) = 0$ satisfies $g(n) = \frac{M}{2} - \frac{M}{2}(n-2)^2 = 0$ or $1 - (n-2)^2 = 0$ so the solutions are $3, 1$. Because we have $g'(n) = -M(n-2) < 0$ for $n > 2$ we know that the root $n = 3$ is stable, i.e. $g(2) > g(3) = 0 > g(4)$. This confirms that for linear marginal costs, $n = 3$ is the unique non-trivial IEA size. ■

**Proof of Remark 3:** The equilibrium level of aggregate abatement with the coalition is $3(3M) + (N-3)M = M(6+N)$, so the global benefit of abatement is $MN(6+N)$. The aggregate equilibrium cost of abatement is $(N-3)\frac{M}{2} + 3\frac{9M}{2} = \frac{1}{2}M(N+24)$, so global equilibrium net benefits are $MN(6+N) - \frac{1}{2}M(N+24) = \frac{1}{2}M(11N + 2N^2 - 24)$. In the absence of a coalition, each nation abates at $M$, so global net benefits are $N\left(NM - \frac{M}{2}\right)$. The global gain due to the formation of the coalition is therefore $\frac{1}{2}M(11N + 2N^2 - 24) - N\left(NM - \frac{M}{2}\right) = 6M(N-2)$. If every nation were to join the coalition, aggregate welfare would be $N\left(N^2M - \frac{N^2M}{2}\right) = \frac{1}{2}N^3M$. The actual gain relative to the potential gain, i.e.

the fraction of the potential gain from cooperation actually achieved, is

$$\frac{6M\,(N-2)}{\frac{1}{2}N^3M - N\left(NM - \frac{M}{2}\right)} = 12\frac{N-2}{N\,(N^2-2N+1)}. \qquad \blacksquare$$

**Proof of Remark 4:** We solve for $K$ such that $\bar{n}$ is a stable equilibrium. As in fn. , the area of $UA(\bar{n}) = \frac{M}{2}(\bar{n}-2)^2$, while the area of $CR(\bar{n})$ is now $(q_s^{\bar{n}} - q_f^{\bar{n}})/2 = K/2$. Hence $CR(\bar{n}) = UA(\bar{n})$ if $K = (\bar{n}-2)^2M$. Moreover, because for $n > \bar{n}-1$, $g(n) = \frac{M}{2}\left((\bar{n}-2)^2 - (n-2)^2\right)$, $q'(\cdot) < 0$ on $(\bar{n}-1, \infty)$. Thus $g(\bar{n}) \geq 0 > g(\bar{n}+1)$, verifying that condition (4) for $\bar{n}$ to be a stable equilibrium is satisfied. $\blacksquare$

**Proof of Proposition 1:** Let $n = 2$ and let the tangent to the marginal cost curve, evaluated at $q_s^2$, be $MC'(q_s^2) = \frac{1}{M}$. The area of triangles $T^{CR}(3)$ and $T^{UA}(3)$ both equal $\frac{M}{2}$. Since by strict convexity, $MC(q)$ lies strictly above the tangent line through $q_s^2$, for all $q \neq q_s^2$, $CR(3) < \frac{M}{2} < UA(3)$. Therefore, $g(3) = CR(3) - UA(3) < 0$. Now assume $n > 2$. From (16),

$$\frac{dg\,(n)}{dn} = S(n) - S(n-1) - (n-1)\,\frac{dS\,(n-1)}{dn}.$$

But since $S(\cdot)$ is strictly concave, $S(n) < S(n-1) + \frac{dS(n-1)}{dn}$. Hence

$$\begin{aligned}\frac{dg\,(n)}{dn} &< S(n-1) + \frac{dS\,(n-1)}{dn} - S(n-1) - (n-1)\,\frac{dS\,(n-1)}{dn}\\ &= -\,(n-2)\,\frac{dS\,(n-1)}{dn},\end{aligned}$$

which is negative for $n > 2$. Since $g(3) < 0$ and $g(n)$ is strictly decreasing for $n \geq 3$, there can be no integer $n > 2$ such that $g(n) \geq 0$. Therefore, there are no stable equilibria with $n$ greater than 2. $\blacksquare$

