Coalitional Power and Public Goods

Debraj Ray

New York University

Rajiv Vohra

Brown University

We study the provision of public goods when all agents have complete information and can write binding agreements. This framework is in deliberate contrast to a traditional view of the free-rider problem based on hidden information or voluntary provision. We focus on coalition formation as a potential source of inefficiency. To this end, we develop a notion of an *equilibrium coalition structure*, based on the assumption that each coalition that forms does so under a rational prediction of the society-wide coalition structure. In a simple model, we characterize the (unique) equilibrium coalition structure. Only in some cases does the equilibrium involve full cooperation, resulting in efficient provision of the public good. In other cases, the equilibrium consists of several coalitions and inefficient provision. However, the degree of inefficiency and the number of possible coalitions are bounded.

I. Introduction

The normative theory of public-goods provision, leading from Lindahl (1919) to Samuelson (1954) to Foley (1970), provides a clear and rigorous characterization of efficiency in public-goods economies. However, a characterization of efficiency, despite the Lindahlian connection with prices, leaves unanswered the question of *how* an economy may

Ray acknowledges financial support under National Science Foundation grant SBR-9709254 and a John Simon Guggenheim fellowship. Vohra acknowledges support from a Salomon research award, Brown University. We thank two anonymous referees for comments on an earlier version.

[*Journal of Political Economy*, 2001, vol. 109, no. 6]
 0022-3808/2001/10906-0003$02.50

attain efficiency. Accordingly, much of the more recent literature on public goods has concentrated on the positive theory of incentives associated with the free-rider problem. The underlying premise in this literature is that the only possible impediment to achieving efficiency stems from the planner's lack of information regarding agents' characteristics. The central concern then is to elicit private information from the agents or, more generally, to design mechanisms that induce the agents to act in a way that leads to an efficient level of provision of public goods. Remarkable progress has been made in this literature (see, e.g., Green and Laffont 1979; Laffont 1987). Following Clarke (1971) and Groves (1973), to use Sonnenschein's (1998) words, "one does not so lightly say that it is not possible to design cooperation" (p. 10).

Nevertheless, we argue in this paper that there is an important sense in which the free-rider problem remains unresolved. (Indeed, this is a point that applies more generally to mechanism design.) *Coalitions* of agents might reject the proposed mechanism in favor of agreements (or mechanisms) among themselves. Our primary aim is to analyze the nature of cooperation, and possible inefficiency, that may arise in equilibrium (contrary to the Coasian prediction) when coalitions can form.

To make this point in the simplest possible way, we deliberately consider a world of complete information, which renders trivial the problem of mechanism design. Thus all relevant information is commonly known to all the agents. Moreover, we assume that agents can write binding agreements regarding their contributions toward the provision of a (pure) public good.

Now, suppose that a Lindahl allocation is proposed for the economy. Is it possible that a coalition may be able to do better for its members by seceding from the grand coalition and selecting some feasible allocation for itself? We are aware that a possible answer is "no," simply on the basis of the fact that a Lindahl allocation belongs to the "core" (more precisely, the α core) of the economy and is therefore immune to coalitional "deviations." However, this notion of the core is ill-suited to deal with positive issues of coalition formation. An allocation in the α core is coalitionally stable in the sense that no coalition can improve on it *if* it assumes that the agents outside this coalition will then decide to make no contributions whatsoever. But this may well be a very unreasonable forecast of what the others, *in their own interest,* might contribute.[1]

This is by no means a new criticism of the α core in a public-goods

[1] In the language of mechanism design, it is possible to include in the mechanism the possibility that coalitions will adopt their own mechanism simply by also incorporating the rule that the others will then agree to contribute nothing. But this is clearly not credible.

economy.[2] It was discussed in the general context of externalities in Rosenthal (1971) and more specifically for public-goods economies in Richter (1974), Roberts (1974), and Champsaur, Roberts, and Rosenthal (1975). As Roberts concludes, "the simple adaptation of the definition of the core which has proven appropriate for private goods economies may not be suitable with public goods economies" (p. 39).

So it was well recognized that a satisfactory theory of coalitional behavior in a public-goods economy must rely on coalitions to make reasonable predictions regarding the reaction of agents in the complementary coalition. But little progress has been made in this direction. In fact, the following statement from Roberts continues to be relevant close to three decades later and provides a good motivation for the present paper: "The task of developing an alternative core definition (or some other formalization of the intuitive notion of social stability) which better recognizes the structure of the public goods problem is a very delicate one. Guaranteeing that solution allocations will exist in a significant class of economies proves to be a particularly difficult problem" (p. 40).

We develop a notion of equilibrium that allows agents to form cooperative agreements, that is, form coalitions and make collective decisions about their level of contributions within a coalition. This equilibrium notion has two components. First, given some coalition structure, which is just a partition of all agents who write binding agreements into subsets, we take it that *cross-coalition* interaction is noncooperative. (For instance, the coalition structure of singletons would simply induce the familiar "voluntary contributions" game.) Second, we require that each coalition make a rational prediction of the overall coalition structure.[3] For a coalition contemplating a deviation, *this* is the prediction to be kept in mind when evaluating payoffs, not some arbitrary description of the way in which nondeviating agents might retaliate.

[2] Note that this is not a criticism of the standard notion of the core in a model in which a coalition's payoffs are well defined independently of the actions of agents not in the coalition. We have no quarrel with the standard notion of the core in a model in which there is some natural way of defining the characteristic function.

[3] This analysis builds (in a self-contained way) on Bloch (1996) and Ray and Vohra (1999). There have been other recent attempts to develop equilibrium notions of coalitional stability that rely on more "reasonable" assumptions (compared to the α core) about the behavior of agents outside a coalition. Carraro and Siniscalco (1993) and Chander and Tulkens (1997) analyze models of pollution control similar to the one we shall study below. But both papers make specific assumptions about the coalitional behavior of the outsiders in response to the formation of a new coalition. While Carraro and Siniscalco assume that the remaining coalitions do not change in any way, Chander and Tulkens assume that the outsiders disintegrate into singletons. In contrast, we develop an equilibrium notion in which the *entire* coalition structure is endogenously determined in equilibrium.

We recognize the possibility that our equilibrium concept may lead us to conclude that any proposed (efficient) allocation in the grand coalition may be subject to a coalitional deviation. However, in our view, this is not a reason to abandon the entire endeavor. All it suggests is that our equilibrium concept should not be based on the *presumption* of efficiency. In particular, it is not appropriate to refer to our equilibrium notion as some particular modification of the core. As we shall explain below, it is better viewed as a notion of an equilibrium coalition structure that emerges sequentially. To be sure, agents *within* a coalition maximize surplus (in the sense of playing best responses) and achieve efficiency. Inefficiency arises, therefore, if the grand coalition of all agents fails to form, as in Ray and Vohra (1997) and Dixit and Olson (1998).

Our main objective is to provide a complete characterization of the equilibrium coalition structure in a public-goods model. For this reason we shall describe in some detail the process through which agents negotiate to form coalitions. While a formal description of the process is contained in Section III and a discussion of the robustness of the negotiating process is provided in Section V, the main features are as follows. (Readers who are especially interested in, or skeptical of, the negotiating process and its connections with Coase [1960] are invited to read Sec. V before proceeding.)

The formation of a coalition (indeed, the very *definition* of "coalition" in this paper) will mean that member-agents make a binding agreement regarding their individual contributions toward the public good. Further, we shall assume that once a coalition is formed, it cannot change its composition.[4] Of course, the formation of a coalition will require the unanimous consent of all its members. In agreeing to such an arrangement, all members of the coalition must view this as their best alternative. Finally, in evaluating any such coalitional agreement, the members of the coalition must make a prediction (which in equilibrium will be correct) of the contributions of agents in the complementary coalition. The actions of the complement will be based on similar considerations. Some or all of the remaining agents may form a coalition of their own, again predicting the behavior of the remainder, and so forth.

An equilibrium will therefore determine a coalition structure, reflecting cooperation among agents within each coalition in the structure. Thus full cooperation refers to the formation of the grand coalition, and no cooperation refers to the coalition structure of singletons. The latter case corresponds to a Nash equilibrium of a game of voluntary contributions, as, for example, in Bergstrom, Blume, and Varian (1986).

[4] This assumption matters. But it is unclear how to drop it. On this, see Sec. V.

The former case corresponds to an equilibrium in which the outcome is efficient. Thus an equilibrium coalition structure describes an endogenously determined collection of cooperative agreements within each coalition in the coalition structure. This framework is designed to allow for the possibility that while full cooperation is, in principle, possible, it may not emerge in equilibrium.

We apply this general methodology, developed in detail in Ray and Vohra (1999), to a simple scenario with one private good, one public good, and several identical agents. As we shall see, even this elementary structure yields fairly complex outcomes.

Full cooperation and no cooperation are, of course, easy to describe in such a model. Because the model is so simple, it is also easy to see what the outcome will be for any arbitrary coalition structure. (In fact, we shall set things up so that each coalition has a dominant production strategy.) We can concentrate, therefore, on the central issue of concern: deriving the equilibrium coalition structure.

Three main insights underlie our findings. The first is this: if full cooperation is available as a *possible* outcome, equilibrium free-riding can never be "too extreme." In particular, the ability to cooperate will imply that agents choose *never* to play the individualistic Nash equilibrium involving voluntary contributions. Thus the degree of equilibrium inefficiency cannot be "too high," in a sense that we formalize for our simple model.

