Self-Enforcing Voting in International Organizations

October 7, 2005

Abstract

Some international organizations are governed by unanimity rule, others by some kind of majority system. The existing voting models, which generally assume that decisions made by voting are perfectly enforceable, have a difficult time explaining the observed variation in governance mode, and in particular the widespread occurrence of the unanimity system. We present a model whose main departure from standard voting models is that there is no external enforcement mechanism: each country is sovereign and cannot be forced to follow the collective decision, or in other words, the voting system must be self-enforcing. The model yields unanimity as the optimal system for a wide range of parameters, and delivers rich predictions on the variation in the mode of governance across organizations.

1 Introduction

Most international organizations lack an external enforcement mechanism. In particular, if an organization relies on a voting system to make decisions, a government cannot be forced to comply with the collective decision. It will do so only if the short-term gain from defecting is outweighed by the future loss of cooperation. Motivated by this observation, in this paper we propose a theory of self-enforcing voting systems.

In the real world of international organizations, there is a wide variation in the mode of governance, both across organizations and over time. Some organizations, such as NATO, WTO and Mercosur, are governed by unanimity rule.[1] Others, such as most United Nations agencies, are governed by simple or qualified majority rules. Still others have seen changes of governance mode over time: for example, the European Union has recently switched from unanimity to qualified majority in several policy areas, and the International Standards Organization has switched from unanimity to a supermajority rule in the 1970s. It is important to note that the unanimity rule is qualitatively different from any non-unanimous rule: the former requires only *coordination*; the latter requires also *enforcement*, to keep in check the dissenting members' temptation to defect.

There is a vast theoretical literature on voting systems, but most of the existing models share the assumption that the outcome of the vote can be perfectly enforced. These enforceable-voting models have a difficult time explaining the above-mentioned variation in governance mode, and in particular the frequent occurrence of the unanimity system. We will present a simple framework whose main departure from standard voting models is the presence of a self-enforcement constraint. This model yields unanimity as the optimal system for a wide range of parameters, and yields rich predictions on the determinants of the optimal mode of governance.

Next we preview the structure of the model and the main results.

We consider an infinite-horizon game where, at the outset, governments anticipate that there will be a sequence of binary collective choices. In each instance, one alternative will be the status quo and the other will be some collective action. The collective action is effective only if all members participate. *Ex ante,* each member attaches some probability to the event that she will be in favor or against changing the status quo for each future issue. Members' preferences on future issues can be correlated.

[1] For the WTO, of course this statement applies only to rule-making activities, not to the dispute settlement system, which is concerned with the enforcement of the agreed-upon rules.

The voting rule is chosen ex ante, under a veil of ignorance about future issues. Thus the optimal voting rule maximizes the ex-ante expected utility of the representative member subject to a self-enforcement constraint: a government must have incentive to comply with the collective decision even if it happens to disagree with it. This requires that the future gains from cooperation outweigh the one-time gain from defecting.

A key parameter in the model is the governments' discount factor. We show that, if the discount factor is higher than some critical level, the best self-enforcing governance mode is the first-best voting rule, which in this context is typically some non-unanimous rule. But if the discount factor is lower than this critical level, the best self-enforcing governance mode is the unanimity system. The discount factor can be interpreted as capturing not only the players' pure time preferences, but also the probability that a player will still be in the game next period, and the frequency with which the organization makes decisions. Thus, our model predicts that a non-unanimous rule is more likely to be adopted in organizations where governments are more stable, and in "busier" organizations.

Another important parameter in the model is the correlation among members' preferences, that is the likelihood that members will agree on future issues. One might expect that higher correlation favors unanimity over other voting rules, but we find that the opposite is true: a higher degree of correlation expands the range of discount factors for which the first-best rule is sustainable. The model thus predicts that a non-unanimous rule is more likely to be adopted in more homogeneous organizations.

In reality, a number of international organizations have different voting rules for different types of issues. For example, the European Union applies the unanimity rule for particularly sensitive issues, and a (simple or qualified) majority rule for more "technical" issues. Our model suggests a theoretical explanation for this kind of dual decision making systems. We consider an organization that makes decisions on two types of issues, high-stake issues and low-stake issues, and find that for intermediate values of the discount factor the optimal voting rule is unanimity for high-stake issues and the first-best rule for low-stake issues.

Next we consider the role of international transfers. Transfers can make it easier to satisfy the self-enforcement constraint, because they can be used to mitigate the dissenting members' temptation to defect. We show that the availability of transfers expands the range of discount factors for which the first-best rule is sustainable. Thus the model suggests that we should be more likely to observe a non-unanimous rule in organizations that have the flexibility to enact

monetary transfers among its members. However, we also find that transfers cannot completely solve the enforceability problem: if the discount factor is low enough, unanimity remains the best self-enforcing rule. This suggests that enforcement issues can explain the unanimity rule even when monetary transfers are available.

Finally we consider an extension of the model where the collective action may be effective even if not all members participate (the case of "impure" collective action). We show that, under some plausible conditions, the main results of the model continue to hold, with one important difference: in some states of the world it may be optimal to involve in the collective action only those countries that are in favor of it (a "coalition of the willing"), without requiring the participation of the dissenting members.

We should emphasize that the model takes the membership of the organization as given. The question of what determines the membership of an international organization is an important one but is beyond the scope of this paper. We view our model as a first and necessary step toward a more general theory of self-enforcing institutions with endogenous membership.[2]

Our paper contributes to two literatures. The first one is the literature on self-enforcing international agreements. To the best of our knowledge, all the models in this class are repeated-game models where there is no scope for voting.[3] Our innovation with respect to this literature is that we consider a multilateral repeated game where it is efficient to make decisions by voting. This is because players have private information about their preferences, and a voting scheme can be used to aggregate information and make efficient collective choices.

Second, our paper contributes to the literature on social choice and voting. All the voting models that we are aware of ignore the enforceability problem. For this reason, these models are useful to examine issues of domestic institutions and constitutional design, but their applicability to international organizations is limited.

In this literature, a paper that is related to ours is Barbera and Jackson [7]. They consider a binary collective choice model where members' preferences on future issues are uncertain, and each player is characterized by a distinct probability of being in favor of the status quo.

[2] In our working paper version (Maggi and Morelli [20]) we consider a simple extension of the model where the size of the organization is endogenous. There we assume that in every period there is a random number of new candidates for membership, and current members choose whether to admit the new candidates. In that setting we show that, under some conditions, the optimal self-enforcing voting rule is unanimity up to some (random) date and then switches to a majority rule.

[3] For models of self-enforcing trade agreements, see for example the survey by Staiger [26]. For models of international lending, see for example the survey by Eaton and Fernandez [14].

They study self-stable voting rules, i.e. voting rules such that there is no alternative rule that would beat the given voting rule if the given voting rule is used to choose between the rules. Our main departures from Barbera and Jackson's model are that (i) we examine self-enforcing voting rules, whereas they assume perfect enforcement, and (ii) we assume that the voting rule is chosen under a veil of ignorance, so that in our case the natural criterion to select a voting rule is the maximization of the members' common ex-ante utility. Another paper that is related to ours is Ledyard and Palfrey [17]. They study a situation in which a group of individuals must decide whether to produce a discrete public good and how to pay for it, and each individual's preferences may be of two types. Among other things, they show that an efficient public good decision can be achieved by a majority voting rule.[4] They do not consider the implications of repeated interaction for the optimal mechanism.[5]

Our theory provides a new rationale for the unanimity rule, which is the lack of enforceability. This is certainly not the first attempt to rationalize the use of the unanimity rule. The classic contributions by Wicksell [28] and Buchanan and Tullock [9] proposed a simple argument in favor of unanimity. Their argument was based on an ex-post Pareto-efficiency criterion: unanimity is the only rule under which collective action is taken only if it is a Pareto-improvement over the status quo. In contrast, we adopt an ex-ante efficiency criterion within a veil-of-ignorance setting. In this setting, if external enforcement is available, the ex-ante efficient rule is generally a non-unanimous rule,[6] but unanimity may become optimal if there is no external enforcement.

The paper is organized as follows. In section 2 we present the static model. First we solve for the first-best outcome (section 2.1), then we characterize the equilibria of the one-shot game without enforcement (section 2.2), and finally we consider the one-shot game when external

[4] Ledyard and Palfrey [18] show that, under plausible conditions, simple binary voting is asymptotically efficient as the number of voters becomes large, even when voters' preferences can take a continuum of values.

[5] In the literature there are some papers that are concerned with voting in dynamic environments. Messner and Polborn [22] consider an overlapping-generations model of voting on projects that require upfront investments and yield delayed benefits. Carrubba and Volden [11] examine the optimal choice of voting rule in a model of repeated logrolling in legislative institutions. Roberts [24] and Barbera, Maschler and Shalev [8] study the dynamics of an organization in which current members have heterogeneous preferences about the admission of new members, and vote on admissions in every period. All of these papers however focus on very different questions than the one considered here, and all assume perfect enforcement of the outcome of the vote.

[6] Also Aghion and Bolton [3] and Guttman [16] argue that, in a veil-of-ignorance setting with perfect enforcement, unanimity is typically dominated by some non-unanimous rule from the standpoint of ex-ante efficiency. Other papers that examine the optimal choice of voting rule are May [21], Rae [23], Taylor [27], Caplin and Nalebuff [10], Austen-Smith and Banks [6] and Dasgupta and Maskin [13]. All of these models assume that the outcome of the vote is perfectly enforceable.

enforcement is available (section 2.3). In section 3 we present the repeated-game version of the model. In section 3.1 we characterize the optimal self-enforcing voting rule, and examine how it depends on the players' discount factor and the degree of correlation in preferences. In section 3.2 we examine the case where the organization may face low-stake issues or high-stake issues. In section 3.3 we consider the role of international transfers. In section 3.4 we extend the analysis to the case of "impure" collective action. In section 4 we discuss the interpretation of our results and their robustness to some extensions of the model.

2 The Static Model

Consider an organization with N members. Each member chooses a binary action, $a^i \in \{0, 1\}$ $(i = 1, ..., N)$. Taking the action $(a^i = 1)$ is interpreted as participating in a collective action, such as going to war, or adopting a common currency, adopting or modifying a common immigration, taxation, agricultural, or trade policy, or harmonizing a standard. Not taking the action $(a^i = 0)$ is interpreted as preserving the status quo. This model takes the organization membership N as given. The question of the endogenous determination of the organization membership is important but outside the scope of this paper.