**Proof of Proposition 2:** Part (4) has been proved in the text, by the line immediately following (30). We also proved in the text that $\varepsilon^{\lambda_1^*}$ implements at least $\bar{n}$ and that $||\varepsilon^{\lambda_1^*}|| = -g(\bar{n})$. To prove that $-g(\bar{n})$ is the minimal cost reduction for at least $\bar{n}$ (part (1)), it suffices to prove that $||\varepsilon|| \geq -g(\bar{n})$, for any admissible cost reduction $\varepsilon(\cdot)$ that implements at least $\bar{n}$. It will then follow immediately that $\varepsilon^{\lambda_1^*}$ is efficient for at least $\bar{n}$ (part (2)).

Assume now that $\varepsilon(\cdot)$ implements at least $\bar{n}$. Substituting (21a) into (22), we obtain

$$\begin{aligned}||\varepsilon|| &= \Delta^{\varepsilon} g\,(\bar{n}) + (\bar{n}-1)\,\varepsilon\,(\bar{n}-1) + \int_0^1 \varepsilon(p)dp + \int_{\bar{n}}^{N+1} \varepsilon\,(p)\,dp\\ &\underbrace{\geq}_{\text{from (24a)}} -g(\bar{n}) + (\bar{n}-1)\,\varepsilon\,(\bar{n}-1) + \int_0^1 \varepsilon(p)dp + \int_{\bar{n}}^{N+1} \varepsilon\,(p)\,dp.\end{aligned} \tag{33}$$

Since admissibility requires $\varepsilon(\cdot) \geq 0$, (33) implies that $||\varepsilon|| \geq -g(\bar{n})$, which is what we needed to prove. This completes the proof of parts (1) and hence (2).

To prove part (3), we first establish sufficiency. Assume that $\int_1^{\bar{n}} \varepsilon(p)dp = -g(\bar{n})$ and that $\varepsilon\Big([0,1) \cup \{\bar{n}-1\} \cup [\bar{n}, \infty]\Big) = 0$. Since $\varepsilon(\bar{n}-1) = 0$, (21a) implies $\Delta^{\varepsilon} g\,(\bar{n}) = -g(\bar{n})$, so that from (24a), $\varepsilon$ implements at least $\bar{n}$. Since $\varepsilon\Big([0,1)\cup[\bar{n},\infty]\Big) = 0$, (22) implies $||\varepsilon|| = -g(\bar{n})$. It now follows that $\varepsilon(\cdot)$ is efficient for at least $\bar{n}$.

We now prove necessity. Assume that $\varepsilon(\cdot)$ is efficient for at least $\bar{n}$, so that from part (1) of the proposition, $||\varepsilon|| = -g(\bar{n})$. This, together with (33) implies

$$0 \quad \geq \quad (\bar{n}-1)\,\varepsilon\,(\bar{n}-1) \quad + \quad \int_0^1 \varepsilon(p)dp \quad + \quad \int_{\bar{n}}^{N+1} \varepsilon\,(p)\,dp. \tag{34}$$

But since $\varepsilon(\cdot)$ is nonnegative, (34) implies

$$\varepsilon\,(\bar{n}-1) \quad = \quad \varepsilon\Big((0,1)\Big) \quad = \quad \varepsilon\Big((\bar{n},N+1)\Big) \quad = \quad 0. \tag{35}$$

while

$$\text{If } \varepsilon(\cdot) \text{ satisfies (20) and } \varepsilon(p) > 0 \text{ then } \exists\delta > 0 \text{ s.t. } \varepsilon(\cdot) > 0 \text{ on } (p, p+\delta). \tag{36}$$