Second (which builds on the first point), if efficiency is not to be had in equilibrium, the equilibrium coalition structure will generally be asymmetric even if all individuals are identical. Inefficiencies that are borne symmetrically by all agents cannot persist if a binding agreement can be written for the grand coalition. (Our first point, which refers to the symmetric structure of singletons, is really a special case of this argument.)

Third, if some healthy form of partial cooperation can be sustained (in some nonsingleton, asymmetric coalition structure), this poses a real threat to the sustenance of full efficiency. In general, efficiency is not to be had even when there are no informational frictions and agents can write binding agreements.

We now turn to the specific findings of the model. Our first result demonstrates that in our model there is a (numerically) unique equilibrium coalition structure.[5] We shall identify an (infinite) strictly increasing sequence of positive integers, $\mathcal{T}^*$, containing 1, such that, for an economy with n agents, the grand coalition is formed in equilibrium

[5] A numerically unique coalition structure corresponds to a description of the size of each coalition. Since our model is symmetric, we cannot expect to predict the identity of the agents in any particular coalition.

if and only if n belongs to $\mathcal{T}^*$. Otherwise the equilibrium coalition structure is described as follows: the last coalition to form has size t, which is the largest integer in $\mathcal{T}^*$ less than n. The second-last coalition to form has size t', which is the largest integer in $\mathcal{T}^*$ no more than $n - t$, and so on.

Of course, equilibrium allocations are efficient if and only if the equilibrium coalition structure is the grand coalition. Thus efficiency obtains in equilibrium for every economy with n agents, where n belongs to $\mathcal{T}^*$; that is, efficiency emerges infinitely often as n varies. On the other hand, since $\mathcal{T}^*$ does not generally consist of *all* the positive integers, it follows that inefficiency is possible in equilibrium. However, there is a well-defined bound on the degree of inefficiency. Our second result shows that in an economy with n agents, the equilibrium level of the surplus is greater than the maximum (efficient) surplus in an economy with $n/2$ agents. Moreover, there is a significant degree of cooperation in equilibrium in the sense that the number of coalitions is bounded; if k is the number of coalitions in equilibrium, then $2^{k-1} < n$.

II. A Model of Public-Goods Provision

A. *Basic Structure and Payoffs*

We develop the simplest symmetric structure for public-goods provision. To fix ideas, we look at a model of pollution control. Suppose that there are n regions (to be interpreted as firms, countries, or geographical centers of decision making that share borders). Each region produces a pure public good—pollution control—the benefits of which accrue equally to *all* regions. Let z denote the public benefit of control activity pursued by any particular region, and assume that its generation involves a cost of $c(z)$, which is private to that region. We take it that c is increasing and strictly convex. Thus (after appropriate normalization), if Z is the total amount of pollution control produced by *all* regions, then the payoff to some region that produces z of it is

$$Z - c(z). \tag{1}$$

In what follows, we depart from the (individualistic) voluntary provisions model by permitting a region to deliberately link up with others. Thus we conceive of an initial phase of negotiations in which a region may make an offer to write a *binding agreement* with some other regions. If all the regions (to which the offer is made) agree on this arrangement, a *coalition* of regions forms, which is then bound to *jointly* decide on the extent of its control activity. That is, the free-rider problem across the regions in a coalition is assumed to be solved by the act of signing the binding agreement. Notice that since $c(z)$ is strictly convex, efficiency

within a coalition will imply that each member-region produces the same level of pollution control. The aggregate payoff to coalition S with cardinality s when each of its members produces z is

$$s[sz - c(z) + Z_{-i}],$$

where Z_{-i} is the aggregate production level of regions not included in S. Thus the problem facing a coalition with cardinality s is to produce pollution control of z per member, where z solves

$$\max_z sz - c(z). \tag{2}$$

This observation follows from the assumed linearity of external effects, so that the optimal production decision of a coalition does not depend on what other regions (outside the coalition) are doing; in terms of production decisions, each coalition has a dominant strategy. Of course, the payoff to each coalition also depends on the actions of the regions outside the coalition.

All the interest, therefore, centers on a description of *which* coalition structure will actually form. To answer this we need to specify a model of coalition formation.

In general, a coalitional agreement may specify not just production levels but also *transfers* across coalitional members.[6] In the next section we shall formally define two versions of a model of coalition formation: a restricted model in which each coalition is constrained to equally divide its surplus among member-regions[7] and a general model without this assumption. We shall now illustrate the basic features of our model through a simple example. Since the basic results remain the same even in the general model, it will be appropriate to assume the absence of transfers in the following example.

B. An Example

Given symmetry and the assumed absence of transfers, an offer essentially boils down to a proposal regarding the *number* of partners that a region seeks. We take it that such coalitions form sequentially: some region (it is unimportant which one) makes the first offer, then some unincluded region (if any) is chosen to make a second offer to other unincluded regions, and so on until all regions are formed into (possibly singleton) coalitions.

Let us examine the implications of this formulation for the special

[6] Notice that allowing for intracoalition transfers is equivalent to allowing arbitrary allocations of coalitional surplus.

[7] This simplifying assumption is actually quite common in the literature (see, e.g., Bloch 1996; Alesina and Spolaore 1997).

case in which the cost function is quadratic, that is, $c(z) = \frac{1}{2}z^2$. Then it is obvious that a coalition of size s will produce $z = s$ per member, or s^2 in all, and will incur a cost per region of $\frac{1}{2}s^2$ in doing so (simply solve [2]). It follows that if there are m coalitions with sizes $\mathbf{n} = \{s_1, s_2, \ldots, s_m\}$, then a coalition of size s_i will enjoy a payoff *per region* of

$$\sum_{j=1}^{m} s_j^2 - \tfrac{1}{2}s_i^2. \tag{3}$$

Since there are no transfers, this average is also the actual payoff to each member of a coalition of size s_i.

To begin with, suppose that there are only two regions. Note that the stand-alone payoff to each region is 1.5, whereas if the two regions form a coalition, then the payoff per region is 2 (simply apply [3]). It should therefore be obvious that any region will wish to team up with the other (and that such an offer will be accepted). Thus the two-region scenario implies full cooperation and efficiency.

The three-region case presents the first nontrivial prediction problem. If a single region contemplates staying on its own, it must predict what the remaining regions will do. But we know that in this case the remaining two regions *will* form a single coalition, so that the average worth of being alone is 4.5. A two-region coalition in this setting would average only 3. Finally, a three-region coalition averages 4.5 as well. This suggests that a region is indifferent between being on its own and being a member of a three-region coalition. Let us break this indifference in favor of the larger coalition (this assumption will not be needed in the general model). We conclude, then, that the three-region scenario is also conducive to efficiency.

Now turn to the four-region problem. By a similar process of computation and prediction, it turns out that if a region stays on its own, then it is in the interest of the other three regions to form a single coalition. Therefore, the average worth of the stand-alone region is 9.5. In contrast, the formation of the grand coalition of four regions yields an average worth of only 8. This suggests that full cooperation cannot occur when there are four regions.

What is the equilibrium coalition structure? With only four regions, it is easy to see what the answer must be. Given that the other three regions hang together when one stands alone, this must be the best coalition structure from the point of view of the stand-alone region. (It is easy to check that the payoff from forming a two- or a three-region coalition yields a lower average payoff.) It follows that in the negotiation game, it will always pay a region (which first gets the opportunity) to commit to standing alone: this will yield the highest return. Thus the "numerical" coalition structure that finally obtains is {1, 3}.

Now—strikingly enough—the five-region problem yields full coop-

eration. If one region were to stand alone, it would not be able to ensure the stability of the remaining four regions, which would configure themselves into the {1, 3} structure. Consequently, the average worth of the original region must be 10.5, whereas the formation of the grand coalition yields 12.5. The formation of a two-region coalition would yield an average worth of 11 for the two regions, which again is lower. Formation of three- and four-region coalitions is similarly ruled out.

However, full cooperation can (thereafter) no longer be reached unless there are at least eight regions, and then not again until there are at least 13 regions.[8] To establish these results, the following concerns are relevant: each region must compare the benefits of making an offer to an arbitrary subset of other regions, and moreover, for each such offer, a prediction of the "remaining" coalition structure is called for. Our general analysis will provide a complete characterization of the entire coalition structure for any "population size," not just the efficient outcomes. Moreover, this characterization embodies a significant reduction in the number of "checks" that need to be performed in order to ascertain the equilibrium structure.[9]

Indeed, if we had to proceed by brute-force methods beyond the five-region case, the problem would very quickly become intractable. The deduction of cooperative outcomes for the other population sizes cited above follows from our more general characterization.

There is, moreover, another reason why a general analysis is indispensable. The example, with its reliance on particular population sizes and seemingly odd cyclicities in efficiency, is not very useful in uncovering certain broad patterns. As a transition to the general analysis, we list two of these features.

First, a coalition structure of singletons—and, more generally, a symmetric inefficient coalition structure—is never an equilibrium. The reason is that the symmetric payoff per region to the grand coalition would always dominate the payoffs to such a structure. Thus if there is inefficiency (as in the four-region case of the example), the equilibrium structure must be asymmetric.

Second, as n increases, the degree of inefficiency cannot become too large. There are several ways to measure this (see below), but one crude measure is the "coarseness" of the coalition structure. With our general results, it can be checked that there can be no more than *seven* coalitions when the number of regions is 100, and no more than *10* coalitions

[8] The obvious comparison to the Fibonacci sequence (suggested to us by Andrew Postlewaite) ends here: the next size supporting full cooperation is 20!

