Self Enforcing Voting in International Organizations *

Giovanni Maggi
Department of Economics
Princeton University
and
NBER

Massimo Morelli
Department of Economics and Political Science
The Ohio State University

April 30, 2003

Abstract

Some international organizations are governed by unanimity rule, some others by a majority system. Still others have moved from one system to the other over time. The existing voting models 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, both across organizations and over time.

*We thank James Anderson, Matt Jackson, Francois Maniquet, Guido Tabellini and Jean Tirole for very helpful comments. Arnaud Costinot provided excellent research assistance. We also benefited from comments by workshop participants at the Minneapolis Fed, Carnegie Mellon University, INSEAD, Namur University, Princeton University and Stockholm University. Morelli thanks the Deutsche Bank for sponsoring him as a member of the Institute for Advanced Study (Princeton) in 2001-02, when the project was conceived.

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. We distinguish between two types of governance mode: unanimity systems and majority systems. Some organizations, such as NATO and WTO, are governed by unanimity rule;[1] some others, such as the United Nations, are governed by some form of majority rule. Still others have seen important changes of governance mode over time: for example, the European Union has recently switched from unanimity to majority in several of its decision-making bodies, and the International Standards Organization has switched from unanimity to majority in the 70s. It is important to note that a unanimity system is qualitatively different from a majority system: the former requires only *coordination*; the latter requires also *enforcement*, to keep in check the minority members' temptation to defect.[2]

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 (or *sovereignty*) constraint. This model yields unanimity as the optimal system for a wide range of parameters, and yields rich predictions on the determinants of the cross-organization and over-time variation in the 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 antic-

[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. We note also that MERCOSUR and NAFTA are governed by unanimity as well.

[2] For this reason, in the notion of "majority" we include both the simple majority rule and supermajority rules.

ipate 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 change (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.

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 sovereignty constraint: a government must always have incentive to comply with the collective decision, even if it happens to be in the minority. 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 majority quota. 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 majority 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 majority, but we find that the opposite is true: a higher degree of correlation expands the range of discount factors for which the optimal self-enforcing institution is a majority system. The model thus predicts that a majority rule is more likely to be adopted in more homogenous organizations.

Next we consider the role of pure international transfers. Transfers can make it easier to satisfy the sovereignty constraint in a majority system, because they can be used to mitigate the minority members' temptation to defect. We show that the availability of transfers expands the range of discount factors for which a majority system is sustainable. Thus the model suggests that we should be more likely to observe a majority system in organizations that have the flexibility to enact

pure transfers between 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 sustainable rule.

The size of the organization interacts in interesting ways with the optimal governance mode. Since international organizations generally do not have open access, and the entry of new members is subject to the approval of current members, it is compelling to think about the organization size as endogenous. Specifically, we assume that in every period there is a random number (possibly zero) of new candidates for membership, and current members choose whether to admit the new candidates. We show that, for intermediate levels of the discount factor, the optimal self-enforcing voting rule is unanimity up to some (random) date and then switches to a majority rule. The reverse switch – from majority to unanimity – can never happen. The model thus offers a theoretical explanation for the "stylized fact" that international organizations tend to move from unanimity to majority, but not vice versa.

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 optimal to make decisions by voting. This is ensured by two ingredients of our model: (i) players have private information about their preferences; voting is then necessary to aggregate information and make efficient collective choices; and (ii) players have conflicting interests ex post but aligned interests ex ante.

Second, our paper contributes to the social choice literature. 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 the social choice literature, the paper that is perhaps colesest to ours is Barbera and Jackson [4]. They have a binary collective choice model where members' preferences on future issues are uncertain,[4] and each player is characterized by a distinct probability of being in favor of the status quo. They study self-stable voting rules, i.e. voting rules such that there is no alternative rule that would beat

[3] For models of self-enforcing trade agreements, see for example the survey by Staiger [18]. For models of international lending, see for example the survey by Eaton and Fernandez [10].

[4] For early discussions and motivations of models with a binary choice between alternatives with uncertain values in voters' minds, see Niemi and Weisberg [15], Badger [3], and Curtis [8].

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.

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 unanimity rule. The classic contributions by Wicksell [20] and Buchanan and Tullock [6] 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 (almost always) a majority rule, and unanimity may become optimal if there is no external enforcement.[5]

Two other papers that are related to ours are Roberts [17] and Barbera, Maschler and Shalev [5]. Both of these papers 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. These models differ from ours at least in three respects. First, the voting rule in these models is exogenous, and assumed to be a simple majority rule. Second, voting is only on admissions of new members, not on policy issues. Third, there are no issues of enforceability.