We assume that the collective action is effective only if all members participate, otherwise the status quo is kept. In particular, each of the N players receives a positive benefit B if $a^i = 1$ for all i, and zero benefit otherwise. We will often refer to this case as "pure" collective action. In a later section we will discuss the case in which the collective action may be effective even if some of the members do not participate (the case of "impure" collective action). For the moment we note that there are situations in reality for which the assumption of pure collective action is not unrealistic. Consider for example an economic union where goods and factors move freely across countries, and suppose the union decides to tighten its immigration policy vis-a-vis outside countries. This policy change requires the participation of all the member countries: if one country fails to patrol its borders, this will completely undermine the collective action of the union. Another example is given by trade policies in a customs union. If the union decides to increase the common external tariff and one member country does not go along, the effects of the tariff hike will be undone. More generally, we think that many collective action problems in international organizations are characterized by strong coordination economies, and the assumption of pure collective action is a good approximation for situations of this kind.

For each member, participating in the collective action is costly. For some members the cost is lower than the benefit, but for others the cost exceeds the benefit. This is a simple way of capturing situations where the members' interests over the collective action may diverge. Formally, we assume that player i's cost of action θ^i takes value θ_L or θ_H, with $\theta_L < B < \theta_H$. Thus, a low-θ member is in favor of the collective action, a high-θ member is against it.[7] The parameter θ^i is player i's private information. This can be interpreted as the economic or the political cost of changing the status quo for country i.[8]

To summarize, player i has the following utility function:

$$U(a^i, n, \theta^i) = B \cdot I_{[n=N]} - a^i \theta^i \tag{1}$$

where $n \in \{a^i, ..., N - 1 + a^i\}$ denotes the total number of members taking action.[9]

What we have described so far is the *ex post* stage of the model. We now step back to an ex-ante perspective. Ex ante, players are under a veil of ignorance about future issues. The idea is that the nature of future issues is uncertain, and therefore each player does not know which side of the issue she will be on. Even for the issues of well-known nature, like raising or lowering a tax rate or an interest rate, a member will be in favor or against the change depending on the realization of the political and economic state of the world for the country he or she represents in the organization. We capture this idea by assuming that at the ex-ante stage $\boldsymbol{\theta} = (\theta^1, ..., \theta^N)$ is a random vector distributed according to the common-knowledge probability distribution $P(\boldsymbol{\theta})$ over support $\boldsymbol{\Theta} = \{\theta_L, \theta_H\}^N$. This distribution is symmetric with respect to its N arguments, which implies that the N players are ex-ante symmetric with respect to the future issue. We

[7] The assumption of two types (which is relatively common in the literature on optimal voting rules; see for example Barbera and Jackson [7] and Ledyard and Palfrey [17]) simplifies the analysis because it allows us to abstract from issues of preference intensity: all the players in favor of collective action feel equally strongly about it, and the same is true for those who are opposed to it. Given this assumption, a simple binary voting rule is sufficient to communicate all the relevant information. If there were more than two cost levels, one would need a more complicated mechanism to elicit all the relevant information. But we note that, within the class of binary voting rules, our results are robust to a more general type space.

[8] An equivalent assumption would be that the cost is common and the benefits are private information.

[9] We have assumed that, if member i takes action ($a^i = 1$), he incurs cost θ^i regardless of the other members' actions. This assumption can be weakened substantially: we only need to assume that a small fraction $\epsilon > 0$ of the cost is incurred regardless. More formally, we can generalize the utility function to $U = [B - (1 - \varepsilon)\theta^i a^i] \cdot I_{[n=N]} - \varepsilon\theta^i a^i$. The interpretation is that, if the collective action is not undertaken ($n < N$), member i can recover a fraction $(1 - \varepsilon)$ of the cost, while a fraction ε of the cost cannot be recovered. For any $\varepsilon \in (0, 1]$, our results hold exactly as stated. An alternative setting that would yield the same results is the following two-stage game. In the first stage, players decide whether to participate in the collective action. In the second stage, each player can confirm or reverse the decision, but in the latter case he incurs a small cost. This could be thought of as a "ratification" game, where not ratifying the initial decision implies a small political cost.

can think of $\boldsymbol{\theta}$ as summarizing the relevant state of the world. In a later section we will discuss more thoroughly the veil of ignorance assumption and how results are likely to change if players are ex-ante asymmetric.

2.1 First-best Outcome

The symmetric first-best outcome is the mapping from states to actions that maximizes the members' common expected utility, or in other words, the ex-ante Pareto-efficient outcome that gives the same expected utility to all members. Our focus on the symmetric first best outcome seems natural given that players are ex-ante symmetric. For simplicity, in what follows we will simply speak of "first best," omitting the qualifier "symmetric."

To characterize the first best outcome, let $N_1(\boldsymbol{\theta})$ denote the number of members that have a low cost realization (and hence are in favor of the collective action). Also, let $\lceil x \rceil$ denote the smallest integer greater than or equal to x.

Proposition 1 *The first-best outcome is:* $a^i = 1$ *for all* i *if* $N_1(\boldsymbol{\theta}) \geq q^*$, $a^i = 0$ *for all* i *if* $N_1(\boldsymbol{\theta}) < q^*$, *where* $q^* \equiv \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil$.

Proof. Given our assumptions on payoffs, we can focus on two vectors of actions, the one where everyone takes the action and the one where nobody does. We can then formulate the problem as choosing a mapping from the state of the world $\boldsymbol{\theta}$ to a collective action $a \in \{0, 1\}$. Given that players are ex-ante identical, we can maximize the members' aggregate expected utility, that is

$$\max_{a(\boldsymbol{\theta})} \sum_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P(\boldsymbol{\theta}) a(\boldsymbol{\theta}) [N_1(\boldsymbol{\theta})(B - \theta_L) + (N - N_1(\boldsymbol{\theta}))(B - \theta_H)]$$

Clearly, it is optimal to take the collective action in all the states where its aggregate benefit, $B \cdot N$, exceeds its aggregate cost, $N_1(\boldsymbol{\theta})\theta_L + (N - N_1(\boldsymbol{\theta}))\theta_H$. This implies that it is efficient to take the collective action if and only if N_1 exceeds the quota q^*. **QED.**

Note that ex-ante efficiency generally requires some players to act against their own interest ex-post. A simple two-player example can illustrate this point. Suppose $B = 1$, $\theta_L = .5$ and $\theta_H = 1.2$. Then, from an ex-ante point of view, it is desirable for both players to take the action whenever one of them would like to. To see this, note that maximizing the players' common ex-ante utility is equivalent to maximizing the sum of their utilities in each state. Consider a

state in which the players disagree, that is one player has cost .5 and the other has cost 1.2. If they both take the action, the joint payoff is $(1 - .5) + (1 - 1.2) = .3$, whereas the alternative is zero, therefore both should take the action.

2.2 One-Shot Game without Enforcement

Let us consider the basic game in which the organization members choose their actions a^i only once, and no external enforcement is available.

Since players have private information, it is compelling to allow for communication before actions are chosen. A natural way to introduce communication in this context is to consider the following timing: after observing her type θ^i, each player simultaneously sends a public message $V^i \in \{\theta_L, \theta_H\}$; then players simultaneously choose actions. We interpret $V^i = \theta_L$ as a vote in support of collective action (a "yes" vote), and $V^i = \theta_H$ as a "no" vote.

A natural equilibrium notion for this kind of game is that of Perfect Bayesian Equilibrium. The game admits multiple equilibria. We are interested in characterizing the "best" equilibrium, i.e. the one that maximizes the players' common ex-ante utility, and the "worst" equilibrium, i.e., the one that gives players the lowest ex-ante utility. The best equilibrium is interesting because it represents an upper bound to what players can accomplish without the help of external enforcement or reputation mechanisms. The worst equilibrium will be important as a punishment when we analyze the repeated game.

The worst equilibrium is one in which messages are ignored and the status quo is never changed: $a^i = 0$ for all i regardless of the state. This is clearly an equilibrium: knowing that no one takes action, it is individually optimal not to take action. It is also clear that there can be no worse equilibrium than this, because it holds each player at its maximin payoff, which is zero. We will refer to this as the "status-quo equilibrium".

The best equilibrium is one in which each player votes sincerely ($V^i = \theta^i$) and then takes action ($a^i = 1$) if and only if all players have voted in favor of action. This can be viewed as a "unanimity equilibrium": players vote (sincerely), and then the collective action is taken if and only if all players vote in favor. To see that this is indeed an equilibrium, note that (i) no player has incentive to take a different action, given the other players' actions and given that all players have reported truthfully, and (ii) no player has incentive to lie about his preferences, given the subgame strategies. To see that there can be no better equilibrium, note the following: to achieve a more efficient outcome, it would be necessary for some player to play $a^i = 1$ when

$\theta^i = \theta_H$, but this can never be individually rational, hence there would be an incentive to deviate. The following proposition summarizes the worst and best equilibrium outcomes:

Proposition 2 *The worst equilibrium of the one shot game is:* $a^i = 0$ *in all states (*status quo *equilibrium). The best equilibrium of the one shot game is: each member i votes sincerely, and takes action if and only if all members have voted "yes" (*unanimity *equilibrium).*

The unanimity equilibrium is more efficient than the status-quo equilibrium, because it yields the status quo for $N_1 < N$ and a more efficient outcome for $N_1 = N$, but in general it does not deliver the first-best outcome. It is important to emphasize that no external enforcement is needed to sustain the unanimity equilibrium. However, playing this equilibrium requires a certain amount of coordination, thus we think of this equilibrium as capturing a simple form of *organization*.

2.3 One-Shot Game with Enforceable Voting

We now consider the benchmark scenario in which external enforcement is available, in the sense that any contract based on verifiable information can be directly enforced.

Since the θ^i values are private information, hence not verifiable, the parties cannot write a contract that is contingent on the realization of $\boldsymbol{\theta}$. However, they can achieve the first best outcome through the following simple *voting rule*: after $\boldsymbol{\theta}$ is realized, each player casts a vote $V^i \in \{\theta_L, \theta_H\}$, and then all members participate in the collective action if and only if at least q^* members have voted in favor. The key is to note that, given the proposed voting rule, each player has incentive to vote sincerely.[10] Sincere voting then immediately implies the claim.

Proposition 3 *If external enforcement is available, the first-best outcome can be implemented by a voting rule with threshold* $q^* = \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil$.

Note the role of external enforcement: if the organization votes in favor of the collective action, the members that disagree are forced to participate. Without external enforcement, this would not be possible. As we will argue later, the degree of correlation will play a more critical role in the absence of external enforcement.

[10] It is easy to see that voting sincerely is a weakly dominant strategy for each player. Non-sincere voting equilibria exist, but these are characterized by weakly dominated strategies.

It is possible that the optimal enforceable voting rule is unanimity, that is $q^* = N$. This however is a rather special case, which obtains when B is close to θ_L. Thus, if external enforcement is available, unanimity is typically dominated by some other rule. We will argue in the next section that the parameter region where unanimity is optimal expands dramatically when collective decisions must be self-enforcing. To focus on the interesting case, we will assume henceforth that $q^* < N$.