[9] For instance, to ascertain the "stability" of the grand coalition when $n = 6$, it suffices to compare a stand-alone payoff under the structure {1, 5} with payoff per region under the grand coalition. Without the characterization, of course, numerous other structures would need to be considered.

when the number of regions is 1,000! (It is of interest to note, moreover, that these bounds do not depend on the quadratic cost function used in this example.) So while there may be *some* inefficiency as the number of regions grows, there cannot be too much.

Finally, the example illustrates starkly the kinds of commitment underlying our model. To understand this, consider any inefficient equilibrium structure, say {1, 3}. One might attempt to argue that the equal-division assumption (or any model of negotiation that yields equal division) is problematic here. The three worse-off regions could attempt to form the grand coalition with region 1 by bribing it with the stand-alone payoff (plus a signing bonus). This would be a way to circumvent the inefficient outcome.

However, this is a useless exercise ex ante. The model is symmetric, and in principle *each* of the four regions could play region 1. All four cannot be simultaneously bribed. Enlarging the ability of coalitions to divide their surplus unequally makes no difference (and this is what our general model of negotiations will show). On the other hand, the exercise has value ex post, when some region has *already committed* to standing alone but is open to renegotiation. The assumption of no renegotiation across coalitions—once a coalition has formed, it cannot expand in any way—is important. We return to a discussion of this issue in Section V.

III. Equilibrium Coalition Structures

The purpose of this section is to deliver a general characterization of equilibrium coalition structures. Accordingly, we return here to a general cost function. As for the model of coalition formation, we present the reader with two options. First, we retain the assumption of equal division of coalitional surplus. The reader who is comfortable with this postulate as a working hypothesis can read the paper in an entirely self-contained way. Second, we describe a general model of coalition formation that permits arbitrary allocations within coalitions. Using results from Ray and Vohra (1999), we prove that the predictions of this general model are no different from those of the simpler model.

In our view, it is important to allow transfers across coalitional members, especially if one is interested in questions of efficiency. The general model we describe allows for all sorts of intracoalitional allocations to be implemented, *in principle*. There is no a priori reason why such transfers cannot matter. (Indeed, we employ the particular public-goods structure to argue that they do not, but this is not generally the case.) Put another way, the general model provides the foundations of the equal-division assumption. Moreover, it tells us how to generalize these results to asymmetric situations, though the details of such an extension

are beyond the scope of the current exercise. In contrast, the equal-division hypothesis suggests no obvious extension to the asymmetric case.

A. *Coalition Formation*

We now describe a model of coalition formation. For a discussion of some conceptual issues underlying this model, see Section V.

In this model, regions are players. They will make proposals to coalitions (of regions) and respond to proposals made to them. To this end, we describe a negotiation *protocol:* if some set T of regions is yet to form into coalitions, a particular region in T gets to be the initial proposer. For each conceivable coalition of regions to which a proposal might be made, the protocol pins down an order of responses among the member-regions of that coalition. Thus an entire ordering of proposal and response (for each collection of negotiating regions) is laid down at the very outset: this is the negotiation protocol.

With the protocol in place, a bargaining game may be described. Some initial proposer starts the game. She chooses a coalition (of which she is a member) and then makes a proposal to this coalition.

In general, a proposal is a complicated object. It will include production plans, as well as a description of transfers among member-regions. However, as already noted, production within a coalition must be insensitive to coalition structure: a coalition of regions S with cardinality s will surely choose to produce pollution control of z per member, where z solves (2). To be sure, proposed *transfers* might continue to depend on the final coalition structure.

Once a proposal to coalition S is made by the initial proposer, attention shifts to the respondents in S (in the order prescribed by the protocol). By a response we mean simply an acceptance or rejection of the going proposal. If all respondents accept, the regions in S form their coalition and retire from the game, which continues among the remaining set of regions. In the case of a rejection, it is assumed that the first rejecter gets to make the next proposal.

If and when all agreements are concluded, a coalition structure forms. Each coalition in this structure is now required to allocate its surplus among its members as dictated by the proposals to which they were signatories. If bargaining continues forever, it is assumed that all regions receive a (normalized) payoff of zero.

At this point, we introduce our two options.

a. A restricted game in which its is assumed that a proposal always involves equal division of the surplus, that is, no transfers. In this case a proposal may simply be identified with a specification of a coalition (given equal division and the fact that production plans must solve [2]).

Furthermore, by symmetry, all that matters is the specification of coalitional *size* (see Bloch 1996). We shall also use the convention that if a player is indifferent between making acceptable proposals to coalitions of different sizes, he will opt to choose the largest of these sizes.[10]

b. A general game in which proposals may specify any arbitrary (feasible) division of the surplus among coalition members. To analyze this more general scenario, we adopt an extension of Rubinstein (1982) and Chatterjee et al. (1993); this is the negotiation model described in Ray and Vohra (1999).[11] This model differs from the simpler one in the following ways: (1) A proposer can propose to divide coalitional surplus in arbitrary ways. She can condition such divisions on the coalition structure that eventually forms. (2) A rejection entails the lapse of a certain amount of time, which imposes a geometric cost on all regions and is captured by a common discount factor $\delta \in (0, 1)$.

From now on, in the context of the restricted game, the word "equilibrium" will refer to a subgame-perfect equilibrium, and in the context of the general game, it will refer to a stationary perfect equilibrium.[12] Our main question is, Can we describe equilibrium coalition structure(s)?

Because our game is symmetric (insofar as production technologies and payoff functions are concerned), there can be no hope for a prediction that links *particular* regions to *particular* coalitions. What we look for, rather, is a description of *numerical* coalition structures. In other words, we describe, for every integer n (which denotes the total number of regions), a *decomposition* of that integer into other integers, so that the number of elements in the decomposition denotes the number of coalitions that form, and the value of each integer in the decomposition denotes the membership size of each coalition.

To do this, we begin with some preliminary observations regarding the outcome when a particular coalition structure forms.

[10] The more general model to follow will not require this convention.

[11] See also Chwe (1994), Bloch (1996, 1997), Yi (1996), Ray and Vohra (1997), and Huang and Sjostrom (1999), among others.

[12] A *(stationary) strategy* for a region will condition its (possibly probabilistic) proposal only on the current state of the game—the current set of negotiating regions and the coalitions that have already formed. It also requires that the accept-reject decision for proposals made to it by other regions not depend on anything else but the current set of regions, the coalitions that have already left, as well as the identity of the proposer and the nature of the proposal. A *stationary (perfect) equilibrium* is defined to be a collection of stationary strategies such that there is no history at which a region benefits by a deviation from its strategy. Since such a game may possess many subgame-perfect equilibria (see, e.g., Chatterjee et al. 1993), it is important to restrict attention to stationary equilibria.

B. Average Worths

Let S be a coalition of regions with cardinality s. Let $z(s)$ be its output per region, that is, the solution to (2),[13] and let

$$f(s) \equiv sz(s), \quad h(s) = c(z(s)), \quad g(s) = f(s) - h(s).$$

Thus $f(s)$ denotes the aggregate output of coalition s, $h(s)$ the corresponding cost per member of provision of the public good, and $g(s)$ the payoff per member from the activity of the coalition (with external effects neglected). If $\pi = \{S_1, \ldots, S_m\}$ is a coalition structure and $S \in \pi$, then the *average worth* of a region in S is just

$$\alpha(S, \pi) \equiv \sum_{j=1}^{m} f(s_j) - h(s). \tag{4}$$

Because all worths depend only on the *sizes* of the coalitions involved, we can abuse notation a little to write

$$\alpha(s, \mathbf{n}) = \sum_{j=1}^{m} f(s_j) - h(s), \tag{5}$$

where $\mathbf{n}$ is some *numerical* coalition structure and $s \in \mathbf{n}$. The average worth function will be crucial in describing a particular decomposition of positive integers that characterize equilibrium coalition structure.

We shall make the following regularity assumptions on the technology, which will be in force throughout the paper.[14]

ASSUMPTION 1. There exists a unique solution $z(s)$ to the problem (2) such that $z(s)$ is strictly positive for all $s > 1$ and strictly increasing in s.

ASSUMPTION 2. The function $f(s)$ is convex in s.

The nature of this regularity assumption can be made clear by observing that the maximization problem (2) can be viewed as a standard profit maximization problem, where s is any positive real number, interpreted as the price of the output, and y is interpreted as the output. Notice that it is sufficient for assumption 1 to assume that $c(\cdot)$ is twice continuously differentiable, $c' > 0$, $c'' > 0$, and $c'(0) < 1$. In assumption 2, we require f, the revenue function, to be convex.[15] It is easy to see that assumption 2 holds for all convex, exponential cost functions,

[13] We shall presently impose restrictions on $c(\cdot)$ to ensure that there exists a unique solution to (2).

[14] Assumption 2 is used in the proofs of lemma 1 and theorem 2.

[15] The standard result that the profit function is convex goes quite far in implying convexity of f, but, in general, not far enough.