The paper is organized as follows. In section 2 we present a static model of collective action. First we solve for the first-best outcome, then we characterize the equilibria of the one-shot game, with and without enforcement. In section 3 we characterize the most cooperative self-enforcing voting rule in the repeated version of the game, when the size of the organization is exogenous and constant. In section 4 we study how the most cooperative voting rule evolves when the organization has opportunities to expand its membership over time. In section 5 we extend the

[5] Also Aghion and Bolton [1] and Guttman [12] argue that, in a veil-of-ignorance setting with perfect enforcement, majority generally dominates unanimity from the standpoint of ex-ante efficiency. These papers do not consider issues of enforceability. Other papers that provide theoretical justifications for (simple or super) majority rules are May [13], Rae [16], Taylor [19], Caplin and Nalebuff [7], Austen-Smith and Banks [2], Dasgupta and Maskin [9], and Messner and Polborn [14]. All of these models assume enforceability of majority decisions.

analysis to situations where the collective action may be effective even if not all members participate (the case of "impure" collective action). In section 6 we offer some concluding remarks.

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, or changing a common agricultural policy, or harmonizing a standard. Not taking the action $(a^i = 0)$ is interpreted as preserving the status quo. In this section and the next we assume that N is fixed, that is, we do not consider the possibility of entry or exit. In section 4 we will extend the model to allow for endogenous N.

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 are free to move within the union and member countries have a harmonized set of policies in areas such as trade, immigration and taxation. Any change in the common policies must be decided collectively, and if any member does not go along with the change, its benefits are compromised for the whole union. If for example the union decides to increase the common external tariff and one member does not go along, the effects of the tariff hike may be undone.[6]

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

[6] In principle, a country that is not willing to participate in the collective action could exit the union, and at that point the $N - 1$ countries will be able to act without the participation of the ex-member. However, exit from an economic union takes time and is costly, because it involves reinstating customs or other types of barriers between the exiting country and the union, thus it remains true that a defection by one country has disruptive effects on the collective 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. 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.

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 $n \in \{1, 2, ..., N-1+a^i\}$ denotes the total number of members taking action.[7]

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 on which side of the issue she will be on. We capture this idea by assuming that at the ex-ante stage $\theta = (\theta^1, ..., \theta^N)$ is a random vector distributed according to the common-knowledge probability distribution $P(\theta^1, ..., \theta^N)$ over support $\Theta = \{\theta_L, \theta_H\}^N$. 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 can think of θ as summarizing the relevant state of the world. In what follows we will often refer to θ simply as the "state".

2.1 First Best Outcome

The first best outcome is the mapping from states to actions that maximizes the members' common expected utility. Given our assumptions on payoffs, we can focus

[7] 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. If $\varepsilon = 0$, the only change is that complete inaction ($a^i = 0$ always) is a weakly dominated equilibrium, whereas with $\varepsilon > 0$ this is an undominated equilibrium. This would imply a different punishment strategy in the repeated game, but would not change the key insights of the model. 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 (possibly small) cost. This could be thought of as a "ratification" game, where not ratifying the initial decision implies a political cost. Also in this game, the status-quo outcome ($a^i = 0$ for all i) would be an undominated equilibrium.

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 θ 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

$$\max_{a(\theta)} \sum_{\theta \in \mathbf{\Theta}} P(\theta) a(\theta) [N^1(\theta)(B - \theta_L) + (N - N^1(\theta))(B - \theta_H)]$$

where $N^1(\theta)$ is the number of members that support the collective action. Note an assumption that is implicit in this formulation: we are restricting our attention to anonymous mappings $a(\theta)$; we are not considering for example rules such as "the action is taken whenever player 1 has $\theta = \theta_L$". Given this anonimity restriction, the unique Pareto-optimal allocation is identified by the above maximization problem. Alternatively, this can be interpreted as a symmetry restriction requiring that all members get the same expected utility.

Clearly, it is optimal to take the collective action in all the states where its aggregate benefit, $B \cdot N$, exceeds its aggregate cost, $N^1(\theta)\theta_L + (N - N^1(\theta))\theta_H$. This implies that it is efficient to take the collective action if and only if N^1 exceeds the quota $q^* \equiv \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil$, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

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

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, recall 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 strategy 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 payoff outcomes:

Proposition 2 *The worst equilibrium of the one shot game is: $a^i = 0$ in all subgames (*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).*

Note that 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 it does not deliver the first best outcome in general. 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 realizations of the θ^i. However, it is not hard to show that the first-best outcome can be implemented with the following voting rule: after uncertainty 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 $q^* = \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil$ members have voted in favor. The key is to note that, given the proposed voting rule, each player has incentive to vote sincerely.[8] Sincere voting then immediately implies the claim. Note the role of external enforcement: if the majority of the group votes in favor of the collective action, all the members that disagree are forced to participate. Without external enforcement, the minority could not be forced to go along with the majority. The following proposition summarizes:

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

We emphasize that the optimal enforceable voting rule is independent of the distribution over possible states, and in particular of the degree of correlation in the members' preferences. We also note that, if one disregards the integer constraint, the relative quota q^*/N is independent of the organization size N. As we will argue

[8] It is easy to see that voting sincerely is a weakly dominant strategy for each player. It should also be noted that there exist non-sincere voting equilibria, but these are characterized by weakly dominated strategies. For example, given $q^* > 1$ it is an equilibrium for everyone to vote "no" independently of θ^i, because in this case the probability of being pivotal for each player is zero.

later, the degree of correlation and the size of the organization will play a more critical role in the absence of external enforcement.

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 $q^* = \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil < N$.