We note that the first-best rule may be a simple majority rule ($q^* = \lfloor N/2 \rfloor + 1$, where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x), a super-majority rule ($q^* > \lfloor N/2 \rfloor + 1$) or a sub-majority rule ($q^* < \lfloor N/2 \rfloor + 1$), depending on the preference parameters. The possibility of a sub-majority rule seems counterfactual, since it is hard to think of real organizations that have adopted this kind of rule, but we could enrich the model in a way that would bring this aspect more in line with reality. In our model a sub-majority rule presents no problems because there is a well-defined status quo ("inaction") and a single action that can be taken, but if the choice were between two actions, neither of which is the status quo, then a sub-majority rule would be problematic.[11] We chose not to pursue this line of modeling because it would complicate the analysis without adding to the qualitative insights of the theory.

3 Self-Enforcing Voting

We now consider the case that is most relevant for an international organization, namely the case in which no external enforcement is available. In other words, the organization members cannot commit to give up sovereignty. Under these circumstances, the only way to enforce cooperation is through repeated interaction. We follow the tradition of the literature on self-enforcing agreements by casting the problem in a repeated-game framework.

[11] For example, suppose that the organization faces a choice between two alternatives A and B, but neither of them is a status quo, and the exact nature of the two alternatives is not known *ex ante*. Then a well-specified rule must not depend on the "name" of the alternatives, that is, it must be of the type "choose one of the alternatives if at least q members are in favor of it." Now consider a sub-majority rule, for example $q = N/3$. This rule leads to non-feasible prescriptions for some states of the world: for example, if the organization is split in the middle (50-50) on the issue at hand, the above rule implies that the organization should choose A *and* B. Note that we do not have this problem in our model: a rule of the kind "take the collective action if at least q members are in favor of it, otherwise do nothing" is well-defined for any q.

3.1 Baseline model

Suppose that the game described in section 2.2 is repeated infinite times. In each period t, each member privately observes the realization of $\theta_t^i \in \{\theta_L, \theta_H\}$, then sends a public message $V_t^i \in \{\theta_L, \theta_H\}$, and then chooses an action $a_t^i \in \{0, 1\}$. For future reference we let $\mathbf{a}_t$ denote the vector of actions. The distribution of the cost vector $\boldsymbol{\theta}_t$ is symmetric with respect to its N arguments and is *iid* across periods. The assumption of symmetric and *iid* distribution is a simple way of extending the notion of a veil of ignorance to a repeated game setting: players do not know which side of a future issue they will be on, and today's issue is no indication of what future issues will be like.[12]

All governments have discount factor δ. This parameter can be interpreted as capturing the governments' degree of stability as well as the frequency with which decisions are made within the organization. Other things equal, δ will be higher if governments are more stable and if issues come up more frequently.

A natural equilibrium notion for this type of game is that of Public Perfect Equilibrium (PPE).[13] The first observation is that we can focus without loss of generality on *truthful* equilibria, i.e. equilibria characterized by sincere voting ($V_t^i = \theta_t^i$ for all i and t). It can be shown that, if there is a PPE that yields the vector of ex-ante payoffs $\mathbf{u}^0$, then there is also a truthful PPE that yields the same vector of ex-ante payoffs $\mathbf{u}^0$. With an abuse of terminology, from now on we will omit the qualifier "truthful" when we speak of equilibria.

A PPE induces a mapping from public histories to continuation payoffs. A player's continuation payoff at time t is defined as the expected present value of her future payoffs as viewed from the end of period t, that is before $\boldsymbol{\theta}_{t+1}$ is realized. We will restrict our attention to *symmetric* Public Perfect Equilibria (SPPE), that is, PPE in which following any public history all players get the same continuation payoff.[14] For example, in our game the symmetry restriction

[12] We are assuming that the costs and benefits of a collective action accrue istantaneously. A more realistic assumption would be that the implementation of a project may take more than one period and there may be more than one project going on at a given point in time. In an earlier version of this paper (available upon request) we considered an extension of the model where it takes multiple periods to implement a project and where projects may overlap, and showed that our qualitative results continue to hold.

[13] Public perfect equilibrium essentially requires that play following each history be a Nash equilibrium (see Fudenberg and Tirole [15] for a more formal definition). The notion of "public perfection" is the natural analog of the notion of subgame perfection for repeated games with private information of this type. Other papers that analyze PPE in repeated games with private information are Athey and Bagwell [4], Athey, Bagwell and Sanchirico [5] and Levin [19].

[14] We note that also Abreu, Pearce and Stacchetti [1] and Athey, Bagwell and Sanchirico [5] focus on equilibria that are symmetric in the same sense as here.

rules out equilibria where today's high-cost players expect a higher future payoff than today's low-cost players. In the final section we will discuss how results may change if one allows for asymmetric equilibria.

It is natural to focus on *optimal* SPPE, that is SPPE that maximize the players' common ex-ante payoff. Optimal equilibria are of particular interest because they represent the best an organization can do without the help of external enforcement.

We can assume without loss of generality that, following any deviation from the optimal equilibrium actions, players revert permanently to the status-quo equilibrium (*trigger* punishment). This is because the status-quo equilibrium keeps the deviator at his maxmin payoff and, since actions are perfectly observed, it is best to punish deviations most severely.

Having pinned down the behavior of players off the equilibrium path, we can now turn to the equilibrium path of the game. The following lemma shows that the equilibrium path of an optimal SPPE must be *stationary*, in the sense that actions at time t depend only on the current state of the world $\boldsymbol{\theta}_t$.

Lemma 1 *Any optimal SPPE is characterized by a stationary equilibrium path.*

Proof: See Appendix.

The basic intuition for this result is the following. Suppose, by way of contradiction, that an optimal SPPE had a nonstationary equilibrium path. Then in this proposed equilibrium the players' common continuation payoff would vary with the history of the game. Letting $u^{\max}$ denote the maximum continuation payoff in the proposed equilibrium, one can construct a new SPPE where the continuation payoff is independent of the history and equal to $u^{\max}$, and therefore yields a higher ex-ante payoff than the proposed equilibrium.

Lemma 1 allows us to focus on stationary equilibrium paths. Given this result, we can further focus without loss of generality on equilibria where players behave according to a simple *voting rule*, and vote sincerely. More precisely, we can focus on strategies with the following structure: (i) $V^i = \theta^i$ for all players i; (ii) $a = 1$ if $N_1 \geq q$ and $a = 0$ if $N_1 < q$, where $N_1 = \#\{i : \theta^i = \theta_L\}$ is the number of low-cost players and $q \in \{1, 2, ..., N\}$ is the approval threshold. This class of strategies is indexed by the voting rule q. We say that a voting rule q is *self-enforcing* if it is part of an equilibrium strategy. Within the set of self-enforcing voting rules, we look for the one that maximizes the players' common expected payoff (or the *optimal* one). The following proposition characterizes such a voting rule.

Proposition 4 *There exists a critical level* $\underline{\delta} \in (0,1)$ *such that the optimal self-enforcing voting rule is* $q = q^*$ *for* $\delta \geq \underline{\delta}$ *and* $q = N$ *for* $\delta < \underline{\delta}$.

Proof: Note first that, within the class of equilibria we are considering, there is never an incentive to lie for any q, thus we only need to worry about the temptation to defect at the action stage. The only incentive to cheat that we need to consider is for a member i that is supposed to take action when he prefers the status quo, i.e., when $\theta^i = \theta_H$ and $N_1 \geq q$. The gain from cheating is $\theta_H - B$, and the discounted loss from cheating is $\frac{\delta}{1-\delta}U(q)$, where $U(q) = \frac{1}{N}\sum_{\{\boldsymbol{\theta}:N_1(\boldsymbol{\theta})\geq q\}} P(\boldsymbol{\theta})[N_1(\boldsymbol{\theta})(B-\theta_L) + (N - N_1(\boldsymbol{\theta}))(B-\theta_H)]$ is the one-period common expected utility given voting rule q. Clearly, the unanimity rule $q = N$ need not satisfy any constraint, thus the problem boils down to

$$\begin{array}{ll} & \max_q U(q) \\ \text{s.t.} \quad \theta_H - B \leq & \frac{\delta}{1-\delta}U(q) \text{ if } q < N \end{array} \tag{2}$$

Note that, since the RHS of (2) is maximized for $q = q^*$, we can restrict attention to two voting rules, $q = q^*$ and $q = N$. If $q = q^*$ satisfies (2), it is also the optimal self-enforcing voting rule. If $q = q^*$ does not satisfy (2) then the optimal self-enforcing voting rule is unanimity ($q = N$). Clearly, there is a critical level $\underline{\delta} \in (0,1)$ such that $q = q^*$ satisfies (2) if and only if $\delta \geq \underline{\delta}$. The claim follows. **QED.**

Notice the *bang-bang* nature of the result: it is never optimal to choose a voting rule that is intermediate between the first best q^* and unanimity. This is because increasing q does not reduce the gain from defecting, unless it is increased all the way to $q = N$, in which case defections are no longer an issue.[15]

The above proposition suggests that a non-unanimous rule is more likely to be adopted in organizations where governments are more patient or stable, and in organizations that make decisions with higher frequency.

An interesting question is how the optimal self-enforcing voting rule is affected by the correlation among members' preferences. As we saw in section 2.1, the optimal enforceable voting rule does not depend on the correlation among the preference shocks θ^i. However, the

[15] One might object that the result relies heavily on the assumption that N is fixed. If N could be changed quickly and costlessly, in each period the high-cost countries could step out of the organization and re-enter at the beginning of the next period, and this would yield a higher ex-ante payoff to all players. However, frictionless adjustment of N would be a highly unrealistic assumption: changing the membership of an international organization takes time and is costly, thus expelling and re-admitting countries in every period is likely to be unfeasible or too costly.

range of discount factors for which the first-best rule q^* is self-enforcing does depend on such correlation. Indeed, we can show that, under mild assumptions on the probability distribution, the range of discount factors for which q^* is self-enforcing expands when the correlation among members' preferences is increased.