(36) restates statement A-4 in the online appendix, where it is proved. (35) and (36) imply $\varepsilon\,(0) = \varepsilon\,(\bar{n}) = 0$, verifying that $\varepsilon\Big([0,1)\cup\{\bar{n}-1\}\cup[\bar{n},\infty]\Big) = 0$. Hence $-g(\bar{n}) = ||\varepsilon|| \equiv \int_0^{N+1}\varepsilon(p)dp = \int_1^{\bar{n}}\varepsilon(p)dp$. ■

**Proof of Proposition 4:** Let $C$ be an initial strictly convex cost function. We distinguish below between "script Q's" (denoted $\mathcal{Q}$) and regular $Q$'s: The $\mathcal{Q}^r$'s are solutions to certain equations relating $B'$ and $C'$, while the $Q^r$'s, as usual, denote aggregate abatements, i.e., $Q^r = rq^r_{\mathrm{s}} + (N-r)q^r_{\mathrm{f}}$. Having defined our $\mathcal{Q}^r$'s, we will construct a modified cost function $\tilde{C}$ with the property that for each $r$, aggregate abatement $Q^r$ with $r$ signatories equals $\mathcal{Q}^r$.

Now fix $\bar{n} \leq N$ and $\delta \geq 0$. Let $\mathcal{Q}^{\bar{n}-1}(\delta) = Nq^{\bar{n}-1}_{\mathrm{s}}(\delta) - (N-(\bar{n}-1))\delta$, and let $\mathcal{Q}^{\bar{n}}(\delta) = Nq^{\bar{n}}_{\mathrm{s}}(\delta) - 2(N-\bar{n})\delta$, where

$$q^{\bar{n}-1}_{\mathrm{s}}(\delta) \text{ is defined by the condition } C'(q^{\bar{n}-1}_{\mathrm{s}}(\delta)) = (\bar{n}-1)B'(\mathcal{Q}^{\bar{n}-1}(\delta)) \tag{37a}$$

$$q^{\bar{n}}_{\mathrm{s}}(\delta) \text{ is defined by the condition } C'(q^{\bar{n}}_{\mathrm{s}}(\delta)) = \bar{n}B'(\mathcal{Q}^{\bar{n}}(\delta)). \tag{37b}$$

Thus $\quad \mathcal{Q}^{\bar{n}}(\delta) - \mathcal{Q}^{\bar{n}-1}(\delta) \quad = \quad N(q^{\bar{n}}_{\mathrm{s}}(\delta) - q^{\bar{n}-1}_{\mathrm{s}}(\delta)) \quad - \quad \delta(N-(\bar{n}-1)). \qquad (38)$

Note that for some $\epsilon > 0$, $q^{\bar{n}}_{\mathrm{s}}(0) - q^{\bar{n}-1}_{\mathrm{s}}(0) \geq 2\epsilon$, since otherwise, $\mathcal{Q}^{\bar{n}}(0) - \mathcal{Q}^{\bar{n}-1}(0) \leq 0$, and, since $B$ is strictly concave, $\frac{\bar{n}B'(\mathcal{Q}^{\bar{n}}(0))}{(\bar{n}-1)B'(\mathcal{Q}^{\bar{n}-1}(0))} \geq \frac{\bar{n}}{\bar{n}-1}$ while, since $C$ is strictly convex, $\frac{C'(q^{\bar{n}}_{\mathrm{s}}(0))}{C'(q^{\bar{n}-1}_{\mathrm{s}}(0))} < 1$, in which case equalities (37a) and (37b) could not be satisfied simultaneously. Hence by continuity, there exists $\bar{\delta}$ s.t. if $\delta < \bar{\delta}$, then $q^{\bar{n}}_{\mathrm{s}}(\delta) - q^{\bar{n}-1}_{\mathrm{s}}(\delta) \geq \epsilon$ and $\mathcal{Q}^{\bar{n}}(\delta) - \mathcal{Q}^{\bar{n}-1}(\delta) > 0$.