We close this section with a remark. Under some conditions, the optimal voting rule may be submajoritarian, that is $q < N/2$. However, as the previous literature has pointed out, submajority rules can create instability in the decision-making process.[9] Since we think of a voting rule as a long-term decision-making procedure that must deal with many different issues, the nature of which is uncertain ex ante, it is reasonable to suppose that the designers of the institution would want to avoid any potential instability problem, and would therefore rule out submajority rules. This could be captured in the model by imposing a feasibility constraint $q \geq N/2$ on the choice of voting rule. Our results would then change in the direction of predicting a simple majority rule ($q = N/2$) when the unconstrained optimum is a submajority rule. To keep the exposition lean, however, we will simply assume $q^* = \lceil \frac{\theta_H - B}{\theta_H - \theta_L} N \rceil \geq N/2$, rather than imposing this constraint. In sum, we will assume throughout the paper that $N/2 \leq q^* < N$, or in words that the first-best voting rule is a majority rule.

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.

[9] Suppose for example that half of the members support the status quo ($a = 0$) and half of the members support change ($a = 1$) . A vote under a submajority rule will result in $a = 1$ being the new status quo. But then the other half of the members would support a change back to $a = 0$. This can potentially give rise to cycles back and forth between $a = 0$ and $a = 1$. See Barbera and Jackson [4] for a more thourough discussion of the problems associated with submajority rules.

We now suppose that the game described in section 2.2 is repeated infinite times, and introduce time subscripts in the notation. In each period, 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 $a_t^i \in \{0, 1\}$. The distribution of the vector θ_t (state of the world) 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. This is an assumption of *deep uncertainty* about the nature of future issues: players do not know which side of future issues they will be on, and today's issue is no indication of what future issues will be like.

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.[10] As in the previous section, we focus on equilibria that yield the same expected payoff to each member. Moreover we focus on stationary equilibria, in the sense that votes and actions at time t can depend only on the current state of the world θ_t. Within this class, we seek to characterize the *most cooperative* equilibrium, that is the one that maximizes the members' common expected payoff. This equilibrium is of particular interest for us, because it represents the upper bound to the efficiency that can be achieved without the help of external enforcement.

The first observation is that we can focus on punishment strategies that prescribe a permanent reversion to the status-quo equilibrium following any deviation (*trigger* punishment). This is because (a) the status quo equilibrium keeps the deviator at his maxmin payoff, and (b) since there will be no punishment episodes on the equilibrium path, it is best to punish deviations most severely. A theoretical limitation of trigger punishments is that they are not renegotiation-proof; we refer the reader to the concluding section for a discussion of this issue.

As far as the equilibrium path is concerned, it can be shown that there is no loss of generality in focusing on strategies where players behave according to a simple voting rule, and vote sincerely. More precisely, we can focus on equilibrium strategies with the following structure: (i) $a_t^i = 1 \ \forall i$ if $V_t^1 \geq q$ and $a_t^i = 0 \ \forall i$ if $V_t^1 < q$, where $V_t^1 = \#\{j : v_t^j = \theta_L\}$ and q is some integer in $\{1, 2, ..., N\}$; (ii) $v_t^i = \theta_t^i$. This class

[10] See Fudenberg and Tirole [11] for a definition and a discussion of this equilibrium notion.

of equilibrium strategies is indexed by the voting rule q. The search for the most cooperative equilibrium therefore boils down to a search for the most cooperative voting rule q. 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. The key step is to write the no-defect conditions for a given voting rule q. 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 $\theta_t^i = \theta_H$ and $V_t^1 \geq q$. The gain from cheating is $\theta_H - B$, and the discounted loss from cheating is $\frac{\delta}{1-\delta}U(q)$, where $U(q)$ denotes the one-period expected utility of the representative member given voting rule q. Clearly, the unanimity rule $q = N$ need not satisfy any constraint, thus the problem boils down to

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

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 most cooperative voting rule. If $q = q^*$ does not satisfy (2) then the most cooperative 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.**

This result suggests that a majority 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.

Notice also the *bang-bang* nature of the result: it is never optimal to choose a majority quota 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.

3.1 Correlation

A natural question is whether and how the correlation of members' preferences affects the optimal self-enforcing voting rule. 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 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 $\Pr(\theta^i = \theta_L) \equiv p$ for all i.

We have the following result:

Proposition 5 *Given (3), $\underline{\delta}$ is a decreasing function of ρ.*

Proof. Note that the one-period expected utility from using q^* can be written as

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

(3) implies that, as ρ increases, the first square bracket increases and the second square bracket decreases. It follows that $U^\rho(q^*)$ is increasing in ρ. Hence, the right hand side of (2) is increasing in ρ for every δ. The claim follows. **QED.**

Intuition might have suggested that a higher degree of correlation in the members' preferences makes unanimity more attractive relative to a majority 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.[11]

This result has an immediate corollary: if the discount factor is not too close to zero or to one, the optimal self-enforcing rule is unanimity for low values of ρ and majority for high values of ρ. We can make this statement more precise. Let $\underline{\delta}(\rho)$ denote the critical level of δ as a function of ρ.