We continue to assume that the joint distribution $P(\theta^1, ..., \theta^N)$ is symmetric with respect to its N arguments, and parametrize correlation in the following way. Let N_1^{-i} be the number of members in favor of action excluding member i. This is a random variable with support $\{0, 1, ..., N-1\}$. Let $P^\rho(N_1^{-i}|\theta^i)$ be the probability distribution of N_1^{-i} conditional on θ^i. The superscript ρ denotes a correlation parameter. A natural assumption is that ρ affects this conditional distribution in a first-order stochastic way. Formally, if $\rho' > \rho''$ then

$$\begin{array}{ccc} P^{\rho'}(N_1^{-i}|\theta^i = \theta_L) & \text{FSD} & P^{\rho''}(N_1^{-i}|\theta^i = \theta_L) \\ P^{\rho''}(N_1^{-i}|\theta^i = \theta_H) & \text{FSD} & P^{\rho'}(N_1^{-i}|\theta^i = \theta_H) \end{array} \tag{3}$$

We take the extreme values of ρ to be $\rho = 0$ (independence) and $\rho = 1$ (perfect correlation). It is also natural to assume that ρ does not affect the marginal probability $p^L = \Pr(\theta^i = \theta_L)$. We have the following result:

Proposition 5 *Higher correlation of preferences facilitates enforcement of the first-best rule. In particular, there exist critical levels δ' and δ'', with $0 < \delta' < \delta'' < 1$, such that:*
(i) If $\delta < \delta'$, the optimal self-enforcing rule is $q = N$ for all ρ;
(ii) If $\delta' \leq \delta < \delta''$, the optimal self-enforcing rule is $q = N$ for low values of ρ and $q = q^*$ for high values of ρ;
(iii) If $\delta \geq \delta''$, the optimal self-enforcing rule is $q = q^*$ for all ρ.

Proof: The one-period expected utility given q^* can be written as

$$U^\rho(q^*) = p^L P^\rho(N_1^{-i} \geq q^* - 1|\theta^i = \theta_L)(B - \theta_L) - (1 - p^L)P^\rho(N_1^{-i} \geq q^*|\theta^i = \theta_H)(\theta_H - B)$$

Assumption (3) implies that, as ρ increases, $U^\rho(q^*)$ increases. Hence, the right hand side of (2) is increasing in ρ for given δ. The claim follows directly. **QED.**

Intuition might have suggested that a higher degree of correlation makes unanimity more attractive relative to a non-unanimous rule. The analysis however points in the opposite direction, and the reason is the following. As ρ increases, the value of the q^* rule relative to unanimity (as captured for example by the ratio $U^\rho(q^*)/U^\rho(N)$) may well decrease, but what

matters for the optimal self-enforcing rule is only the *absolute* value of the q^* rule. When the members of an organization are more likely to have the same preferences regarding future collective actions, the value of the relationship is higher, therefore the cost of defecting is higher, and hence the organization is more likely to adopt the first-best rule.[16]

Thus the model broadly predicts that organizations whose members have more homogenous preferences are more likely to be governed by a *non*-unanimous rule. We emphasize that this result is due specifically to the presence of a self-enforcement constraint, since the optimal enforceable voting rule q^* is independent of the degree of correlation.

3.2 Sensitive vs. Technical Issues

Thus far we have assumed that all the issues faced by the organization are characterized by similar *stakes*. Formally, we have assumed that the possible cost realizations (θ_L, θ_H) and the benefit B are the same for all issues. Suppose now that there are two types of issue: (i) high-stake (or *sensitive*) issues, and (ii) low-stake (or *technical*) issues. Let (θ_L^S, θ_H^S) and (θ_L^T, θ_H^T) denote the possible cost realizations respectively for sensitive and technical issues, and assume for simplicity that the benefit B is the same for all issues. A sensitive issue is one that players feel more strongly about, so we posit $\theta_H^S > \theta_H^T$ and $\theta_L^S < \theta_L^T$.[17]

Intuitively, under a non-unanimous system, the incentive to defect by high-cost members is stronger for sensitive issues than for technical ones. This suggests that under some conditions the optimal self-enforcing governance system should be one in which sensitive issues are decided by unanimity and technical issues by the first-best rule. In what follows we examine this intuition more rigorously.

We assume that sensitive and technical issues come up randomly over time. Formally, we assume that in each period the cost vector $\boldsymbol{\theta}_t$ is drawn from one of two distributions: with probability p_S it is drawn from a distribution $P_S(\boldsymbol{\theta})$, which has support $\{\theta_L^S, \theta_H^S\}^N$ and is symmetric with respect to its N arguments; and with probability $1 - p_S$ it is drawn from a distribution $P_T(\boldsymbol{\theta})$, which has support $\{\theta_L^T, \theta_H^T\}^N$ and is also symmetric with respect to its N arguments. We continue to assume that $\boldsymbol{\theta}_t$ is *iid* across periods.

The individual cost realizations are private information, but the nature of the issue (sensitive

[16] In the extreme case of perfect correlation, of course, the q^* rule and the unanimity rule are equivalent, so the problem is not interesting.

[17] The second inequality will actually not matter for the result, so we could alternatively define a sensitive issue simply as one for which θ_H is higher.

or technical) is common knowledge, thus the voting rule can be conditioned on the nature of the issue. To focus on the interesting case, we assume that the first-best voting rule is a non-unanimous rule for both types of issue: $q_k^* < N$ for $k = S, T$, where $q_k^* = \lceil \frac{\theta_H^k - B}{\theta_H^k - \theta_L^k} N \rceil$. All other assumptions of the model are unchanged.

We consider decision-making systems where sensitive issues are decided by a voting rule q_S and technical issues are decided by a (possibly different) voting rule q_T. We look for the pair (q_S, q_T) that maximizes the common expected utility subject to the relevant incentive constraints. We can write the common one-period expected utility given (q_S, q_T) as

$$U(q_S, q_T) \equiv p_S U_S(q_S) + (1 - p_S) U_T(q_T)$$

where $U_k(q_k)$ is the common one-period expected utility if the issue is of type k. The gain from defecting for a high-cost type, given an issue of type k, is $\theta_H^k - B$, therefore we can write the problem as follows

$$\begin{array}{rl} & \max_{q_S, q_T} U(q_S, q_T) \\ s.t. \quad \theta_H^k - B \leq & \frac{\delta}{1-\delta} U(q_S, q_T) \text{ if } q_k < N, \ k = S, T \end{array} \tag{4}$$

Note that the right hand side of the incentive constraint is identical across issues, whereas the left hand side is higher for sensitive issues. This implies, using the logic of Proposition 4, that there exists an intermediate interval of δ such that the optimal pair of rules is (N, q_T^*). The following proposition records this result:

Proposition 6 *If issues can be of two types, sensitive and technical, there is an intermediate interval of δ for which the optimal self-enforcing rule is unanimity for sensitive issues and the first-best rule q_T^* for technical issues.*

We emphasize that this result is driven by the presence of a self-enforcement constraint. Under perfect enforceability, both types of issue would be governed by the first-best rule q_k^*. In fact, it is possible that under perfect enforcement the optimal quota is higher for technical issues than for sensitive issues ($q_S^* < q_T^*$), and introducing the self-enforcement constraint reverses this ranking.

This result suggests that enforcement considerations may contribute to explain why, in some international organizations, some issues are decided by unanimity and some others by majority rule. In the EU, for example, most issues are decided by (simple or qualified) majority rule, but a number of sensitive issues are decided by unanimity rule, for example issues of foreign policy, security policy, agreements with external countries and accession of new members.

3.3 International Transfers

In the basic model we have not considered the possibility of international transfers. This assumption is realistic in some settings but not in others. For example, in the European Union there has been an increasing use of monetary transfers over time, while in the WTO monetary transfers are not used. In this section we amend the baseline model by supposing that international transfers are available.

Transfers may help sustain the first-best outcome: if the members who are in favor of the collective action are willing to compensate the members who are against it, the latter might be convinced to participate. However, the use of transfers is subject to two limitations. First, transfers have to be self-enforcing, just as the decision to participate in the collective action: in other words, a member may refuse to pay. Second, players will have an incentive to vote strategically: a member who favors collective action may be tempted to vote against the collective action, hoping to get compensation.

Assume that transfers enter utility additively, and consider the following timing for the stage game: after players observe their θ^i values, they vote, then they choose actions, and finally transfers are made.[18] Let t^i be the (positive or negative) amount paid by player i. The overall budget constraint is $\sum_{i=1}^{N} t^i \geq 0$. Note that we allow for budget surpluses.[19]

As in the case without transfers, we can focus without loss of generality on truthful equilibria. Also, since actions and transfers are perfectly observed, we can assume that any deviation from the prescribed actions or transfers is followed by a permanent reversion to the status quo equilibrium.

We will focus on the set of efficient PPE. We say that a PPE is *efficient* if the associated ex-ante payoffs $\mathbf{u}^0$ lie on the frontier of the equilibrium payoff set. This is the standard notion of constrained Pareto efficiency in repeated games. Note that in this section we allow for asymmetric PPE, unlike in the no-transfer case.

A full characterization of the efficient PPE for all values of δ is a very difficult task, but

[18] In principle one could consider the alternative sequence in which transfers are made before the collective action is taken. However in this case transfers cannot help, because a high-cost type will face the same temptation to cheat as in the absence of transfers.

[19] Technically, t^i is a choice of player i only if it is positive; a player cannot choose to receive a transfer. Our formulation can be viewed as a reduced form of a richer game where each player i chooses a vector of nonnegative transfers $t^{ij} \geq 0$, where t^{ij} is the transfer from player i to player j, and a transfer to some outside party, $t^{i0} \geq 0$ (to allow for the possibility of budget surpluses). It can be shown that this game is payoff-equivalent to our reduced form.

we will show the following two results: (a) transfers help sustain the first-best outcome, in the sense of expanding the range of discount factors for which the first-best rule is sustainable, but (b) for sufficiently low values of the discount factor, unanimity remains the best self-enforcing governance mode.

To build intuition toward these results, let us focus initially on equilibria of a simple type, where (i) the equilibrium path is stationary; (ii) on the equilibrium path, the collective action is taken if and only if at least q members vote in favor; (iii) if the collective action is undertaken, each high-cost member gets a transfer t; and (iv) the budget is balanced, so each low-cost member contributes an amount $\frac{(N-V_1)t}{V_1}$, where V_1 is the number of votes in favor of action.

Within the class just described, an equilibrium can be identified with a pair (q,t). Observe that equilibria of this type give all players the same expected payoff in each period. Since the budget is balanced and utility is transferable, the members' common expected utility associated with a pair (q,t) is simply $U(q)$.

A pair (q,t) is part of an equilibrium if it satisfies the following incentive constraints:

$$\theta_H - B - t \leq \frac{\delta}{1-\delta}U(q) \text{ for } q < N \tag{5}$$

$$\frac{(N-q)t}{q} \leq \frac{\delta}{1-\delta}U(q) \tag{6}$$

$$\sum_{V_1=q+1}^{N}\left(t+\frac{(N-V_1)t}{V_1}\right)\Pr(N_1=V_1|\theta^i=\theta_L)-(B-\theta_L)\Pr(N_1=q|\theta^i=\theta_L)\leq 0 \tag{7}$$

The first constraint requires that a high-cost member have incentive to participate: the one-time gain from cheating is $\theta_H - B - t$, and the loss from cheating is the future value of cooperation, $\frac{\delta}{1-\delta}U(q)$. The second condition requires that a low-cost member have incentive to make the required payment. The gain from cheating in this case is given by the contribution, $\frac{(N-V_1)t}{V_1}$, and the loss from cheating is again the future value of cooperation; note that, since the gain from cheating is highest when $V_1 = q$, we can replace V_1 with q in the constraint. The third constraint requires that a low-cost member have incentive to vote sincerely. By voting strategically, this member would gain $t + \frac{(N-V_1)t}{V_1}$ (he would get the transfer and avoid the contribution) in the event that he is not pivotal, that is when $N_1 > q$; and he would lose $(B - \theta_L)$ in the event that he *is* pivotal, that is when $N_1 = q$.