Fig 6 illustrates a modified cost function $\tilde{C}(\cdot|\delta) \leq C(\cdot)$, for $0 < \delta < \bar{\delta}$. We first identify two intervals on which $C$ and $\tilde{C}(\cdot|\delta)$ agree, and define $\tilde{C}(\cdot|\delta)$ at some critical points in between.

$$\tilde{C}'(q|\delta) \quad = \quad \begin{cases} C'(q) & \text{if } q \leq (C')^{-1}(B'(\mathcal{Q}^{\bar{n}}(\delta)) - \delta) \\ B'(\mathcal{Q}^{\bar{n}}(\delta)) & \text{if } q = q_s^{\bar{n}-1}(\delta) - 2\delta \\ B'(\mathcal{Q}^{\bar{n}-1}(\delta)) & \text{if } q = q_s^{\bar{n}-1}(\delta) - \delta \\ C'(q) & \text{if } q \geq q_s^{\bar{n}-1}(\delta) \end{cases} . \tag{39}$$

For the remainder of the proof we will suppress the argument $\delta$ when refering to the functions that depend on it. It is straightforward to verify that $\tilde{C}'(\cdot)$ strictly increases with $q$ over the restricted domain for which it is has so far been defined. For the remaining intervals, define $\tilde{C}'(\cdot)$ so that the entire function is strictly increasing and differentiable. Observe also that since $C'$ is strictly increasing, if $\delta > 0$ is sufficiently close to zero, then $\tilde{C}'(\cdot)$ will lie strictly below $C(\cdot)$ on the interval $\Big((C')^{-1}(B'(\mathcal{Q}^{\bar{n}}) - \delta), q_s^{\bar{n}-1}\Big)$.

By construction, under the cost function $\tilde{C}$, a signatory to an IEA with $\bar{n}-1$ members chooses $q_s^{\bar{n}-1}$ while a non-signatory chooses $q_f^{\bar{n}-1} = q_s^{\bar{n}-1} - \delta$. To verify this, note that

$$Q^{\bar{n}-1} \quad \equiv \quad (\bar{n}-1)q_s^{\bar{n}-1} + (N-(\bar{n}-1))q_f^{\bar{n}-1} \quad = \quad Nq_s^{\bar{n}-1} - (N-(\bar{n}-1))\delta \quad \equiv \quad \mathcal{Q}^{\bar{n}-1} \tag{40}$$

so that

$$\tilde{C}'(q_f^{\bar{n}-1}) \quad \equiv \quad \tilde{C}'(q_s^{\bar{n}-1} - \delta) \quad \underbrace{=}_{\text{from (39)}} \quad B'(Q^{\bar{n}-1})$$

$$\tilde{C}'(q_s^{\bar{n}-1}) \quad \underbrace{=}_{\text{from (39)}} \quad C'(q_s^{\bar{n}-1}) \quad \underbrace{=}_{\text{from (37a)}} \quad (\bar{n}-1)B'(\mathcal{Q}^{\bar{n}-1}) \quad \underbrace{=}_{\text{from (40)}} \quad (\bar{n}-1)B'(Q^{\bar{n}-1}).$$

Similarly, a signatory to an IEA with $\bar{n}$ members chooses $q_s^{\bar{n}}$ while a non-signatory chooses $q_f^{\bar{n}} = q_s^{\bar{n}-1} - 2\delta$, since

$$Q^{\bar{n}} \quad = \quad \bar{n}q_s^{\bar{n}} + (N-\bar{n})q_f^{\bar{n}} \quad = \quad Nq_s^{\bar{n}} - 2(N-\bar{n})\delta \quad = \quad \mathcal{Q}^{\bar{n}}, \tag{41}$$

so that

$$\tilde{C}'(q_f^{\bar{n}}) \quad \equiv \quad \tilde{C}'(q_s^{\bar{n}-1} - 2\delta) \quad \underbrace{=}_{\text{from (39)}} \quad B'(Q^{\bar{n}})$$

$$\tilde{C}'(q_s^{\bar{n}}) \quad \underbrace{=}_{\text{from (39)}} \quad C'(q_s^{\bar{n}}) \quad \underbrace{=}_{\text{from (37b)}} \quad (\bar{n})B'(\mathcal{Q}^{\bar{n}}) \quad \underbrace{=}_{\text{from (40)}} \quad \bar{n}B'(Q^{\bar{n}}).$$