Corollary 1 *If $\delta < \underline{\delta}(1)$, the optimal self-enforcing rule is unanimity for all ρ; if $\delta \in [\underline{\delta}(1), \underline{\delta}(0))$, there exists a critical level $\hat{\rho}$ such that the optimal self-enforcing rule is unanimity for $\rho < \hat{\rho}$ and $q = q^*$ for $\rho \geq \hat{\rho}$; if $\delta \geq \underline{\delta}(0)$, the optimal self-enforcing rule is $q = q^*$ for all ρ.*

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

3.2 Transfers

Thus far we have implicitly assumed that pure transfers are not available.[12] 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 transfers have hardly ever been used. Suppose now that such transfers are possible, and that they enter utility additively. Transfers may help sustain the first best outcome: if the majority wants to take the collective action and is willing to compensate the minority, the minority might be convinced to participate. However, the use of transfers is subject to two limitations. First,

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

[12] We are purposedly speaking of *pure* transfers, for the following reason. One might be tempted to argue that, even in the absence of pure transfers, if the organization has to decide on many different issues then there are ample opportunities for compensations between members, and therefore utility is effectively transferable. However this argument is correct only if the organization deals with many issues *simultaneously*. If issues come up sequentially – which we would argue to be more often the case – then the multiplicity of issues is already taken into account in our model, and the analysis of the previous section applies.

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 war may be tempted to vote against the collective action, hoping to get compensation.

A complete characterization of the optimal self-enforcing mechanism for all values of δ when transfers are available is a very difficult task, but we will show the following two results: (i) transfers help sustain the first best outcome, in the sense of expanding the range of discount factors for which the first best majority rule is sustainable, but (ii) for sufficiently low values of the discount factor, unanimity remains the best possible governance mode.

Consider the following timing for the stage game. After players observe their θ^i values, they vote. If the number of votes in favor of the collective action exceeds the quota q, the collective action is taken. Then each minority member gets a transfer.[13] Of course, the cost of transfers must be financed by the majority members. One can show that there is no loss of generality in assuming that the transfer is equal for all minority members, and the cost is split evenly among the majority members.[14] Note that the transfer can depend on the number of votes in favor of action, V^1. For this reason, we will speak of a "transfer scheme," and denote $t(V^1)$ the transfer received by each minority member. Observe right away that budget balance implies that each majority member must contribute an amount $\frac{(N-V^1)t(V^1)}{V^1}$. Also note that, since $t(V^1)$ is balanced and utility is transferable, the members' common expected utility associated with a pair $(q, t(\cdot))$ is simply $U(q)$.

We first show that there exists a transfer scheme such that q^* can be sustained for a wider range of δ than in the absence of transfers, unless preferences are perfectly correlated. 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. We will construct an equilibrium which entails the first best rule q^* for $\delta = \underline{\delta} - \varepsilon$, where $\varepsilon > 0$ is a small number. Consider a simple trigger punishment strategy whereby the organization reverts to the worst one-shot equilibrium if anyone refuses to participate in the collective action or to make a required transfer. The punishment is conditional only

[13] 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 member of the minority will face the same temptation to cheat as in the absence of transfers.

[14] This is because, starting from an uneven transfer scheme, we can always redistribute from the one that receives (pays) the most (least) to the one that receives (pays) the least (most) and still satisfy all incentive constraints.

on past actions and transfers, not on past votes. Then a pair $(q, t(\cdot))$ is part of an equilibrium if it satisfies the following incentive constraints:

$$\max_{V^1 \geq q} \left(\theta_H - B - t(V^1)\right) \leq \frac{\delta}{1-\delta} U(q) \tag{4}$$

$$\max_{V^1 \geq q} \left(\frac{(N - V^1)t(V^1)}{V^1}\right) \leq \frac{\delta}{1-\delta} U(q) \tag{5}$$

$$\sum_{V^1=q+1}^{N} \left(t(V^1 - 1) + \frac{(N - V^1)t(V^1)}{V^1}\right) \Pr(N^1 = V^1 | \theta^i = \theta_L) - (B - \theta_L) \Pr(N^1 = q | \theta^i = \theta_L) \leq 0 \tag{6}$$

The first constraint requires that a member who is against action have incentive to participate: the one-time gain from cheating is $\theta_H - B - t(V^1)$, and the loss from cheating is the future value of cooperation, $\frac{\delta}{1-\delta} U(q)$. The second condition requires that a member who is in favor of action have incentive to contribute to the cost of transfers. The gain from cheating here is given by the contribution, $\frac{(N-V^1)t(V^1)}{V^1}$, and the loss from cheating is again the future value of cooperation. The third constraint requires that a member who is in favor of action have incentive to vote sincerely. By voting strategically, this member gains $t(V^1 - 1) + \frac{(N-V^1)t(V^1)}{V^1}$ (he gets the transfer and avoids the contribution) in the event that he is not pivotal, that is when $V^1 > q$; and loses $(B - \theta_L)$ in the event that he *is* pivotal, that is when $V^1 = q$.

Notice that, if $\delta = \underline{\delta}$, the pair $(q, t(\cdot)) = (q^*, 0)$ satisfies (4) 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 constant 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 proves our first claim.