Let $\underline{\delta}$ be the minimum level of δ such that the first-best rule q^* can be sustained in equilibrium in the game without transfers. Notice that, if $\delta = \underline{\delta}$, the pair $(q,t) = (q^*, 0)$ satisfies (5) with

equality and the other two constraints with slack, provided $\Pr(N_1 = q^* | \theta^i = \theta_L) > 0$. This condition is generically satisfied, except if preferences are perfectly correlated. Now consider a small transfer t. For t small enough, (q^*, t) satisfies all three constraints with slack if $\delta = \underline{\delta}$. But this implies that (q^*, t) satisfies the three constraints also if δ is slightly lower than $\underline{\delta}$. This establishes that the first best rule q^* can be sustained for a wider range of δ than in the absence of transfers.

Next we argue that unanimity remains the best possible governance mode if δ is sufficiently low. A necessary condition for a pair (q, t) to be part of a PPE is that it satisfy constraints (5) and (6). Consider a non-unanimous rule, $q < N$. Combining constraints (5) and (6), we get

$$\theta_H - B - \frac{\delta}{1-\delta} U(q) \leq t \leq \frac{q}{N-q} \cdot \frac{\delta}{1-\delta} U(q)$$

Clearly, if δ is small enough this condition will be violated, and hence the best self-enforcing rule is unanimity. Intuitively, when δ is small a low-cost member's temptation to renege on the transfer is strong, therefore only a small transfer is incentive-compatible. But the transfer needed to convince a high-cost member to participate in the collective action is not small, and therefore it is impossible to satisfy all incentive constraints at the same time if $q < N$.

It turns out that the insights we just gained by focusing on a simple class of equilibria are valid for the whole class of efficient PPE. In particular, the result is valid also if one allows for non-stationary equilibrium paths. The following proposition states the result:

Proposition 7 *(i) The availability of transfers expands the range of δ for which the first best rule q^* is self-enforcing. (ii) If δ is sufficiently small, any efficient PPE is characterized by the unanimity rule.*

Proof: See Appendix.

Broadly interpreted, this result suggests that an organization is more likely to be governed by a non-unanimous rule if it has the flexibility to make pure transfers between its members.

We conclude this section with a remark: transfers are less helpful when the organization size N is larger. More specifically, the availability of transfers lowers the critical level $\underline{\delta}$ by a smaller amount when N is larger. To see this, note that if N is large, the probability of being pivotal is small, and hence the truthtelling constraint (7) implies that only a small transfer t is enforceable. But this implies that $\underline{\delta}$ can be lowered only by a small amount relative to the no-transfer case.

3.4 Impure collective action

In this section we consider situations where the collective action may be effective even if not all members participate, in which case we speak of "impure collective action".

Formally, consider our baseline model (without transfers) and suppose that the benefit accruing to an individual member is $B(n)$, where n is the number of members that participate in the action. Assume $B(n)$ is increasing in n, with $B(0) = 0$. Also assume that $nB(n)$ is weakly convex in n. This is a relatively mild condition, which is satisfied for example if $B(n)$ has constant elasticity.

We assume that member i has the following utility function:

$$U^i(a^i, n, \theta^i) = a^i \left(B(n) - \theta^i\right) \tag{8}$$

All other assumptions of the model are unchanged. Note the implicit assumption that the benefits from the collective action are excludable: the members who do not participate receive no benefit from collective action. There are some situations in which this assumption is realistic: for example, when the EU decided to adopt a common currency, the United Kingdom chose not to participate, and it arguably did not share in the benefits from the venture. But the primary reason for this assumption is theoretical. Our model focuses on a simple type of cooperation problem, where there is an ex-post incentive to defect for high-cost members. If the benefits of collective action are non-excludable, an additional cooperation problem is introduced, that is a temptation to free ride even by low-cost types, that is members who *favor* collective action ex-post. Furthermore, this could potentially introduce incentives to vote strategically. While this might be an interesting direction for future extensions, here we prefer to shut down this additional free-rider problem. We also note that our results will remain unchanged if some of the benefits spill over to non-participants, as long as this spillover is not too large; or if the benefits are fully non-excludable but the cost of action is relatively small.[20]

We present our results through a series of remarks, which are all proved in Appendix. Let us start by looking at the one-shot game without enforcement. Let $n^{\min}$ denote the first integer n such that $B(n) \geq \theta_L$. The threshold $n^{\min}$ can be interpreted as the "critical mass" of low-cost players below which it is not efficient to undertake collective action. Also let V_1 the number of "yes" votes.

[20] To see this latter point, note that a collective action problem with fully non-excludable benefits can be captured by the utility function $U = B(n) - a^i\theta^i$. If θ_L is small enough, a low-cost type internalizes enough of the benefits that she will not have incentive to free-ride on other low-cost types.

Remark 1 *The worst equilibrium of the one-shot game is the status-quo equilibrium ($a^i = 0$ in all states). The best equilibrium of the one-shot game is the following: all members vote sincerely; then, if $V_1 < n^{\min}$, no one takes action; if $V_1 \geq n^{\min}$, the members who have voted "yes" take action (*coalition of the willing *equilibrium).*

In the coalition-of-the-willing equilibrium, action is taken only by the members who vote in favor of the project, provided this group exceeds the critical mass $n^{\min}$. Note that if $B(N-1) \leq \theta_L$ then $n^{\min} = N$ and hence this equilibrium coincides with the "unanimity" equilibrium that we found in the case of pure collective action. The condition $B(N-1) \leq \theta_L$ means that the benefit drops by a sufficient amount as one member drops out of the collective action, or in other words, the "impurity" in the collective action is small.

Even if $n^{\min} < N$, however, there is a close affinity between this coalition-of-the-willing rule and the unanimity rule: the coalition-of-the-willing rule satisfies the Buchanan-Tullock *ex post* Pareto criterion (no member should receive less than the status-quo utility for any $\boldsymbol{\theta}$). Just as in the case of unanimity, no player participates in the collective action if she is opposed to it.

Next we suppose that external enforcement is available, and characterize the first-best voting rule.

Remark 2 *If external enforcement is available, the first-best outcome can be implemented by a voting rule with the following structure: if $V_1 < q_1$, no one takes action; if $q_1 \leq V_1 < q_2$, the members who have voted "yes" take action; if $V_1 \geq q_2$, all members must take action (where $1 \leq q_1 \leq q_2 \leq N$). The intermediate interval of V_1 is empty ($q_1 = q_2$) if $B(\tilde{q}^*) \leq \theta_L$, where $\tilde{q}^* \equiv \lceil \frac{\theta_H - B(N)}{\theta_H - \theta_L} N \rceil$.*

This rule is similar to the first-best rule in the case of pure collective action, except that there may be an intermediate interval of V_1 for which action is taken only by a coalition of the willing.

The proposition also provides a sufficient condition under which there is a single threshold for action ($q_1 = q_2$), so that the first-best rule has the same structure as in the case of pure collective action. This condition is satisfied if the benefit B drops sufficiently as the participation rate drops, or in other words if the "impurity" in the collective action is sufficiently small.[21]

[21] Note that $\tilde{q}^*$ is the rule that would be optimal if the collective action were "purified," that is if we replaced $B(n)$ with a schedule $\tilde{B}(n)$ that is equal to zero for $n < N$ and equal to $B(N)$ for $n = N$.

Note that this condition is weaker than the condition under which the best equilibrium of the one-shot game is the pure unanimity rule ($B(N-1) \leq \theta_L$).

Next we consider self-enforcing voting rules. The question is to what extent proposition 4 extends to situations of impure collective action. We have the following result:

Remark 3 *There exists a critical level $\hat{\delta}$ such that: (i) for $\delta \geq \hat{\delta}$, the optimal self-enforcing voting rule is the first-best rule described in remark 2; (ii) for $\delta < \hat{\delta}$, the optimal self-enforcing voting rule is the "coalition-of-the-willing" rule described in remark 1.*

This is a generalization of the bang-bang result that we obtained in the basic model with pure collective action. If δ is relatively high, the first-best voting rule can be sustained, but if δ is relatively low, the most that can be achieved is the best equilibrium of the one-shot game.

One might ask whether the result that it may be optimal to keep some countries out of the collective action is consistent with what we observe in reality. First, as we discussed above, if the "impurity" of the collective action problem is relatively small the optimal decision rule has the same structure as in the case of pure collective action, with all players always taking the same action; and we argued in section 2 that a variety of collective action problems faced by international organizations are not far from "pure". Second, one can think of situations where not all members of an international organization participate in its collective actions. One example is given by the peacekeeping missions of the United Nations, where often only a subset of members participate in the mission.

4 Discussion

In this final section we discuss the interpretation of our results and some extensions of the model.

First we need to discuss the interpretation of "voting" in our model, and how it relates to the voting systems we observe in real organizations. In our model, voting is a way for the organization members to communicate preferences to one another. We have assumed that this communication takes the simplest possible form, that is, governments simultaneously cast a yes/no vote. But now consider another way in which communication could occur. Suppose that governments vote in two stages, first an informal vote and then an "official" vote, and all the information is communicated at the informal stage. In this case, the informal vote is the

"real" one, and the official vote becomes irrelevant. Clearly, if communication occurs in this way, our results on the optimal self-enforcing voting rule are still valid but they apply to the informal vote, not to the formal one. This suggests that we should be cautious in relating our model to reality, because in reality we may be able to observe only the official voting rule, while our model yields predictions about the "real" decision rule, which may be hidden behind the scenes.

Our basic model cannot explain why it might be desirable to have a two-stage voting rule, but a possible explanation might be that the organization wants to transmit a unified image to the rest of the world.[22] This argument could be captured in a stylized way with a simple extension of the model. Suppose the organization gets a benefit if the rest of the world observes that all members vote in the same way, but an individual member suffers a cost if she misrepresents her view to the rest of the world. Then under some conditions it will be optimal to structure the deliberation process in the following way: first an informal vote where positions are revealed only within the organization but not to outsiders, and then an "official" vote where decisions are made by unanimity. Note that this would effectively add a second layer of collective action to the model, since the official votes would be actions with direct payoff relevance. In this setting, considerations of "image" could explain unanimity even in the absence of enforcement problems (e.g. if δ is high).