$c(z) = bz^{\beta}$, where $b > 0$ and $\beta > 1$. How our results may change if assumption 2 is not satisfied remains an open question.[16]

C. *Decompositions of Positive Integers*

Here is a rule that generates numerical coalition structures from any given number n of regions. Let $\mathcal{T} = \{m_1, m_2, \ldots\}$ be an ordered collection of increasing positive integers, where $m_1 = 1$. For any integer $n \geq 2$, define the $\mathcal{T}$ *decomposition* of n as a collection $\mathbf{s}(n) \equiv (t_1, \ldots, t_k)$ of (possibly repeated) elements of $\mathcal{T}$ satisfying the following properties: (1) t_k is the largest integer in $\mathcal{T}$ that is strictly smaller than n; (2) for any $i \in \{1, \ldots, k-1\}$, t_i is the largest integer in $\mathcal{T}$ no greater than $n - \sum_{j=i+1}^{k} t_j$. In other words, the $\mathcal{T}$ decomposition $\mathbf{s}(n)$ is obtained by subtracting the largest integer in $\mathcal{T}$ (which is strictly smaller than n) from n, then subtracting the largest possible integer in $\mathcal{T}$ (no greater than the remainder) from the remainder, and so on. Notice that since $1 \in \mathcal{T}$, $\mathbf{s}(n)$ is well defined and unique for any positive $n > 1$.

Now consider a special collection $\mathcal{T}^* = \{m_1, m_2, \ldots\}$ of positive integers with the property that $m_1 = 1$, and for each $i \geq 1$, m_{i+1} is the *smallest* integer n, with the property that $n > m_i$ and

$$g(n) \geq \alpha(t_1, \mathbf{s}(n)) = \sum_{i=1}^{k} f(t_i) - h(t_1), \tag{6}$$

where $(t_1, \ldots, t_k)$ is the $\mathcal{T}^*$ decomposition of n. It is very easy to see that there is a unique sequence $\mathcal{T}^*$ satisfying this property, and in fact, this sequence is computable recursively.[17]

For any positive integer n, define its *strict decomposition* to be just its $\mathcal{T}^*$ decomposition and its *decomposition* to be either its strict decomposition if $n \notin \mathcal{T}^*$ or the singleton set $\{n\}$ if $n \in \mathcal{T}^*$. The notation $\mathbf{d}(n)$ ($\mathbf{s}(n)$) will refer to the decomposition (strict decomposition) of n. For notational convenience, the decomposition of zero is empty.

Given any decomposition and integers such as s and t, we shall use concatenation such as $s \cdot \mathbf{d}$ or $t \cdot s \cdot \mathbf{s}(n)$ to denote the numerical coalition structure generated by putting together these integers with the decomposition. (The order of appearance of the integers is unimportant.)

We now record for later use the following properties of $\mathcal{T}^*$ and $\mathbf{d}(n)$.

[16] As will become clear below, our results (through lemma 1 below) actually rely on the weaker assumption that $g(s) + f(n - s)$ is convex.

[17] The collection $\{m_1, \ldots, m_i\}$ is all that is needed to compute the $\mathcal{T}^*$ decomposition of any integer n that exceeds m_i, so recursion is possible. All we need to do is make sure that the inequality (6) *is* satisfied for some $n > m_i$. To see this, simply note that $f(n) - h(n) \to \infty$ as $n \to \infty$.

OBSERVATION 1. If $\mathbf{d}(n) = (n_1, \ldots, n_k)$, where $k \geq 2$, then $n_i \neq n_j$ for all $i, j \in \{1, \ldots, k\}$, $i \neq j$. If $\mathcal{T}^* = \{m_1, \ldots, m_i, m_{i+1}, \ldots\}$, then $m_1 = 1$, $m_2 = 2$, and $m_{i+1} < 2m_i$ for all $i \geq 2$.

The proof of this observation is a simple consequence of our definition of a decomposition. Suppose that there exist $m_i, m_j \in \mathbf{d}(n)$ such that $i \neq j$ and $m_i = m_j = m$. Then it is easy to see that $\mathbf{d}(2m) = (m, m)$. In particular, $2m \notin \mathcal{T}^*$, so that (6) is negated, which means that

$$f(m) + g(m) > g(2m). \tag{7}$$

Recall that $f(m) = mz(m)$, where $z(m)$ is the solution to (2) for $s = m$, and $f(m) + g(m) = 2mz(m) - c(z(m))$. But $g(2m)$, by definition, equals the *maximum value* of $2mz - c(z)$ over all z, so that (7) cannot hold. This is a contradiction, so the first part of observation 1 must be true. The second part of the observation now follows immediately.

D. *Equilibrium Coalition Structures*

For both the restricted game and the general game, we obtain the same characterization of the equilibrium (numerical) coalition structure.

THEOREM 1. Fix a set of regions n. Then there exists $\delta^* \in (0, 1)$ such that, for all $\delta \in (\delta^*, 1)$, there is a unique numerical coalition structure, which is just the decomposition of n.

We discuss this result, postponing its formal proof to the Appendix.

Combining observation 1 and theorem 1, we see that the equilibrium of this (ex ante) symmetric game typically displays a high degree of asymmetry.

COROLLARY 1. Suppose that the number of regions is n and the grand coalition does not form in equilibrium; that is, the equilibrium coalition structure is $\mathbf{d}(n) = (n_1, \ldots, n_k)$, where $k \geq 2$. Then $n_i \neq n_j$ for all $i, j \in \{1, \ldots, k\}$ and $n_k > n/2$.

The main idea behind the argument in theorem 1 is that whenever a coalition forms, it will tend to divide its worth *equally* among its members. This is true by assumption in the restricted game. For the general game, it is an *outcome* when discount factors are close enough to unity. Indeed, *any* other model of negotiation with the feature that average worth plays an important role will generate the same results.

With this in mind, a coalition S that forms will do so with the intention of maximizing its average worth, which we have denoted by $\alpha(S, \pi)$. But note well the dependence of this worth on both the coalition S and the entire coalition structure π of which it is a part. Thus it is not enough for S to simply form: its members must also attempt to *predict* the entire coalition structure in which S will be embedded.

Prediction is assisted by the understanding that *other* coalitions, when they form, will have the same motivation as S does. They will all seek

to maximize their average worths and will be confronted with a similar problem of prediction. This suggests that a solution may be found by a backward recursion argument: if only two regions are left, the problem of prediction is trivial (given that most of the coalition structure has already formed). The solution to the two-region case will, in turn, inform predictions in the three-region case: for instance, if a region decides to "go it alone," it will be able to predict what the other two regions will do. In this way, we can solve out for equilibrium coalition structures.

To understand how this recursive procedure boils down to a coalition structure that is just the decomposition of n, return to the quadratic example. It is easy to write down the first few terms in $\mathcal{T}^*$: they are {1, 2, 3, 5, …}.

Now follow the procedure. To construct the equilibrium structure when there are, say, six regions, it will suffice to compare $\alpha(1, (1, 5))$, which is 25.5, to $\alpha(6, (6))$, which is 18. (According to our result, it is not necessary to compute $\alpha(2, (2, 4))$, $\alpha(2, (1, 2, 3))$, or the average worth of a region in any of the several other coalition structures.) This shows that {1, 5} is the equilibrium coalition structure when $n = 6$. It also means that $6 \notin \mathcal{T}^*$. Thus, when $n = 7$, we need only compare $\alpha(2, (2, 5))$, which is 27, to $\alpha(7, (7))$, which is 24.5. This establishes that {2, 5} is the equilibrium coalition structure when $n = 7$ and, in particular, that $7 \notin \mathcal{T}^*$. Therefore, when $n = 8$, the comparison is to be made between $\alpha(3, (3, 5))$, which is 29.5, and $\alpha(8, (8))$, which is 32. This signals the return of the grand coalition: $8 \in \mathcal{T}^*$.

With this sort of argument, it is easy enough to proceed further and to show that the next two elements in $\mathcal{T}^*$ are 13 and 20.

Our decomposition achieves a significant reduction in the number of checks for an equilibrium coalition structure. What is established—and this is the part of the formal argument that requires work—is that regions that get the opportunity to move initially must attempt to create opportunities for the *largest* possible blocs to form in their wake. The idea of a decomposition captures this. For any integer n, its strict decomposition is obtained by first taking the largest possible integer in $\mathcal{T}^*$ (less than n), then taking the largest possible integer in $\mathcal{T}^*$ no less than the remainder, and thereafter repeating the same operation with the remainder (if any). Think of the initial coalition size that forms as the *last* integer in this decomposition. This captures (inductively) the best that a region can do if it were not to propose full cooperation. The payoff from this strict decomposition is then compared with the payoff from full cooperation, which is captured formally in the rule that includes n as a member of $\mathcal{T}^*$ (see [6]).

IV. How Much Inefficiency?

It should be clear from the discussion above that our model of coalition formation in the provision of public goods admits some inefficiency. For efficient outcomes are to be had only through the formation of the grand coalition; only a binding agreement among the set of *all* regions can cause all the effects of provision to be fully internalized.

Yet it should also be clear that there cannot be *too much* inefficiency. If a high degree of inefficiency were anticipated in equilibrium, then some region would move to make a suitable proposal to the grand coalition, and all the regions, fearing a huge loss if the proposal falls through, would accept. Therefore, a destruction of the fully efficient outcome can be achieved only at not too great a loss in efficiency. The purpose of this section is to make these ideas clearer through the use of our model.

We look at two efficiency criteria. First, we compare the total surplus generated by the equilibrium coalition structure (by simply adding up over all surpluses) to the surplus that would have been generated by the grand coalition were it to form. Second, we attempt to place a bound on the number of coalitions that can form in equilibrium. In both these comparisons we are particularly interested in the case of a large number of regions, where the problems of inefficiency are likely to be more severe.