Our second claim is that, for sufficiently low values of δ, unanimity is the best possible governance mode. To see this, note that a necessary condition for a pair $(q, t(\cdot))$ with $q < N$ to be part of a PPE is that it satisfy constraints (4) and (5). Constraint (4) implies

$$\min_{V^1 \geq q} \; t(V^1) \geq \theta_H - B - \frac{\delta}{1-\delta} U(q) \tag{7}$$

This means that given any value of δ and any realization of V^1, the transfer must be strictly positive. Upon inspection of (5), however, it is easy to see that if $\delta = 0$,

$\max_{V^1 \geq q} t(V^1) \leq 0$. Therefore it is impossible to satisfy both (4) and (5) when δ goes to zero. It follows that unanimity is the only sustainable outcome.

The following proposition summarizes the main insight of this section:

Proposition 6 *The availability of transfers expands the range of δ for which the first best rule q^* is sustainable, but for low enough values of δ unanimity remains the optimal self-enforcing rule.*

Broadly interpreted, this result suggests that an organization is more likely to be governed by majority rather than unanimity if it has the flexibility to make pure transfers between its members.

4 Endogenous organization size

In this section we examine how the mode of governance evolves over time when the organization has opportunities to expand.

We think of the expansion of the organization as endogenous, in the following sense. The initial membership, N_0, is exogenously given. In each period $t \geq 1$, a random number $z_t \geq 0$ of new countries become candidates for membership. These may be countries that for some reason become interested in joining the organization at time t, or countries that become eligible for membership at time t as they meet requirements such as a good human rights record, a democratic system etc. We assume that z_t is *iid* across periods.

Candidates have the same utility function as current members. We assume that, if a candidate is rejected at t, it is eligible again at $t+1$ (thus if the union keeps rejecting applicants, the pool of applicants keeps getting larger). This is a natural assumption given that candidates are all ex-ante symmetric with respect to future issues, so that there is no point selecting among candidates. Letting $Z_t \equiv N_0 + \sum_{\tau=1}^{t} z_\tau$, the admission decision by the N_{t-1} current members boils down to choosing a number $N_t \leq Z_t$.

In this setting it is convenient to define the cost vector $\theta_t = \{\theta^i\}_{i=1}^{\infty}$ as including all potential members. We assume that θ_t is *iid* across periods. Also, in line with the assumption that all potential members are ex-ante symmetric, we assume that the distribution of θ_t is symmetric with respect to its components θ_t^i.

At the beginning of each period, z_t is realized, then current members decide how many of the candidates (if any) to admit. Then the state θ_t is realized, and the

organization votes on the "issue of the day." Since admissions take place before θ_t is realized, the current members' preferences on admissions are homogenous, so the admission decision is made to maximize the current members' common expected utility.

We allow the benefit from collective action to depend on N, and denote it by $B(N)$.[15] Consistently with the analysis in the previous sections, we assume (i) $\theta_L < B(N) < \theta_H$ for all N; this ensures that the collective action problem is interesting for all N; and (ii) $\frac{N}{2} \leq \lceil \frac{\theta_H - B(N)}{\theta_H - \theta_L} N \rceil < N$ for all N; that is, the first-best voting rule (conditional on N) is a majority rule.

Let $U(q, N)$ be the expected utility given voting rule q and membership N and $\hat{U}(N) = \max_q U(q, N)$. To rule out ties that would make the analysis more complicated, we assume the following genericity condition: $\hat{U}(N) \neq U(N', N')$ for all N and N'.

As in the case of fixed organization size, we can assume without loss of generality that any deviation from the equilibrium path is followed by a trigger punishment (permanent reversion to the status-quo equilibrium). We focus on equilibrium paths where the voting rule q and the organization size N depend only on the current value of the state variable Z_t. Given this restriction, we look for the most cooperative equilibrium path, that is the pair of functions $(q(Z_t), N(Z_t))$ that maximizes the members' common expected utility. This is given by the solution to the following constrained maximization problem:

$$\max_{q(Z_t), N(Z_t)} \sum_{t=1}^{\infty} \delta^{t-1} E[U(q(Z_t), N(Z_t))|Z_1] \tag{8}$$

subject to

$$\theta_H - B(N(Z_t)) \leq \sum_{\tau=1}^{\infty} \delta^{\tau} E[U(q(Z_{t+\tau}), N(Z_{t+\tau}))|Z_t] \text{ for all } Z_t \text{ such that } q(Z_t) < N(Z_t) \tag{9}$$

$$N(Z_t) \leq Z_t \text{ for all } Z_t \tag{10}$$

where the expectation E is taken with respect to future values of Z_t. Condition (9) is the incentive constraint at time t; it requires that a high-cost type is willing to participate in the collective action, given the future path of $q(Z_t)$ and $N(Z_t)$. As in the case of constant N, the unanimity rule ($q = N$) need not satisfy any incentive constraint. We have the following result:

[15] We could allow also the cost parameters θ_L and θ_H to depend on N, but this would make the analysis more cumbersome without adding much to the qualitative insights.