Our model offers *one* theoretical explanation for the unanimity rule – based on the presence of enforcement considerations – but there is at least one other possible explanation, that is based on considerations of image vis-a-vis the rest of the world. It is not obvious whether and how these two explanations can be sorted out empirically: if we observe that an organization has an official voting rule that requires unanimity, one possible explanation is that this is a "real" unanimity rule and is motivated by self-enforcement considerations, and another explanation is that it is motivated by considerations of image, and the real decision rule behind the scenes is not unanimous. However there are some observed features that we think can be explained more easily with our theory than with the alternative one. For example, in some international organizations high-stake issues are decided by unanimity while lower-stake issues are decided by (simple or qualified) majority. This feature is easily explained with considerations of enforcement, as we argued in section 3.2, while it would be harder to explain it

[22] See Seidmann [25] for a paper that makes a similar point.

with considerations of "image".[23] Hence we are inclined to think that the use of unanimity in international organizations is explained at least in part by enforcement considerations.

Next we discuss our restriction to equilibria where players get the same continuation payoff after any history (SPPE). It is possible that an asymmetric equilibrium might yield a higher ex-ante payoff than the optimal SPPE. In particular, it may be easier to sustain the first-best outcome if today's high-cost members receive a higher future payoff than today's low-cost members, because this will reduce the high-cost members' incentive to defect. For example, the equilibrium might specify that the players who vote "no" at time t but participate in the collective action receive a higher voting weight at time $t+1$.[24] Note however that this type of mechanism would introduce incentives to vote strategically: a low-cost player would be tempted to vote "no". A full analysis of asymmetric equilibria would be very difficult, but it is not hard to show that our results will still be valid for δ sufficiently high and for δ sufficiently low. If δ is sufficiently high, the first-best rule q^* will of course still be sustainable. And if δ is sufficiently low, unanimity will still be the optimal rule; the simplest way to see this is to note that adding the possibility of transfers to this game expands (weakly) the PPE payoff set, and we have shown that even when transfers are available, unanimity is the optimal rule for low values of δ. Thus results could change – if at all – only for intermediate values of δ.[25]

We have focused on equilibria where deviations are punished with a permanent reversion to the status-quo equilibrium. This type of punishment suffers from a problem of collective credibility: once the game is in the punishment phase, there is a strong incentive for players to collectively reconsider the plan of action. We do not have a complete analysis of renegotiation-proof equilibria. However, we have in mind a simple alternative punishment strategy that is much more robust to renegotiation than the trigger punishment we considered in the previous sections: a player that deviates could be expelled from the organization. This punishment

[23] The same can probably be said for the fact that the EU has recently switched from unanimity to qualified majority in many policy areas. It would not be easy to explain this switch if unanimity was motivated by "image" considerations: the importance of conveying a unified image to the rest of the world for an organization like the EU has probably increased, not decreased, in recent years. On the other hand, our theory can easily explain this change of voting rule, as highlighted in our working paper (Maggi and Morelli, 2003).

[24] Another example would be a system of storable votes, as proposed by Casella [12].

[25] We have assumed that players vote openly. How would results change if players voted by secret ballots? Of course the use of secret ballots cannot increase efficiency in our model, thus a rationale for this voting procedure must lie in considerations outside our model. But conditional on using secret ballots, there would be an interesting implication for the analysis: secret ballots would have the effect of ruling out asymmetric equilibrium paths, because they would make it impossible to identify a member's individual vote, and hence her cost realization, thus the continuation payoff on the equilibrium path would necessarily be the same for all players. In this case, therefore, the restriction to SPPE would be without loss of generality.

would give the deviator her maximin payoff, which is the same as under a trigger punishment, and hence the incentive constraints would be exactly the same as under a trigger punishment. At the same time, the remaining $N-1$ players would suffer only a modest reduction of utility relative to the equilibrium path, so the incentives to renegotiate in the punishment phase would be limited.

The reason we did not work directly with expulsion punishments is that this would require expanding the strategy space in a way that makes the expulsion of a member a meaningful strategy. One way of doing this would be to assume that, for the organization to be effective, each member must be *connected* with all other members (e.g., it must have an active communication line). At the beginning of each period, each member has the option of cutting the communication line with one or more other members. If a member is disconnected from all others, it cannot participate in the collective action, and is effectively "expelled". In this extended game, after a player has deviated, it is an equilibrium for the players who have not deviated to cut the connection with the deviator and continue cooperating among themselves. Rather than expanding the game in this fashion and make the notation more complicated, we opted to keep the more basic version of the game and work with the simpler trigger punishment.

Another key assumption of the basic model is that the organization members design the institution under a veil of ignorance. We believe that in some environments this assumption is not unreasonable. Even if members are asymmetric at the stage in which they negotiate the rules of the organization, this asymmetry will play a negligible role if it is difficult to predict the exact nature of the issues that the organization will face in the future, so that it is difficult to predict what the relevant payoff functions will be. In this type of environment, the veil-of-ignorance assumption is a reasonable approximation. In other environments, it may be more reasonable to assume that members have asymmetric payoff functions at the institution-design stage *and* present asymmetries are powerful predictors of future asymmetries for the issues to come, in which case one needs to depart from the veil-of-ignorance setting. Next we discuss how the presence of ex-ante asymmetries of this kind may impact the results of the model.

One interesting possibility is that countries have different outside options. Countries with higher outside options can be thought of as more "powerful," in the sense that they have less to lose from a break-up of the organization. Another interesting possibility is that governments differ in the discount factor, perhaps because they differ in the degree of stability.

Let us suppose that countries have heterogenous outside options and/or discount factors,

and consider the class of simple voting rules, where a collective action is approved if and only if at least q members vote in favor. For a given $q < N$, a country with a higher outside option or a lower discount factor clearly has a stronger incentive to defect. This has a simple implication: the binding incentive constraint will be the one for the most "problematic" country, that is the country that has the highest temptation to cheat given a high-cost realization ($\theta^i = \theta_H$). As a consequence, the presence of even just one country with a low discount factor or a high outside option can force the organization to make decisions by unanimity.

The above observation is relevant for situations where the constitution writers, for any reason, choose to focus on egalitarian voting rules, that is rules where all members have equal voting power. However, if non-egalitarian voting rules are allowed, there may be systems that do better than unanimity. One possibility would be to adopt a non-unanimous voting rule but give the "problematic" countries *veto power*. For example, suppose that some member countries have a higher outside option than others. Then the unanimity rule might be dominated by a system where the countries with higher outside options have veto power. Intuitively, since the self-enforcement constraint is more stringent for the high-outside-option countries, giving them veto power would remove their incentive to defect. This might help explain why some international organizations such as the UN Security Council grant veto power to a subset of its members.

In principle there is another potential way to deal with issues of ex-ante heterogeneity, namely, countries could form multiple organizations. For example, if there are two groups of countries, one characterized by a higher incentive to defect than the other, it might conceivably be optimal for them to form two separate organizations, one governed by unanimity and one governed by a non-unanimous rule. However this line of reasoning leads to complex issues of endogenous organization membership and coalition formation, which are fascinating but outside the scope of this paper.

5 Appendix

Proof of Lemma 1:

Following Abreu, Pearce and Stacchetti's ([1], [2]) factorization approach, a PPE can be represented in compact form as a pair $(\mathbf{A}; \boldsymbol{\nu}(\phi))$, where $\mathbf{A}$ is the vector of first-period strategies and $\boldsymbol{\nu}(\phi)$ is the vector of continuation payoffs as a function of the public outcome at the end of the first period.[26] Given sincere voting, we can think of $\mathbf{A}$ as containing simply the schedule $\mathbf{a}(\boldsymbol{\theta})$, and the public outcome ϕ as containing the realization of $\boldsymbol{\theta}$ and the chosen actions $\mathbf{a}$.

We must distinguish between two kinds of deviation: (i) deviations on actions, which are perfectly observed; and (ii) insincere voting. Deviations of type (i) are best punished with the most severe punishment possible, that is a permanent reversion to the status quo. We can then decompose the continuation value function $\boldsymbol{\nu}(\phi)$ in two parts. If a deviation on actions has been observed, then all players get a continuation payoff of zero (of course, this continuation payoff remains off the equilibrium path). If, on the other hand, no deviation on actions has been observed, players get continuation payoffs $\mathbf{u}(\boldsymbol{\theta})$. These can be interpreted as continuation payoffs "on the equilibrium path." In order to be part of a PPE, the schedules $\mathbf{a}(\boldsymbol{\theta})$ and $\mathbf{u}(\boldsymbol{\theta})$ must of course be such that players have no incentive to lie or to deviate from the equilibrium actions.

Given our restriction to symmetric equilibria, we can write $\nu^i(\phi) = \nu(\phi)$ and $u^{0i} = u^0$ for all i. We show that, at an optimal SPPE, the continuation payoff function $u(\boldsymbol{\theta})$ is constant and equal to the ex-ante payoff: $u(\boldsymbol{\theta}) = u^0$. Stationarity of the equilibrium path is an immediate consequence of this.