A. *Full Efficiency, Sometimes*

Indeed, along an infinite subsequence of region populations, given by $\mathcal{T}^*$, the outcome *is* fully efficient.

Full efficiency also obtains if the number of regions does not exceed some upper bound. For instance, in the case of a quadratic cost function discussed in Section III, the grand coalition of regions must form when the number of regions is no more than three. To see how this is computed, return to the general case. Then the initial range of populations for which full efficiency obtains is just all the values of n for which a *single* region does not want to be on its own under the assumption that all the other regions stay together. (It is easy enough to verify the truth of this assertion by a recursive argument, which we omit.) This boils down to the condition that[18]

$$f(n-1) + g(1) \leq g(n). \tag{8}$$

To make this condition more transparent, assume that the cost function

[18] Of course, if (8) does not hold and $(n-1) \in \mathcal{T}^*$, then $n \notin \mathcal{T}^*$; full efficiency is not to be expected in general.

takes the constant elasticity form $c(z) = (1/\alpha)z^{\alpha}$ for some $\alpha > 1$. Then (8) is equivalent to the requirement that

$$\lambda(n-1)^{\lambda} - n^{\lambda} + 1 \leq 0,$$

where $\lambda \equiv \alpha/(\alpha - 1)$. It is easy enough to see that the largest n satisfying this condition is 3 for the quadratic case. (As another example, if $\alpha = 1.2$, the largest run of initial consecutive integers yielding efficiency is four.)

Efficiency is also to be had along a subsequence of n, but the argument here is more complicated. Now a group of "deviating regions" must attempt to predict the entire coalition structure that will be left behind after their defection and compare the gains from full cooperation with the return to be had from defection. With theorem 1 already obtained, it is easy to see that this condition is given by (6). Because this condition is met again and again along a subsequence (another way of saying this is that the set $\mathcal{T}^*$ is infinite), full efficiency must return (infinitely often) as n varies.

B. *The Extent of Inefficiency*

The discussion above also tells us that the degree of inefficiency cannot be too large in equilibrium: the formation of the grand coalition of regions is always an option. This allows us to place a lower bound on what we might call the *efficiency ratio:* the ratio of equilibrium surplus to the highest potential surplus.

Corollary 1 has an obvious implication for computing a lower bound on the degree of inefficiency: in the decomposition of n there must be one coalition with a size of at least $n/2$. This coalition generates a per capita surplus that is no less than $g(n/2)$, whereas every other coalition enjoys a per capita surplus of at least $f(n/2)$ (the output of the largest coalition). This, in turn, is at least as large as $g(n/2)$, so that the ratio of equilibrium to potential surplus is at least $g(n/2)/g(n)$. In the case of a quadratic cost function, this yields an "efficiency ratio" of at least 25 percent. With additional work, however, one can obtain a tighter bound, which we report in theorem 2 below. This theorem also provides an upper bound on the total number of coalitions that can form.

THEOREM 2. (1) For each n, let $e(n)$ denote the ratio of equilibrium to potential surplus. Then

$$e(n) > \frac{4}{3}\frac{g(n/2)}{g(n)}. \tag{9}$$

(2) If k is the number of equilibrium coalitions, then

$$k < \log_2 n + 1. \tag{10}$$

Part 1 of the theorem can be applied to the quadratic case to get a tighter lower bound on the efficiency ratio: one-third. Note that the bound is independent of n, the number of regions. Whether it is possible to obtain an improvement on this bound (which is also uniform in n) remains an open question.

Finally, part 2 of the theorem is a close descendant of corollary 1. In any decomposition, the size of each successive coalition (counting from the largest) must exceed half the number of the remaining regions. A little manipulation then reveals that the total number of regions must be (approximately) at least two raised to the power of the number of coalitions, which yields the required result. Note that this bound predicts substantial cooperation: for instance, no more than seven coalitions can form when there are 100 negotiating regions. Whether an upper bound on the number of coalitions can be given that is *independent* of the number of regions is an open question.

V. Discussion and Extensions

This admittedly stylized model raises a number of questions regarding robustness.

1. *How important is the particular model of negotiation used?*—What is central to the results in this paper is the idea that members of a coalition are influenced by the *average* and not the total payoff of that coalition. This was the explicit assumption in the restricted game. In the general game, the particular model we use yields the outcome that (as discount factors go to unity) a formed coalition must exactly divide its worth among its members. Any other model that has the same feature will do. It is, however, of some significance that in a well-specified noncooperative bargaining model, equal division turns out to be a result.

Another important implication of (both versions of) our coalition formation model is that it allows us to circumvent the usual problems of coordination failure. In particular, the completely noncooperative, Nash, outcome is not an equilibrium in our model. This is in sharp contrast to Dixit and Olson (1998), who rely on a much simpler model of coalition formation.[19] In the first stage of their game, agents decide,

[19] In the second stage of their model, agents who have formed a coalition act efficiently. In this respect their approach is the same as the one followed here and in Ray and Vohra (1997, 1999).

independently, whether or not to join one potential coalition.[20] There are always two pure-strategy equilibria in their model: one involving a complete coordination failure and another nonsymmetric equilibrium in which an efficient allocation emerges. They mainly focus on a symmetric mixed-strategy equilibrium, and they show that it leads to grossly inefficient outcomes in large economies. It is easy to see that our equilibrium concept applied to their public-goods model would yield a coalition structure in which one (last) coalition is just large enough to find it worthwhile to provide the public good, predicting full efficiency.

2. *Is the additive representation of payoffs important?*—We have chosen the payoff functions so that pollution control by any region (or regions) enters the return to a particular region in an additive way. The main implication of this assumption is that we can apply the decomposition rule to "subintegers"—subgroups of the original set of negotiators—regardless of which coalition structure has *already* formed. If external effects do not enter in an additive way, the particular structure of already formed coalitions will influence the structures that emerge among the remaining set of negotiators. For the recursive rule that we employ, this poses no problem at all. Ray and Vohra (1999) analyze these and other issues in a more abstract setting. However, it will become more difficult to provide a transparent characterization of equilibrium structure (the characterization in the additive case is complicated enough as it is).

3. *If binding agreements are possible, why does the Coase theorem not apply to generate full efficiency, regardless of the number of regions?*—This is an issue that requires careful discussion. Our model of negotiations permits coalitions to form freely and divide their worth freely, but *once formed, a coalition is not permitted to break up or expand, even if all its members unanimously agree to do so.* To understand the implications of this assumption, let us return to the quadratic case of Section III and focus on the situation with only four regions. Our model predicts that the coalition structure in this case is {1, 3}. However, consider the subgame in which the one-region coalition has already committed to forming. If the three-region coalition forms thereafter, the stand-alone region will receive a payoff of 9.5, whereas the three-region coalition averages 5.5. Now calculate the average worth of the four-region coalition; it is 8. So its *total* worth is 32. There is room, then, for the three remaining regions to make a mutually beneficial offer to the stand-alone region. For instance, the stand-alone region could be offered a payoff of 14. This would still leave a payoff of 18 for the remaining three regions: an average of 6,

[20] They study a model of a discrete public good with identical consumers. The decision to join a coalition is effectively a decision to undertake the efficient choice for the coalition, sharing the cost equally. Given a population of N consumers, one can assume that there is an integer $M < N$ such that a coalition of size M finds it profitable to provide the public good.

which exceeds the 5.5 that they would receive in the absence of full efficiency. Thus full efficiency should *finally* obtain.

Note the emphasis on the word "finally." For it is impossible to obtain this outcome *without* an intermediate stage in which the stand-alone region commits to standing alone. If an efficient proposal were made prior to this stage, then at least one region in that proposal must get no more than a payoff of 8. As soon as that region obtains the opportunity to respond to the going proposal, it will commit to standing alone (this guarantees a payoff of at least 9.5). Thus full efficiency requires the creation of an intermediate, inefficient stage in which a coalition *structure* comes into being.[21]

We can regard this situation in three different ways. First, a commitment is a commitment and *cannot* be reversed. This will be true of situations in which a commitment must be made by the use of concrete actions (not legal devices) that are prohibitively costly to reverse. For instance, pollution control might require the setting up of environmentally friendly factories that must be built from scratch. A region that does *not* take this route is committing to a reduced level of pollution control in a way that may be too costly to reverse (it may be setting up factories that are not built environmentally friendly, and moving to a greater level of control will require the tearing down of these fixed investments). In this case, our model may be taken to be a "literal" description of the coalition structure that is likely to emerge.

Second, a commitment may be costly but reversible (at some additional cost). As in the previous example, this might occur if pollution control devices are, by and large, modular, so that they can be tacked on to existing installations at moderate cost. In this situation, renegotiation is more likely to occur, but nevertheless the inefficient situation must first obtain (otherwise the beneficiaries of renegotiation will have no power to extract the surplus). In this case our prediction of a coalition structure must be broadened to a prediction of a *coalitional power structure,* which describes the initial constellation of groupings that must occur before a final efficient agreement comes about. Note, however, that efficiency can be said to obtain only in an "end state sense": the final outcome is efficient, but the negotiating process *must* be costly and involve inefficiencies.