Proposition 7 *There exist critical levels δ^l and δ^h, with $0 < \delta^l \leq \delta^h < 1$, such that:*
(I) For $\delta < \delta^l$, the optimal self-enforcing voting rule is unanimity for all t;
(II) For $\delta^l \leq \delta < \delta^h$, the optimal self-enforcing voting rule is unanimity up to some (random) date $\hat{t}$ and then switches to a majority rule;
(III) For $\delta \geq \delta^h$, the optimal self-enforcing voting rule is a majority rule for all t.

Proof. See Appendix.

Parts (I) and (III) of this proposition are quite intuitive, given the previous result of proposition 4. Part (II) is more subtle. The key aspect of this result is that it can never be optimal to switch from majority to unanimity, hence if there is any regime switch it must be from unanimity to majority. The general intuition for this result is the following. The arrival of new candidates over time enlarges the organization's opportunity set (since candidates can be rejected), therefore the value of the organization cannot decrease over time, and hence the right hand side of the incentive constraint cannot decrease over time. If $B(N)$ is increasing, the left hand side of the incentive constraint $(\theta_H - B(N))$ is decreasing, therefore the incentive constraint can never get tighter over time, and the result follows right away. If $B(N)$ is decreasing, on the other hand, the key observation is that the unanimity payoff $U(N, N)$ is decreasing.[16] This implies that it cannot be optimal to switch from majority to unanimity at the same time as new members are admitted, because this is dominated by keeping N constant. But if N does not change, there is no reason to switch from majority to unanimity.

We emphasize that this result holds regardless of how $B(N)$ depends on N. Thus, the result is not driven by the increase in size *per se*, but by the interaction between the expansion process and the sovereignty constraint: in a world of enforceable voting, the optimal voting rule would always be q^*, which may increase or decrease with N.

The proposition does not ensure that there exist values of δ for which a regime switch occurs, because it may be $\delta^l = \delta^h$. When is $\delta^l < \delta^h$? This inequality will be satisfied if $B(N)$ is sufficiently increasing in the relevant range (more precisely, when $B(N) - B(N_0)$ is sufficiently large for some $N > N_0$). Intuitively, when this is the case, new candidates will be admitted, and as the organization expands the

[16] Recall that $U(N, N) = (B(N) - \theta_L) \Pr(\theta^1 = \theta^2 = ... = \theta^N = \theta_L)$ and note that $\Pr(\theta^1 = \theta^2 = ... = \theta^N = \theta_L)$ is nonincreasing in N.

incentive constraint gets relaxed, because the gain from cheating decreases and the value of the organization increases. Then there will be intermediate values of δ for which the first-best rule is initially not sustainable but it becomes sustainable later on.

At the outset of the paper we highlighted that some international organizations, such as the European Union and the International Standards Organization, have moved over time from unanimity to some type of majority rule. It is hard to think, on the other hand, of organizations that have experienced the reverse switch, that is from majority to unanimity. Our model offers a theoretical explanation for this "stylized fact," based on the interplay between the endogenous expansion of the organization membership and the presence of a self-enforcement constraint.

5 Extensions

5.1 Impure collective action

In this section we consider situations where the collective action may be effective even if not all members participate. We speak in this case of "impure collective action". We will revisit the results of the model, focusing first on the case of fixed membership size N, and then on the case of endogenous N.

Formally, let $b(n)$ denote the individual benefit generated if n members take action. We assume that $b(n)$ is increasing, with $b(0) = 0$. The convexity assumption captures the notion of "collective action," in the sense that the benefits of action are proportionally higher when a larger share of members participates. We note that this assumption is stronger than we need: the same results would obtain under the weaker assumption that $nb(n)$ is weakly convex; this is satisfied for example if $b(n) = n^\alpha$ with any $\alpha > 0$.

The utility function we are assuming is

$$U^i(a^i, n, \theta^i) = a^i \left(b(n) - \theta^i\right) \tag{11}$$

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 from high-θ 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.[17]

We present our results through a series of remarks, which are proved in the appendix. Let us start by looking at the one-shot game without enforcement. Let $n^{\min}$ denote the first integer n such that $b(n) \geq \theta^L$, and V^1 the number of "yes" votes.

Remark 1:

The worst equilibrium of the one-shot game is: $a^i = 0$ in all subgames (status-quo equilibrium).

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.

Note that, if $n^{\min} = N$, this equilibrium coincides with the "unanimity" equilibrium that we found in the case of pure collective action. More generally, action is taken by a "coalition of the willing." The only difference between this equilibrium and the simple unanimity equilibrium is that the minimum efficient size for the coalition of the willing ($n^{\min}$) may be lower than N. For this reason we feel justified in interpreting this as a *modified unanimity* equilibrium.

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

Remark 2:

[17] To see this latter point, note that a collective action problem with fully non-excludable benefits can be captured by the utility function $U = nb(n) - a^i\theta^i$. 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.

If external enforcement is available, the first-best outcome can be implemented by a voting rule with the following structure: There exist quotas q_1 and q_2, with $1 \leq q_1 \leq q_2 \leq N$, such that: 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.