Let $\bar{u}$ denote the maximum payoff attainable by a SPPE. Since the set of equilibrium payoffs is compact (this can be shown as in Abreu, Pearce and Stacchetti [2]), this maximum exists. Recall that, as a general property of Public Perfect Equilibria, the set of continuation equilibria after any history is the same as the set of equilibria at the beginning of the game. It follows that $u(\boldsymbol{\theta})$ cannot exceed $\bar{u}$ for any $\boldsymbol{\theta}$. It is also clear that the minimum SPPE payoff is zero, since (i) this is a player's maxmin and (ii) there is a SPPE that gives zero to all players, namely a permanent reversion to the status quo. Next note that, given our focus on optimal equilibria, we can focus on action schedules of the type $a(\boldsymbol{\theta})$, where $a = 1$ means that all players take action and $a = 0$ means that no player takes action. The pair $(a(\boldsymbol{\theta}), u(\boldsymbol{\theta}))$ is part of a SPPE if

[26] We note that Abreu, Pearce and Stacchetti use the notion of sequential equilibrium rather than that of PPE, but their analysis applies to PPE as well.

and only if it satisfies the following constraints:

$$(B-\theta_j)a(\boldsymbol{\theta})+\delta u(\boldsymbol{\theta})\geq 0,\ \forall\boldsymbol{\theta},\ j=H,L \tag{9}$$

$$E[(B-\theta_L)a(\theta_L,\boldsymbol{\theta}^{-i})+\delta u(\theta_L,\boldsymbol{\theta}^{-i})|\theta^i=\theta_L]\geq E[(B-\theta_L)a(\theta_H,\boldsymbol{\theta}^{-i})+\delta u(\theta_H,\boldsymbol{\theta}^{-i})|\theta^i=\theta_L],\ i=1,..,N \tag{10}$$

$$E[(B-\theta_H)a(\theta_H,\boldsymbol{\theta}^{-i})+\delta u(\theta_H,\boldsymbol{\theta}^{-i})|\theta^i=\theta_H]\geq E[(B-\theta_H)a(\theta_L,\boldsymbol{\theta}^{-i})+\delta u(\theta_L,\boldsymbol{\theta}^{-i})|\theta^i=\theta_H],\ i=1,..,N \tag{11}$$

$$0\leq u(\boldsymbol{\theta})\leq \bar{u},\ \forall\boldsymbol{\theta} \tag{12}$$

where $a(\theta_L,\boldsymbol{\theta}^{-i})$ is the collective action prescribed if player i has cost θ_L and the remaining players have the cost vector $\boldsymbol{\theta}^{-i}$, $u(\theta_L,\boldsymbol{\theta}^{-i})$ is the common continuation value when player i has cost θ_L and the remaining players have the cost vector $\boldsymbol{\theta}^{-i}$; similar definitions apply for $a(\theta_H,\boldsymbol{\theta}^{-i})$ and $u(\theta_H,\boldsymbol{\theta}^{-i})$. The operator $E[\cdot|\theta^i=\theta_L]$ denotes the interim expectation over $\boldsymbol{\theta}$ conditional on player i being a θ_L type, and similarly for $E[\cdot|\theta^i=\theta_H]$. The first constraint ensures that noone has incentive to cheat at the action stage; the second (third) one ensures that a low-cost (high-cost) type has incentive to vote sincerely; the fourth one requires that the continuation payoff $u(\boldsymbol{\theta})$ lie in the equilibrium payoff set.

The ex-ante payoff u^0 associated with $(a(\boldsymbol{\theta}),u(\boldsymbol{\theta}))$ is given by

$$u^0=E[(B-\theta^i)a(\boldsymbol{\theta})]+\delta E[u(\boldsymbol{\theta})]$$

(Note that $E[(B-\theta^i)a(\boldsymbol{\theta})]$ is independent of the player i because of the veil of ignorance, i.e. because the distribution $P(\boldsymbol{\theta})$ is symmetric in its N arguments.)

An optimal SPPE must have $u^0\geq u(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. If it were $u^0<u(\boldsymbol{\theta}')$ for some $\boldsymbol{\theta}'$, we could increase efficiency by replacing the original equilibrium with the equilibrium that yields $u(\boldsymbol{\theta}')$, a contradiction. Next we rule out the possibility that $u^0>u(\boldsymbol{\theta}')$ for some $\boldsymbol{\theta}'$. Suppose this were the case. Then we can construct another equilibrium that does strictly better than the original one. Replace $u(\boldsymbol{\theta})$ with u^0 for all $\boldsymbol{\theta}$, and replace $a(\boldsymbol{\theta})$ with the following

$$a'(\boldsymbol{\theta})=\begin{cases} a^*(\boldsymbol{\theta}) & \text{if } N_1(\boldsymbol{\theta})<N \text{ and } a(\boldsymbol{\theta})=1 \text{ for some } \boldsymbol{\theta} \\ a(\boldsymbol{\theta}) & \textit{otherwise}\end{cases}$$

where $a^*(\boldsymbol{\theta})$ is the first-best action profile. In words, we replace the original action schedule with the first-best schedule if the original one calls for some high-cost type to take action. It is direct to verify that the pair $(a'(\boldsymbol{\theta}),u^0)$ satisfies all the equilibrium conditions, hence it is part of a SPPE. Moreover, since the first-period action $a'(\boldsymbol{\theta})$ involves a first-period payoff at least as

high as $a(\boldsymbol{\theta})$, and we have strictly increased the continuation payoff, the new equilibrium has a strictly higher ex-ante payoff than the original one, thus contradicting its efficiency. We can conclude that $u^0 = u(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. **QED**

Proof of Proposition 7:

As in the proof of the previous result, we follow the factorization approach and represent a PPE as a pair $(\mathbf{A}; \boldsymbol{\nu}(\phi))$, where $\mathbf{A}$ includes the first-period schedules $\mathbf{a}(\boldsymbol{\theta})$ and $\mathbf{t}(\boldsymbol{\theta})$ and ϕ includes the realization of $\boldsymbol{\theta}$ and the chosen actions and transfers. If no deviation on actions or transfers is observed, i.e. if play remains on the equilibrium path, players get continuation payoffs $\mathbf{u}(\boldsymbol{\theta})$. We can then think of an equilibrium path as described by a triplet $(\mathbf{a}(\boldsymbol{\theta}), \mathbf{t}(\boldsymbol{\theta}); \mathbf{u}(\boldsymbol{\theta}))$. These schedules of course have to be such that players have incentive to vote sincerely. Consider an efficient PPE $(\mathbf{a}(\boldsymbol{\theta}), \mathbf{t}(\boldsymbol{\theta}); \mathbf{u}(\boldsymbol{\theta}))$. Clearly we can focus on action schedules of the type $a(\boldsymbol{\theta})$, where $a = 1$ means that all players take action and $a = 0$ means that no player takes action. Letting $\hat{u}^i(\boldsymbol{\theta}) = -t^i(\boldsymbol{\theta}) + \delta u^i(\boldsymbol{\theta})$ denote player i's continuation value inclusive of today's transfer, this PPE must satisfy the following conditions:

$$(B - \theta_j)a(\boldsymbol{\theta}) + \hat{u}^i(\boldsymbol{\theta}) \geq 0, \ \forall \boldsymbol{\theta}, \ i = 1, ..., N, \ j = H, L \tag{IC$_A$}$$

$$\hat{u}^i(\boldsymbol{\theta}) \geq 0, \ \forall \boldsymbol{\theta}, \ i = 1, ..., N \tag{IC$_T$}$$

$$E[(B - \theta_L)a(\theta_L, \boldsymbol{\theta}^{-i}) + \hat{u}^i(\theta_L, \boldsymbol{\theta}^{-i})|\theta^i = \theta_L] \geq E[(B - \theta_L)a(\theta_H, \boldsymbol{\theta}^{-i}) + \hat{u}^i(\theta_H, \boldsymbol{\theta}^{-i})|\theta^i = \theta_L], \ i = 1, .., N \tag{TC$_L$}$$

$$E[(B - \theta_H)a(\theta_H, \boldsymbol{\theta}^{-i}) + \hat{u}^i(\theta_H, \boldsymbol{\theta}^{-i})|\theta^i = \theta_H] \geq E[(B - \theta_H)a(\theta_L, \boldsymbol{\theta}^{-i}) + \hat{u}^i(\theta_L, \boldsymbol{\theta}^{-i})|\theta^i = \theta_H], \ i = 1, .., N \tag{TC$_H$}$$

$$\sum_{i=1}^{N} t^i(\boldsymbol{\theta}) \geq 0, \ \forall \boldsymbol{\theta} \tag{TF}$$

$$u^i(\boldsymbol{\theta}) \geq 0, \ \forall \boldsymbol{\theta}, \ i = 1, ..., N \tag{IR}$$

where we use similar notation as in the proof of Lemma 1. The first constraint ensures that no player has incentive to cheat at the action stage; the second constraint ensures that noone has incentive to cheat at the transfer stage; the third and fourth are the truthtelling constraints for low-cost and high-cost types; the fifth is a transfer feasibility constraint; and the sixth is the individual rationality constraint. To be more precise, we should write that the transfer feasibility constraint applies only if $t^i(\boldsymbol{\theta}) > 0$, i.e. if player i is supposed to make a payment, but we do not need to worry about this, because if $t^i(\boldsymbol{\theta}) \leq 0$ this constraint is redundant given the individual rationality constraint $u^i(\boldsymbol{\theta}) \geq 0$.

We can write the ex-ante payoff for player i associated with the PPE $(a(\boldsymbol{\theta}), \mathbf{t}(\boldsymbol{\theta}); \mathbf{u}(\boldsymbol{\theta}))$ as

$$u^{0i} = E[(B - \theta^i)a(\boldsymbol{\theta}) - t^i(\boldsymbol{\theta})] + \delta E[u^i(\boldsymbol{\theta})]$$

To prove point (i), recall that the first best rule requires that the collective action should be taken if and only if $N_1(\boldsymbol{\theta}) \geq q^*$ (where $N_1(\boldsymbol{\theta})$ is the number of players who have low cost in state $\boldsymbol{\theta}$). Let $a^*(\boldsymbol{\theta})$ denote the corresponding action schedule. Let $\underline{\delta}$ be the minimum value of δ for which the first best rule $a^*(\boldsymbol{\theta})$ satisfies the equilibrium conditions without transfers. This is the value of δ that solves $\theta_H - B = \frac{\delta}{1-\delta}u^*$, where $u^* = E[(B - \theta_j)a^*(\boldsymbol{\theta})]$. In other words, for $\delta = \underline{\delta}$ the IC_A constraints for high-cost types are binding at $a^*(\boldsymbol{\theta})$ with $t(\boldsymbol{\theta}) \equiv 0$.

We now show that, if δ is slightly lowered, say to $\underline{\delta} - \varepsilon$, there exist a transfer schedule $t^*(\boldsymbol{\theta})$ such that $(a^*(\boldsymbol{\theta}), t^*(\boldsymbol{\theta}))$ satisfies all the constraints. Clearly, if $\delta = \underline{\delta}$ the IC_T and TC constraints are not binding at $a(\boldsymbol{\theta}) = a^*(\boldsymbol{\theta}), t(\boldsymbol{\theta}) \equiv 0$. This means that we can raise the transfer t_j for any player by at least a small amount, and still keep these constraints satisfied. Consider having low-cost types make a small payment to high-cost types in every state $\boldsymbol{\theta}$ such that $a^*(\boldsymbol{\theta}) = 1$ (keeping a balanced budget, so that no money is wasted). This will relax the IC_A constraints for high-cost types, without violating the other constraints. But this means that now we can lower δ to $\underline{\delta} - \varepsilon$ without violating any of the constraints. This proves the claim.