Finally, it may be that a commitment, once made, is costless to reverse. This may sound a bit paradoxical (and perhaps it is), but in the interests of playing the devil's advocate, let us pursue this line of thought a bit further. In the context of our four-region example, it may be that the

[21] It can be seen that this observation is perfectly general and independent of *any* process of negotiation that we might choose to write down. A full treatment of the generality of this observation is beyond the scope of the current paper.

first region has access to an international court in which it makes the following declaration: that it commits to forming no binding relationship with the other regions *unless* the other regions are signatories to an agreement that gives it more than its stand-alone payoff. The qualification may then be used to "reverse the commitment," which was really a conditional commitment all along. In this case the final outcome must be truly efficient in the Coasian sense, but it is nevertheless still true that an intermediate power structure must first form before the efficient agreement is realized. Moreover, the final agreement will *not* involve equal division of the surplus. The division will mirror the endogenously generated coalitional power structure. We claim that even in this case (most conducive to efficiency), our model sheds light on this intermediate structure. For the coalitional structure that we predict must form before the final renegotiation (if any) is carried out.

Appendix

Proofs

We begin with an informal outline of the steps for theorem 1. First, we describe an algorithm that assigns a unique numerical coalition structure to each "population size" n. It will be clear that this algorithm picks out the subgame-perfect equilibrium of the restricted game. Proposition 1 (below) will show that this algorithm is identical to the decomposition of n. This will complete the proof of theorem 1 for the restricted game.

To establish the same result for the general game, we shall prove proposition 2 (below). This proposition shows that the numerical coalition structure of the algorithm satisfies a sufficient condition introduced in Ray and Vohra (1999) for that coalition structure to be the unique equilibrium (for discount factors close enough to one).

The algorithm now follows. Formally, to each integer n we assign a choice of integer $T(n) \in \{1, \ldots, n\}$. Applying T to n and then repeatedly to $n - T(n)$ will allow us to break up any integer n into a numerical coalition structure. Let $\mathbf{c}(n, T)$ denote this numerical structure.

STEP 1. Set $T(1) = 1$.

STEP 2. Recursively, for any integer $n > 1$, suppose that we have defined $T(m)$ for all $m = \{1, \ldots, n-1\}$. Choose $T(n)$ to be the *largest* integer t in $\{1, \ldots, n\}$ that maximizes $\alpha(t, t \cdot \mathbf{c}(n-t, T))$. Let $\alpha^*(n)$ be this maximum value.

STEP 3. Complete this recursive definition so that T is now defined on all the positive integers. Define a numerical coalition structure for a situation with n regions as $\mathbf{c}(n, T)$.

This completes the description of the algorithm, and we can now state the two main steps involved in proving theorem 1.

PROPOSITION 1. For any positive integer n, $\mathbf{c}(n, T) = \mathbf{d}(n)$.

Clearly, the restricted game possesses a subgame-perfect equilibrium. Moreover, given our tie-breaking convention, every such equilibrium involves a numerically unique coalition structure. It is also easy to see that this coalition structure is precisely $\mathbf{c}(n, T)$. Proposition 1 therefore completes the proof of theorem 1 for the restricted game.

To prove theorem 1 for the general game, we shall also need the following result.

PROPOSITION 2. For any positive integer n and any $t \in \{1, \ldots, n-1\}$ such that

$$\alpha(t, \mathbf{c}(n-t, T)) \geq \alpha(t', \mathbf{c}(n-t', T)) \quad \text{for all } t' \in \{1, \ldots, t\}, \tag{A1}$$

we have

$$\alpha^*(n) \geq \alpha^*(n-t) + f(t). \tag{A2}$$

Proposition 2 implies condition (6) of Ray and Vohra (1999). Since, by proposition 1, the algorithm and the decomposition yield the same coalition structure, it now follows from theorem 3.4 in Ray and Vohra (1999) that the decomposition characterizes the unique equilibrium coalition structure for the general game (for a sufficiently high discount factor).

To complete the proof of theorem 1, therefore, it remains to prove propositions 1 and 2.

Recall that, for each t, we define $g(t) \equiv f(t) - h(t)$. The following lemma collects some elementary observations regarding the functions f and g.

LEMMA 1. (i) If $g(t) + f(n-t) \geq g(s) + f(n-s)$ for some $1 \leq s \leq t < n$, then $g(t') + f(n-t') \geq g(t) + f(n-t)$ for all $t \leq t' < n$. (ii) If $t \geq s$, then $g(s) + f(t) \geq g(t) + f(s)$, with strict inequality whenever $t > s$.

Proof. As we have already noted in Section III*B*, g is a convex function since it is the value function for the maximization problem described in (2). By assumption 2, f is also convex. It then follows that the function $g(s) + f(n-s)$ must be convex in s, for $s \in [0, n]$. If a convex function is nondecreasing over some interval, it can never decrease thereafter. This proves part i.

To prove part ii, observe that (by assumption 1) optimal output per region $z(s)$ is strictly increasing in s, and therefore so is the cost $h(s)$. It follows that if $t \geq s$, then

$$g(s) + f(t) = f(s) + f(t) - h(s) \geq f(s) + f(t) - h(t) = g(t) + f(s),$$

with strict inequality whenever $t > s$, which proves part ii. Q.E.D.

The next lemma collects some elementary observations regarding decompositions of positive integers.

LEMMA 2. For any positive integer n, (i) if $\mathbf{d}(n) = t_1 \cdot \cdots \cdot t_k$, then, for all $t \in \{1, \ldots, t_1\}$, $\mathbf{d}(n-t) = \mathbf{d}(t_1 - t) \cdot t_2 \cdot \cdots \cdot t_k$. (ii) If t_k is the largest value in $\mathbf{s}(n)$, then $\mathbf{d}(m) = \mathbf{d}(m - t_k) \cdot t_k$ for all m such that $t_k \leq m < n$. (iii) The strict decomposition $\mathbf{s}(n) = \mathbf{d}(n-t) \cdot t$ whenever $t \in \mathbf{s}(n)$.

Proof. These observations follow directly from the definition of a decomposition. Q.E.D.

The following lemma is central to the main argument.

LEMMA 3. Suppose that n has strict decomposition $\mathbf{s}(n) = \{t_1, \ldots, t_k\}$. Then

$$\alpha(t_1, \mathbf{s}(n)) \geq \alpha(t, t \cdot \mathbf{d}(n-t)) \quad \text{for all } t \in \{1, \ldots, t_1 - 1\} \tag{A3}$$

and

$$\alpha(t_1, \mathbf{s}(n)) > \alpha(t, t \cdot \mathbf{d}(n-t)) \quad \text{for all } t \in \{t_1 + 1, \ldots, n-1\}. \tag{A4}$$

Proof. Proceed by induction. Clearly for $n = 2$ the assertion is trivially true because both the sets in question are empty. Suppose, then, that the lemma is fully established for all integers m that lie between 2 and $n-1$, and consider the lemma for n.

First, take $t \in \{1, \ldots, t_1 - 1\}$. Then $\mathbf{d}(n - t) = \mathbf{d}(t_1 - t) \cdot t_2 \cdot \cdots \cdot t_k$ (part i of lemma 2), so that

$$\begin{aligned}\alpha(t,\ t \cdot \mathbf{d}(n - t)) &= \alpha(t,\ t \cdot \mathbf{d}(t_1 - t) \cdot t_2 \cdot \cdots \cdot t_k)\\ &= \alpha(t,\ t \cdot \mathbf{d}(t_1 - t)) + \sum_{j=2}^{k} f(t_j)\\ &\leq \alpha(s_1,\ \mathbf{s}(t_1)) + \sum_{j=2}^{k} f(t_j)\\ &\leq f(t_1) - h(t_1) + \sum_{j=2}^{k} f(t_j)\\ &= \alpha(t_1,\ \mathbf{s}(n)),\end{aligned}$$

where s_1 is the first term in the strict decomposition of t_1,[22] and the inequality in that line holds by the induction hypothesis. The second inequality holds from (6) and the fact that $t_1 \in T^*$.

Next, take $t \in \{t_1 + 1, \ldots, n - 1\}$. There are now two subcases, each of which we consider in turn.

Subcase 1. $t_1 < t \leq n - t_k$. Note that in this case $k \geq 3$. From part ii of lemma 2, the decomposition of $n - t$ is given by $\mathbf{d}(n - t_k - t) \cdot t_k$, so that

$$\begin{aligned}\alpha(t,\ t \cdot \mathbf{d}(n - t)) &= \alpha(t,\ t \cdot \mathbf{d}(n - t_k - t) \cdot t_k)\\ &= \alpha(t,\ t \cdot \mathbf{d}(n - t_k - t)) + f(t_k). \qquad \text{(A5)}\end{aligned}$$

Since $k \geq 3$, it follows that $\mathbf{d}(n - t_k) = \mathbf{s}(n - t_k)$ and, by part iii of lemma 2, $\mathbf{d}(n - t_k) = t_1 \cdot \cdots \cdot t_{k-1}$. Applying the induction hypothesis to the integer $n - t_k$, we see that

$$\alpha(t,\ t \cdot \mathbf{d}(n - t_k - t)) < \alpha(t_1,\ \mathbf{s}(n - t_k)). \qquad \text{(A6)}$$

Combining (A5) and (A6) and using part iii of lemma 2 once again, we may conclude that

$$\alpha(t,\ t \cdot \mathbf{d}(n - t)) < \alpha(t_1,\ \mathbf{s}(n - t_k)) + f(t_k) = \alpha(t_1,\ \mathbf{s}(n)),$$

which completes the proof in this subcase.