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. Note that the intermediate interval of V^1 may be empty ($q_1 = q_2$), in which case the optimal mechanism is the same as in the case of pure collective action.

Next we consider self-enforcing voting rules. The question is to what extent proposition 4 is robust 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 "modified unanimity" rule described in Remark 1.

This is a generalization of the bang-bang result that we obtained in the case of 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. The only difference with respect to the case of pure collective action is that, both in the first-best voting rule and in the modified-unanimity rule, action may be taken by a "coalition of the willing" for certain realizations of the state of the world θ.

Finally, we reconsider the case of endogenous organization size with impure collective action. A similar result as proposition 7 holds:

Remark 4: Consider the model with endogenous size as in section 4, except that the assumption of "pure" collective action is replaced by that of "impure" collective action. Then the optimal self-enforcing rule is the modified-unanimity rule for $t < \hat{t}$ and the first-best rule for $t \geq \hat{t}$ (where $\hat{t}$ may be zero or infinity, in which case the rule never changes).

Intuitively, the same force that drives the result in the case of pure collective action is operating here: as the organization enjoys increasing opportunities to ex-

pand over time, the value of the organization can only grow, and hence sustaining the first-best voting rule can only get easier over time.

5.2 Renegotiation proofness

We conclude with a note on the issue of renegotiation-proof punishments. 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, because the continuation equilibrium is not Pareto-efficient.

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 would give the deviator her maxmin 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. Also note that, if N is endogenous and the organization is at steady state, then punishing a deviator with expulsion has a second-order impact on the utility of the remaining members, because N is at the level that maximizes the expected payoff of the representative member.

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.

6 Appendix

Proof of Proposition 7:

We start by proving the following

Lemma: Let $(q^o(Z_t), N^o(Z_t))$ denote the optimal plan: (i) The value function $V(Z_t) = \sum_{\tau=0}^{\infty} \delta^\tau E[U(q^o(Z_{t+\tau}), N^o(Z_{t+\tau}))|Z_t]$ is nondecreasing in Z_t; (ii) The instantaneous payoff $U(q^o(Z_t), N^o(Z_t))$ is nondecreasing in Z_t. (iii) The continuation value $\sum_{\tau=1}^{\infty} \delta^\tau E[U(q^o(Z_{t+\tau}), N^o(Z_{t+\tau}))|Z_t]$ is nondecreasing in Z_t.

Proof:
(i) We can write

$$\begin{aligned} V(Z_t) &= \sum_{\tau=0}^{\infty} \delta^\tau \int U(q^o(Z_{t+\tau}), N^o(Z_{t+\tau}))dF(Z_{t+\tau}|Z_t) \\ &= \sum_{\tau=0}^{\infty} \delta^\tau \int U(q^o(Z_t + \Delta_{t,\tau}), N^o(Z_t + \Delta_{t,\tau}))dF(\Delta_{t,\tau}|Z_t) \end{aligned}$$

where $\Delta_{t,\tau} = Z_{t+\tau} - Z_t$ and $F(\Delta_{t,\tau}|Z_t)$ is the c.d.f. of $\Delta_{t,\tau}$ conditional on Z_t. Note that, given the *iid* assumption, $F(\Delta_{t,\tau}|Z_t)$ is independent of Z_t.

Now suppose Z_t is increased to $Z'_t = Z_t + \Delta$, with Δ a positive number. Given this change, the new (random) value of $Z_{t+\tau}$ is $Z'_{t+\tau} = Z'_t + \Delta_{t,\tau} = Z_t + \Delta + \Delta_{t,\tau}$.

We need to show that $V(Z_t + \Delta) \geq V(Z_t)$. We do this by displaying a feasible contingency plan that, when starting from state $Z'_t = Z_t + \Delta$, attains value $V(Z_t)$. Consider the following plan: $(q'(Z'_{t+\tau}), N'(Z'_{t+\tau})) = (q^o(Z'_{t+\tau} - \Delta), N^o(Z'_{t+\tau} - \Delta)) = (q^o(Z_t + \Delta_{t,\tau}), N^o(Z_t + \Delta_{t,\tau}))$. In words, we are "ignoring" the increment Δ, so this plan yields the same path for q and N starting from $Z_t + \Delta$ as did the original plan starting from Z_t, for any realization of $\{\Delta_{t,\tau}\}_{\tau=1}^{\infty}$. Clearly, this plan still satisfies (10). The value of the objective with this plan is

$$\sum_{\tau=0}^{\infty} \delta^\tau \int U(q^o(Z_t + \Delta_{t,\tau}), N^o(Z_t + \Delta_{t,\tau}))dF(\Delta_{t,\tau}|Z_t + \Delta)$$

Since $F(\Delta_{t,\tau}|Z_t + \Delta) = F(\Delta_{t,\tau}|Z_t)$, this value is equal to $V(Z_t)$. It remains to argue that the proposed plan is incentive compatible at all dates. This follows

from the facts that (i) the original plan $(q^o(\cdot), N^o(\cdot))$ is incentive compatible at all dates, and (ii) by the *iid* assumption, Δ does not affect the distribution of the future increments $\Delta_{t,\tau}$, $\tau = 1,...,\infty$, therefore for all $s \geq t$ the continuation value $\sum_{\tau=1}^{\infty} \delta^\tau E_s[U(q^o(Z_{s+\tau}), N^o(Z_{s+\tau}))]$ is independent of Δ. We can conclude that Δ does not affect either the LHS or the RHS of the incentive constraint at any $s \geq t$.