We now show that $g(\bar{n}) > 0$. In the chain of relationships below, we use approximation signs ($\approx$) to indicate that the difference between the right and left hand sides of the approximation sign is $O(\delta)$.[10] Because $\delta$ can be chosen to be arbitrarily small, the approximations can be treated for our purposes as effectively equalities. From (3),

$$\begin{aligned} g(\bar{n}) \quad &= \quad B(Q^{\bar{n}}) - \tilde{C}(q_s^{\bar{n}}) \quad - \quad \Big[B(Q^{\bar{n}-1}) - \tilde{C}(q_f^{\bar{n}-1})\Big] \\ &= \quad B(Q^{\bar{n}}) - \left(\tilde{C}(q_s^{\bar{n}-1}) \quad + \quad \int_{q_s^{\bar{n}-1}}^{q_s^{\bar{n}}} \tilde{C}'(q)dq\right) \quad - \quad \Big[B(Q^{\bar{n}-1}) - \tilde{C}(q_f^{\bar{n}-1})\Big]. \end{aligned}$$

Because $\tilde{C}'$ is strictly increasing

$$g(\bar{n}) \quad > \quad B(Q^{\bar{n}}) - \Big(\tilde{C}(q_s^{\bar{n}-1}) \quad + \quad \tilde{C}'(q_s^{\bar{n}})(q_s^{\bar{n}} - q_s^{\bar{n}-1})\Big) \quad - \quad \Big[B(Q^{\bar{n}-1}) - \tilde{C}(q_f^{\bar{n}-1})\Big].$$

[10] Big O notation: We write $x(\delta) - y(\delta) = O(\delta)$. if there exists $\alpha \in \mathbb{R}$ such that as $\delta$ approaches zero, $|x(\delta) - y(\delta)| < \alpha\delta$.

Since $\tilde{C}(q_{\rm f}^{\bar{n}-1}) \approx \tilde{C}(q_{\rm s}^{\bar{n}-1})$

$$\begin{aligned} g(\bar{n}) &\approx B(Q^{\bar{n}}) \;-\; \tilde{C}'(q_{\rm s}^{\bar{n}})(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \;-\; B(Q^{\bar{n}-1}) \\ &= \int_{Q^{\bar{n}-1}}^{Q^{\bar{n}}} B'(q)dq - \tilde{C}'(q_{\rm s}^{\bar{n}})(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}). \end{aligned}$$

Since $B' < 0$ and, from (38), $Q^{\bar{n}} - Q^{\bar{n}-1} = N(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \;-\; \delta(N - (\bar{n}-1)) > 0$,

$$\begin{aligned} g(\bar{n}) &> B'(Q^{\bar{n}})\Big(Nq_{\rm s}^{\bar{n}} - Nq_{\rm s}^{\bar{n}} \;-\; \delta(N-(\bar{n}-1))\Big) - \tilde{C}'(q_{\rm s}^{\bar{n}})(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \\ &\approx B'(Q^{\bar{n}})\Big(Nq_{\rm s}^{\bar{n}} - Nq_{\rm s}^{\bar{n}}\Big) - \tilde{C}'(q_{\rm s}^{\bar{n}})(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \\ &= \Big(NB'(Q^{\bar{n}}) - \tilde{C}'(q_{\rm s}^{\bar{n}})\Big)(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \\ &\geq \Big(\bar{n}B'(Q^{\bar{n}}) - \tilde{C}'(q_{\rm s}^{\bar{n}})\Big)(q_{\rm s}^{\bar{n}} - q_{\rm s}^{\bar{n}-1}) \quad \underbrace{=}_{\text{from (37b)}} \quad 0. \end{aligned}$$

■