Subcase 2. $t > n - t_k$. Suppose that (A4) does not hold for some $t > n - t_k$, that is,

$$\alpha(t,\ t \cdot \mathbf{d}(n - t)) \geq \alpha(t_1,\ \mathbf{s}(n)). \qquad \text{(A7)}$$

Recall that $g(t) \equiv f(t) - h(t)$ for all t. Notice that the highest possible average worth to a coalition of size t arises when the rest of the regions form one single coalition, that is, $\alpha(t,\ t \cdot \mathbf{d}(n - t)) \leq g(t) + f(n - t)$. Combining this information with (A7), we see that

$$g(t) + f(n - t) \geq \alpha(t_1,\ \mathbf{s}(n)). \qquad \text{(A8)}$$

From the previous subcase (or the definition of $\mathbf{s}(n)$ in the case $n = t_1 + t_k$) and from the fact that $\mathbf{d}(t_k) = t_k$, it follows that

[22] This is well defined because in the case under consideration, t_1 must be at least 2.

$$\alpha(t_1, \mathbf{s}(n)) \geq \alpha(n - t_k, (n - t_k) \cdot t_k) = g(n - t_k) + f(t_k). \tag{A9}$$

Combining (A8) and (A9), we have

$$g(t) + f(n - t) \geq g(n - t_k) + f(t_k).$$

Applying part i of lemma 1, we conclude that

$$g(n - 1) + f(1) \geq g(t) + f(n - t) \geq g(n - t_k) + f(t_k). \tag{A10}$$

Applying part ii of lemma 1 to the first and third expressions of the inequality (A10), we conclude that $n - t_k > 1$. This means that $n - 1 \notin T^*$. In other words, we can write $\mathbf{d}(n - 1) = (s_1 \cdot \cdots \cdot s_q)$, where $q \geq 2$.[23] Using (6), we conclude that

$$g(n - 1) < \alpha(s_1, \mathbf{d}(n - 1)) = g(s_1) + \sum_{j=2}^{q} f(s_j).$$

Since we know from part ii of lemma 2 that $g(1) + f(s_1) \geq g(s_1) + f(1)$, this inequality can be rewritten as

$$\begin{aligned} g(n - 1) + f(1) &< g(1) + \sum_{j=1}^{q} f(s_j) \\ &= \alpha(1, 1 \cdot \mathbf{d}(n - 1)) \\ &\leq \alpha(t_1, \mathbf{s}(n)). \end{aligned}$$

But this, along with (A10), implies

$$\alpha(t_1, \mathbf{s}(n)) > g(t) + f(n - t),$$

which contradicts (A8). Q.E.D.

Proof of Proposition 1

Note that this result is trivially true when $n = 1$, so assume inductively that, for some integer $n \geq 2$, $\mathbf{c}(m, T) = \mathbf{d}(m)$ for all $m = \{1, \ldots, n - 1\}$. All we need to show now is that $T(n)$ is just the first term in the decomposition of n. In other words, if $\mathbf{d}(n) = t_1 \cdots t_k$, we need to prove that

$$t_1 = \max\left\{\arg \max_{t \in \{1, \ldots, n\}} \alpha(t, t \cdot \mathbf{c}(n - t, T))\right\}. \tag{A11}$$

Let $\mathbf{s}(n) = s_1 \cdots s_q$. From lemma 3 we know that

$$s_1 = \max\left\{\arg \max_{t \in \{1, \ldots, n-1\}} \alpha(t, t \cdot \mathbf{c}(n - t, T))\right\}. \tag{A12}$$

Consider the two following cases: (1) Suppose $s_1 = t_1$. This means that n does not satisfy (6), that is,

$$\alpha(t_1, t_1 \cdot \mathbf{d}(n - t_1)) > \alpha(n, n).$$

Since, by the induction hypothesis, $\mathbf{c}(n - t_1, T) = \mathbf{d}(n - t_1)$ and $s_1 = t_1$, this

[23] In fact we know that if $t_1 = 1$, then $\mathbf{d}(n - 1) = (t_2 \cdot \cdots \cdot t_k)$, and if $t_1 > 1$, then $\mathbf{d}(n - 1) = [\mathbf{d}(t_1 - 1) \cdot t_2 \cdot \cdots \cdot t_k]$. But this need not concern us in what follows.

along with (A12) implies (A11). (2) Suppose $t_1 > s_1$. This means that $t_1 = n$ and that n satisfies (6), that is,

$$\alpha(t_1,\ \mathbf{d}(n)) = \alpha(n,\ n) \geq \alpha(s_1,\ s_1 \cdot \mathbf{d}(n - s_1)),$$

and again (A12) implies (A11). Q.E.D.

The following lemma is an intermediate step in the proof of proposition 2.

LEMMA 4. For any positive integer m, let $\phi(m)$ be the first term in the strict decomposition of m. Let ϕ^k denote the k-fold composition of ϕ. Then if $t \in \{1, \ldots, n-1\}$ satisfies (A1), $t = \phi^k(n)$ for some integer k.

Proof. Suppose that t satisfies (A1) but the conclusion of the lemma is false. Noting that $\phi^k(n) = 1$ after some finite k, let s be the largest integer of the form $\phi^k(n)$ such that s is smaller than t. Use the convention $\phi^0(n) = n$. Then

$$s = \phi^k(n) < t < \phi^{k-1}(n) \equiv m.$$

It is easy to see, from lemma 2, that because $m = \phi^{k-1}(n)$,

$$\mathbf{d}(n - t') = \mathbf{d}(m - t') \cdot \mathbf{d}(n - m) \quad \text{for all } t' < m. \tag{A13}$$

Since s is the first term in the strict decomposition of m, it follows from lemma 2 that

$$s \cdot \mathbf{d}(n - s) = s \cdot \mathbf{d}(m - s) \cdot \mathbf{d}(n - m) = \mathbf{s}(m) \cdot \mathbf{d}(n - m).$$

Thus

$$\begin{aligned}\alpha(s,\ s \cdot \mathbf{d}(n - s)) &= \alpha(s,\ \mathbf{s}(m - s)) + \sum_{t' \in \mathbf{d}(n-m)} f(t') \\ &> \alpha(t,\ t \cdot \mathbf{d}(m - t)) + \sum_{t' \in \mathbf{d}(n-m)} f(t') \\ &= \alpha(t,\ t \cdot \mathbf{d}(n - t)),\end{aligned}$$

where the inequality uses (A4) of lemma 3 applied to the integer m. Since $\mathbf{c}(n,\ T) = \mathbf{d}(n)$ for every n, we may conclude that

$$\alpha(s,\ s \cdot \mathbf{c}(n - s,\ T)) > \alpha(t,\ t \cdot \mathbf{c}(n - t,\ T)),$$

which contradicts the presumption that t satisfies (A1). Q.E.D.

Proof of Proposition 2

Pick any integer $t \in \{1, \ldots, n\}$ satisfying (A1). We must show that (A2) holds. By lemma 4, $t = \phi^k(n)$ for some $k \geq 1$. Let $m = \phi^{k-1}(n)$ (to be interpreted, by convention, as n if $k = 1$). Then, if s_2 is the *second* term in the strict decomposition of m (t being the first),

$$\alpha(t,\ \mathbf{s}(m)) > \alpha(s_2,\ \mathbf{s}(m)),$$

by virtue of the simple fact that $t < s_2$ and by part ii of lemma 1. Using (A13), we have

$$\begin{aligned}\alpha(t,\ t \cdot \mathbf{d}(n - t)) &= \alpha(t,\ \mathbf{s}(m) \cdot \mathbf{d}(n - m)) \\ &> \alpha(s_2,\ \mathbf{s}(m) \cdot \mathbf{d}(n - m)).\end{aligned} \tag{A14}$$

Since $\mathbf{s}(m) = t \cdot \mathbf{d}(m - t)$, we can appeal to (A13) to assert that

$$\alpha(s_2,\ \mathbf{s}(m)\cdot\mathbf{d}(n-m)) = \alpha(s_2,\ t\cdot\mathbf{d}(m-t)\cdot\mathbf{d}(n-m))$$
$$= \alpha(s_2,\ \mathbf{d}(n-t)) + f(t). \tag{A15}$$

Since $\mathbf{c}(n,\ T) = \mathbf{d}(n)$ for all n, it follows from the definition of α^* that

$$\alpha(s_2,\ \mathbf{d}(n-t)) = \alpha^*(n-t). \tag{A16}$$

Combining equations (A14), (A15), and (A16), we see that

$$\alpha(t,\ t\cdot\mathbf{d}(n-t)) > \alpha^*(n-t) + f(t). \tag{A17}$$

On the other hand, the definition of α^* implies that

$$\alpha^*(n) \geq \alpha(t,\ t\cdot\mathbf{c}(n-t,\ T)) = \alpha(t,\ t\cdot\mathbf{d}(n-t)). \tag{A18}$$

Combining (A17) and (A18), we conclude that

$$\alpha^*(n) \geq \alpha^*(n-t) + f(t),$$

which establishes (A2). Q.E.D.

Proof of Theorem 2

Let the decomposition of n be denoted by $\{t_1, \ldots, t_k\}$. (We know that this is the equilibrium coalition structure as well.)