(ii) Consider a given player, say $i = 1$, and a state $\boldsymbol{\theta}$ where $\theta^i = \theta_H$. From TF,

$$-t^1(\theta_H, \boldsymbol{\theta}^{-1}) \leq \sum_{i \neq 1} t^i(\boldsymbol{\theta})$$

From IC_T, summing up over all players other than i, we get

$$\sum_{i \neq 1} t^i(\boldsymbol{\theta}) \leq \delta \sum_{i \neq 1} u^i(\boldsymbol{\theta})$$

From IC_A, we have

$$-t^1(\theta_H, \boldsymbol{\theta}^{-1}) \geq (\theta_H - B)a(\boldsymbol{\theta}) - \delta u^1(\boldsymbol{\theta})$$

Combining the above conditions, we get

$$(\theta_H - B)a(\boldsymbol{\theta}) - \delta u^1(\boldsymbol{\theta}) \leq -t^1(\theta_H, \boldsymbol{\theta}^{-1}) \leq \delta \sum_{i \neq 1} u^i(\boldsymbol{\theta})$$

which implies

$$(\theta_H - B)a(\boldsymbol{\theta}) \leq \delta \sum_{i=1}^{N} u^i(\boldsymbol{\theta})$$

If δ is sufficiently small, this condition is satisfied only if $a(\boldsymbol{\theta}) = 0$. This implies that there cannot be collective action when there are high-cost types. It is then a small step to conclude that any efficient PPE must be characterized by unanimity: $a(\boldsymbol{\theta}) = 1$ if all players have low cost and $a(\boldsymbol{\theta}) = 0$ otherwise. **QED**

Proof of Remark 1: straightforward.

Proof of Remark 2:

Recall that N_1 is the number of low-cost members, and let n_1 be the number of low-cost members who participate in the action. Let $N_0 \equiv N - N_1$ and n_0 the number of high cost members who participate in the action. In order to find the first-best mapping we need to maximize the joint surplus of the group with respect to n_0 and n_1:

$$\max_{n_0,n_1} J(n_0, n_1) \equiv n_1[B(n_0 + n_1) - \theta_L] + n_0[B(n_0 + n_1) - \theta_H] \quad (13)$$

$$\text{s.t. } 0 \leq n_0 \leq N - N_1, 0 \leq n_1 \leq N_1 \quad (14)$$

Since $nB(n)$ is assumed to be weakly convex, it is easy to see that J is convex in each argument. The convexity of J implies that the solution is corner.

There are only 4 candidate corners: $(n_0, n_1) \in \{(0,0), (N - N_1, 0), (0, N_1), (N - N_1, N_1)\}$. Clearly, $(N - N_1, 0)$ is dominated: it cannot be optimal that high cost types act and low cost types do not. Thus we have to compare the values $J(0,0) = 0$, $J(0, N_1) = N_1[B(N_1) - \theta_L]$, and $J(N - N_1, N_1) = N_1[B(N) - \theta_L] + (N - N_1)[B(N) - \theta_H] = NB(N) - N\theta_H + N_1(\theta_H - \theta_L)$.

The comparison depends on the value of N_1. Consider the functions of N_1 $g^0(N_1) = J(0, N_1)$ and $g^1(N_1) = J(N - N_1, N_1)$. Note that (I) $g^1(0) < 0 = g^0(0)$; (II) $g^0(N) = g^1(N)$; (III) $g^0(N_1)$ is convex and $g^1(N_1)$ is linear; (IV) $g^0(N_1) \leq 0$ for $N_1 < n^{\min}$ and $g^0(N_1) > 0$ for $N_1 > n^{\min}$; (V) $g^1(N_1) < 0$ for $N_1 < \tilde{q}^*$ and $g^1(N_1) > 0$ for $N_1 \geq \tilde{q}^*$. (Note that $\tilde{q}^*$ can be higher or lower than $n^{\min}$.) Using this information we can conclude that there are three possibilities:

1. There is a quota $q_2 > n^{\min}$ such that the optimum is $(N - N_1, N_1)$ for $N_1 > q_2$, $(0, N_1)$ for $n^{\min} < N_1 < q_2$, and $(0,0)$ if $N_1 < n^{\min}$.

2. The optimum is $(0, N_1)$ for $N_1 > n^{\min}$ and $(0,0)$ otherwise.

3. The optimum is $(N - N_1, N_1)$ for $N_1 > \tilde{q}^*$ and $(0,0)$ otherwise.

The convexity of g^0, the linearity of g^1, and the other boundary conditions on these functions guarantee that there are no other possibilities. It is easy to see that each of these cases is consistent with the statement of remark 2: in the first case, all three intervals of the first-best schedule are non-empty; in the second case, the upper interval is empty; and in the third case, the intermediate interval is empty.

It is direct to verify that the first-best outcome we just described can be implemented with the voting rule proposed in remark 2. Finally, one can check that if $B(\tilde{q}^*) \leq \theta_L$ then the optimum is always case (3) above. **QED.**

Proof of Remark 3:

The first observation is that Lemma 1 is valid also in this case of impure collective action, thus we can focus on stationary equilibrium paths. Let $\mathbf{a}(\boldsymbol{\theta})$ denote the action schedule on the equilibrium path.

The key is to argue that we can focus on two action schedules: (i) the one that corresponds to the coalition-of-the-willing rule defined in remark 1 – let $\mathbf{a}^{cow}(\boldsymbol{\theta})$ denote such profile; and (ii) the one that corresponds to the first-best rule defined in remark 2 – let $\mathbf{a}^*(\boldsymbol{\theta})$ denote this action schedule.

Clearly, if $\mathbf{a}^*(\boldsymbol{\theta})$ is self-enforcing then it is optimal. The self-enforcement condition for $\mathbf{a}^*(\boldsymbol{\theta})$ is

$$\theta_H - B(N) \leq \frac{\delta}{1-\delta} U^*$$

where U^* is the per-period expected payoff associated with the first-best rule. The LHS of the above inequality is the one-period gain from cheating, which occurs when a θ_H type is called to action.

An alternative schedule $\mathbf{a}(\boldsymbol{\theta})$ can be preferred to $\mathbf{a}^*(\boldsymbol{\theta})$ only if it implies a one-period gain from cheating strictly lower than $\theta_H - B(N)$ for all states θ. Since $B(n)$ is increasing, a schedule $\mathbf{a}(\boldsymbol{\theta})$ can satisfy this condition only if a θ_H type is never called to action, for otherwise his gain from cheating would be at least $\theta_H - B(N)$. But if θ_H types are never called to action, it is easy to see that we can do no better than $\mathbf{a}^{cow}(\boldsymbol{\theta})$. It is also clear that $\mathbf{a}^{cow}(\boldsymbol{\theta})$ implies no unilateral incentive to deviate, since it is an equilibrium of the stage game. Therefore the only candidates for an optimum are $\mathbf{a}^{cow}(\boldsymbol{\theta})$ and $\mathbf{a}^*(\boldsymbol{\theta})$. We can easily conclude that $\mathbf{a}^*(\boldsymbol{\theta})$ is optimal if δ is higher than a critical level, and $\mathbf{a}^{cow}(\boldsymbol{\theta})$ is optimal otherwise. **QED.**

References

ABREU, D., D. PEARCE and E. STACCHETTI (1986), "Optimal Cartel Equilibria with Imperfect Monitoring", *Journal of Economic Theory*, **39**, 251-69.

ABREU, D., D. PEARCE and E. STACCHETTI (1990), "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, **58**, 1041-63.

AGHION, P. and P. BOLTON (2002), "Incomplete Social Contracts," *Journal of the European Economic Association*, **1**, 38-67.

ATHEY, S. and K. BAGWELL (2001), "Optimal Collusion with Private Information", *RAND Journal of Economics*, **32**, 428-65.

ATHEY, S., K. BAGWELL and C. SANCHIRICO (2004), "Collusion and Price Rigidity", *Review of Economic Studies*, **71**, 317-49.

AUSTEN-SMITH, D. and J. BANKS (1997), "Information Aggregation, Rationality and the Condorcet Jury Theorem," *American Political Science Review*, **90**, 34-45.

BARBERA, S. and M.O. JACKSON (2004), "Choosing how to Choose: Self-Stable Majority Rules and Constitutions," *Quarterly Journal of Economics*, **119**, 1011-48.

BARBERA, S., M. MASCHLER and J. SHALEV (2001), "Voting for Voters: A Model of Electoral Evolution," *Games and Economic Behavior*, **37**, 40-78.

BUCHANAN, J.M. and G. TULLOCK (1967), *The Calculus of Consent, Logical Foundations of Constitutional Democracy*, Ann Arbor, University of Michigan Press.

CAPLIN A. and B. NALEBUFF (1988), "On 64%-Majority Rule." *Econometrica*, **56**, 787-814.

CARRUBBA, C.J. and C. VOLDEN (2000), Coalitional Politics and Logrolling in Legislative Institutions", *American Journal of Political Science*, **44**, 261-77.

CASELLA, A. (2002), "Storable Votes," NBER Working Paper no. 9189.

DASGUPTA, P. and E. MASKIN (1998), "On the Robustness of Majority Rule," mimeo.

EATON, J., and R. FERNANDEZ (1995), "Sovereign Debt," in G. Grossman and K. Rogoff (eds.), *Handbook of International Economics*, vol.3, Amsterdam: North-Holland.

FUDENBERG, D. and J. TIROLE (1991), *Game Theory*, the MIT Press: Cambridge MA.

GUTTMAN, J. (1998), "Unanimity and Majority Rule: the Calculus of Consent Reconsidered", *European Journal of Political Economy*, **14**, 189-207.

LEDYARD, J. and T. PALFREY (1994), "Voting and Lottery Drafts as Efficient Public Goods Mechanisms," *Review of Economic Studies*, **61**, pp.327-356.

LEDYARD, J. and T. PALFREY (2002), "The Approximation of Efficient Public Goods Mechanisms by Simple Voting Schemes," *Journal of Public Economics*, **83.**

LEVIN, J. (2003), "Relational Incentive Contracts", *American Economic Review*, **93**, 835-57.

MAGGI, M. and M. MORELLI (2003), "Self Enforcing Voting in International Organizations," NBER Working Paper no. 10102.

MAY, K.O. (1952), "A set of Independent, Necessary and Sufficient Conditions for Simple Majority Decisions," *Econometrica*, **20**, 680-84.

MESSNER, M., and M. POLBORN, (2004), "Voting on Majority Rules," *Review of Economic Studies*, **71**, 115-132.

RAE, D.W. (1969), "Decision-Rules and Individual Values in Constitutional Choice," *American Political Science Review*, vol. 63, pp. 40-56.

ROBERTS, K. (1999), "Dynamic Voting in Clubs," mimeo, London School of Economics.

SEIDMANN, D.J. (2004), "A Theory of Voting Patterns and Performance in Private and Public Committees", mimeo.

STAIGER, R. (1995), "International Rules and Institutions for Cooperative Trade Policy," in G. Grossman and K. Rogoff (eds.), *Handbook of International Economics*, vol.3, Amsterdam: North-Holland.

TAYLOR, M.J. (1969), "Proof of a Theorem on Majority Rule," *Behavioral Science*, vol. 14, pp. 228-231

WICKSELL, K. (1896), "A New Principle of Just Taxation," *Finanztheoretische Untersuchungen*, Jena.