First, we establish part 1 of the theorem. Using the negation of (6), we see that, for each i,

$$\alpha(t_i,\ \mathbf{d}(n)) = \sum_{j=1}^{i-1} f(t_j) + \alpha\left(t_i,\ \mathbf{d}\left(n-\sum_{j=1}^{i-1} t_j\right)\right)$$
$$\geq \sum_{j=1}^{i-1} f(t_j) + g\left(\sum_{j=i}^{k} t_j\right)$$
$$\geq g\left(\sum_{j=i}^{k} t_j\right). \tag{A19}$$

Now observe that the per capita surplus generated by the decomposition is simply $(1/n)\sum_{i=1}^{k} t_i\alpha(t_i,\ \mathbf{d}(n))$. Expression (A19) permits us to describe a lower bound on this surplus:

$$\frac{1}{n}\sum_{i=1}^{k} t_i\alpha(t_i,\ \mathbf{d}(n)) \geq \frac{1}{n}\sum_{i=1}^{k} t_i g\left(\sum_{j=i}^{k} t_j\right). \tag{A20}$$

The remainder of the proof looks for a lower bound to this expression that is free of endogenous terms such as the particular decomposition of n.

Notice that we can view g as a smooth convex function defined on any nonnegative real n (and not just the positive integers). For *any* sequence of nonnegative numbers $\{n_i\}_{i=0}^{\infty}$ that sum to n, define $S_i \equiv \sum_{j=1}^{i} n_j$ for all $i \geq 1$. The convexity of g implies that

$$\sum_{i=0}^{\infty} n_i g\left(\sum_{j=0}^{i} n_j\right) = n_0 g(n_0) + \sum_{i=0}^{\infty} n_i g(n_0 + S_i)$$

$$\geq n_0 g(n_0) + \sum_{i=1}^{\infty} n_i [g(n_0) + S_i g'(n_0)]$$

$$= ng(n_0) + g'(n_0) \sum_{i=1}^{\infty} n_i S_i. \tag{A21}$$

It is possible to show that[24]

$$\sum_{i=1}^{\infty} n_i S_i \geq \tfrac{2}{3}(n - n_0)^2. \tag{A22}$$

Combining (A21) and (A22), we conclude that

$$\sum_{i=0}^{\infty} n_i g\left(\sum_{j=0}^{i} n_j\right) \geq ng(n_0) + \tfrac{2}{3}(n - n_0)^2 g'(n_0)$$

$$\geq g(n_0)\left[n + \frac{2}{3n_0}(n - n_0)^2\right], \tag{A23}$$

where the second inequality uses the fact that $g'(n_0) \geq g(n_0)/n_0$ by convexity of g. Now consider the function

$$y(n_0) \equiv g(n_0)\left[n + \frac{2}{3n_0}(n - n_0)^2\right].$$

It is easy to check that $y'(n_0) > 0$ for all $n_0 \geq n/2$ (see n. 24). Using this information in (A23), we may conclude that

$$\sum_{i=0}^{\infty} n_i g\left(\sum_{j=0}^{i} n_j\right) > g\left(\frac{n}{2}\right)\left(n + \frac{2}{3}\frac{n^2/4}{n/2}\right)$$

$$= \frac{4n}{3} g\left(\frac{n}{2}\right) \tag{A24}$$

whenever $n_0 > n/2$.

Let $\mathbf{d}(n) = \{t_1, t_2, \ldots, t_k\}$, and define a particular sequence $\{n_i^*\}_{i=0}^{\infty}$ by $n_i^* = t_{k-i}$ for all $i = 0, \ldots, k-1$ and $n_i^* = 0$ for all $i \geq k$. Applying corollary 1, we note that $n_0^* > n/2$. Therefore, (A24) applies to the sequence $\{n_i^*\}_{i=0}^{\infty}$. Combining this information with (A20), we see that

$$\frac{1}{n}\sum_{i=1}^{k} t_i \alpha(t_i, \mathbf{d}(n)) \geq \frac{1}{n}\sum_{i=1}^{k} t_i g\left(\sum_{j=i}^{k} t_j\right)$$

$$= \frac{1}{n}\sum_{i=0}^{\infty} n_i^* g\left(\sum_{j=0}^{i} n_j^*\right)$$

$$> \frac{4}{3} g\left(\frac{n}{2}\right).$$

[24] Details are available at http://econ.pstc.brown.edu/~/rvohra/papers/details98-24.pdf.

We may therefore conclude that

$$e(n) = \frac{(1/n)\sum_{i=1}^{k} t_i \alpha(t_i, \mathbf{d}(n))}{g(n)} > \frac{4}{3}\frac{g(n/2)}{g(n)}.$$

Now we establish part 2 of the theorem. Let $\mathbf{d}(n) = (t_1, \ldots, t_k)$ and define $R_i \equiv \sum_{j=1}^{i} t_j$. By corollary 1, we know that

$$t_i > \tfrac{1}{2}R_i \quad \text{for all } i = 1, \ldots, k. \tag{A25}$$

Furthermore,

$$R_{i+1} = R_i + t_{i+1}. \tag{A26}$$

Combining (A26) with (A25) for index $i + 1$, we may conclude that $R_{i+1} > 2R_i$ or that $R_i > 2^{i-1}$. Now observe that $R_k = n$ to conclude that (10) holds. Q.E.D.

References

Alesina, Alberto, and Spolaore, Enrico. "On the Number and Size of Nations." *Q.J.E.* 112 (November 1997): 1027–56.

Bergstrom, Theodore; Blume, Lawrence; and Varian, Hal. "On the Private Provision of Public Goods." *J. Public Econ.* 29 (February 1986): 25–49.

Bloch, Francis. "Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division." *Games and Econ. Behavior* 14 (May 1996): 90–123.

———. "Non-cooperative Models of Coalition Formation in Games with Spillovers." In *New Directions in the Economic Theory of the Environment,* edited by Carlo Carraro and Domenico Siniscalco. Cambridge: Cambridge Univ. Press, 1997.

Carraro, Carlo, and Siniscalco, Domenico. "Strategies for the International Protection of the Environment." *J. Public Econ.* 52 (October 1993): 309–28.

Champsaur, Paul; Roberts, Donald John; and Rosenthal, Robert W. "On Cores in Economies with Public Goods." *Internat. Econ. Rev.* 16 (October 1975): 751–64.

Chander, Parkash, and Tulkens, Henry. "The Core of an Economy with Multilateral Environmental Externalities." *Internat. J. Game Theory* 26, no. 3 (1997): 379–401.

Chatterjee, Kalyan; Dutta, Bhaskar; Ray, Debraj; and Sengupta, Kunal. "A Noncooperative Theory of Coalitional Bargaining." *Rev. Econ. Studies* 60 (April 1993): 463–77.

Chwe, Michael Suk-Young. "Farsighted Coalitional Stability." *J. Econ. Theory* 63 (August 1994): 299–325.

Clarke, Edward H. "Multipart Pricing of Public Goods." *Public Choice* 11 (Fall 1971): 17–33.

Coase, Ronald H. "The Problem of Social Cost." *J. Law and Econ.* 3 (October 1960): 1–44.

Dixit, Avinash, and Olson, Mancur. "Does Voluntary Participation Undermine the Coase Theorem?" Manuscript. Princeton, N.J.: Princeton Univ., Dept. Econ., 1998.

Foley, Duncan K. "Lindahl's Solution and the Core of an Economy with Public Goods." *Econometrica* 38 (January 1970): 66–72.

Green, Jerry R., and Laffont, Jean-Jacques. *Incentives in Public Decision-Making.* Amsterdam: North-Holland, 1979.

Groves, Theodore. "Incentives in Teams." *Econometrica* 41 (July 1973): 617–31.

Huang, Chen-Ying, and Sjostrom, Tomas. "Consistent Solutions for Cooperative Games with Externalities." Manuscript. University Park: Pennsylvania State Univ., Dept. Econ., 1999.

Laffont, Jean-Jacques. "Incentives and the Allocation of Public Goods." In *Handbook of Public Economics,* vol. 2, edited by Alan J. Auerbach and Martin Feldstein. Amsterdam: North-Holland, 1987.

Lindahl, Erik. "Just Taxation—a Positive Solution." 1919. Reprinted in *Classics in the Theory of Public Finance,* edited by Richard A. Musgrave and Alan T. Peacock. New York: St. Martin's Press (for Internat. Econ. Assoc.), 1967.

Ray, Debraj, and Vohra, Rajiv. "Equilibrium Binding Agreements." *J. Econ. Theory* 73 (March 1997): 30–78.

———. "A Theory of Endogenous Coalition Structures." *Games and Econ. Behavior* 26 (February 1999): 286–336.

Richter, Donald K. "The Core of a Public Goods Economy." *Internat. Econ. Rev.* 15 (February 1974): 131–42.

Roberts, Donald John. "The Lindahl Solution for Economies with Public Goods." *J. Public Econ.* 3 (February 1974): 23–42.

Rosenthal, Robert W. "External Economies and Cores." *J. Econ. Theory* 3 (June 1971): 182–88.

Rubinstein, Ariel. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50 (January 1982): 97–109.

Samuelson, Paul A. "The Pure Theory of Public Expenditure." *Rev. Econ. and Statis.* 36 (November 1954): 387–89.

Sonnenschein, Hugo. "The Economics of Incentives: An Introductory Account." In *Frontiers of Research in Economic Theory: The Nancy L. Schwartz Memorial Lectures, 1983–1997,* edited by Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien. Cambridge: Cambridge Univ. Press, 1998.

Yi, Sang-Seung. "Endogenous Formation of Customs Unions under Imperfect Competition: Open Regionalism Is Good." *J. Internat. Econ.* 41 (August 1996): 153–77.