(ii) Suppose by contradiction that there exist Z_{t-1} and Z_t such that $U(q^o(Z_{t-1}), N^o(Z_{t-1})) > U(q^o(Z_t), N^o(Z_t))$. Now consider an alternative plan $(q'(\cdot), N'(\cdot))$ that is identical to $(q^o(\cdot), N^o(\cdot))$ except that $(q'(Z_t), N'(Z_t)) = (q^o(Z_{t-1}), N^o(Z_{t-1}))$. This strictly increases the instantaneous payoff at Z_t, and hence the overall value $V(Z_0)$. The new plan clearly satisfies (10). Since the current payoff at t is being increased, all incentive constraints for previous dates are still satisfied. We only need to argue that the incentive constraint at t is still satisfied. If $q^o(Z_{t-1}) = N^o(Z_{t-1})$ (unanimity) this is obvious. Focus then on the case $q^o(Z_{t-1}) < N^o(Z_{t-1})$ and suppose the incentive constraint at t is not satisfied. Then, since the original plan is incentive compatible at $t-1$, it must be that the continuation value for the new plan at t is lower than the continuation value for the original plan at $t-1$:

$$\begin{aligned}\sum_{\tau=1}^{\infty} \delta^\tau E\left[U(q'(Z_{t+\tau}), N'(Z_{t+\tau}))|Z_t\right] &= \sum_{\tau=1}^{\infty} \delta^\tau E\left[U(q^o(Z_{t+\tau}), N^o(Z_{t+\tau}))|Z_t\right] \\ &< \sum_{\tau=1}^{\infty} \delta^\tau E\left[U(q^o(Z_{t+\tau-1}), N^o(Z_{t+\tau-1}))|Z_{t-1}\right]\end{aligned}$$

Since $(q'(Z_t), N'(Z_t)) = (q^o(Z_{t-1}), N^o(Z_{t-1}))$, this in turn implies

$$\sum_{\tau=0}^{\infty} \delta^\tau E\left[U(q^o(Z_{t+\tau}), N^o(Z_{t+\tau}))|Z_t\right] < \sum_{\tau=0}^{\infty} \delta^\tau E\left[U(q^o(Z_{t+\tau-1}), N^o(Z_{t+\tau-1}))|Z_{t-1}\right]$$

which is impossible given part (i) of this lemma. We can conclude that plan $(q'(\cdot), N'(\cdot))$ is incentive compatible, and hence it dominates plan $(q^o(\cdot), N^o(\cdot))$. But this contradicts the optimality of $(q^o(\cdot), N^o(\cdot))$.

(iii) Given the *iid* assumption, increasing Z_t shifts the conditional distribution of $Z_{t+\tau}$ in a first-order stochastic sense for all τ, that is, $F_{Z_{t+\tau}|Z_t=Z'}$ FSD $F_{Z_{t+\tau}|Z_t=Z}$ if and only if $Z' \geq Z$. This, together with the fact that $U(q^o(Z_t), N^o(Z_t))$ is nondecreasing in Z_t (part (ii) of this lemma), implies the claim. **Q.E.D.**

We turn now to the proof of the proposition. The first remark is that we can focus without loss of generality on two voting rules: unanimity, that is $q = N$, and the first-best rule conditional on N, that is $q^*(N) = \lceil \frac{\theta_H - B(N)}{\theta_H - \theta_L} N \rceil$. This is because

(a) if any $q < N$ is sustainable at a given date, so is $q^*(N)$ at that date; and (b) if $q^*(N)$ is incentive compatible at a given date then it is also optimal at that date.

The proof will proceed in five steps. We will show that: (i) for δ sufficiently low, $q(Z_t) = N(Z_t)$ for all Z_t at an optimum; (ii) for δ sufficiently high, $q(Z_t) = q^*(N(Z_t))$ for all Z_t at an optimum; (iii) if for $\delta = \delta'$ the optimum entails $q(Z_t) = N(Z_t)$ for all Z_t, then the same is true for all $\delta < \delta'$; (iv) if for $\delta = \delta''$ the optimum entails $q(Z_t) = q^*(N(Z_t))$ for all Z_t, then the same is true for all $\delta > \delta''$; (v) it is never optimal to switch from a majority rule to unanimity at any point in time. The claim will then follow.

(i) If δ is sufficiently close to zero, any rule other than unanimity violates (9), therefore permanent unanimity is the only feasible solution. This is ensured because $B(N)$ is bounded, hence U is bounded.

(ii) Consider the solution of the problem without constraint (9). Clearly, this plan entails $q(Z_t) = q^*(N(Z_t))$ for all Z_t. If δ is sufficiently close to one, (9) is not binding at this plan, hence this is also the solution of the constrained problem. We can conclude that the optimal voting rule is $q(Z_t) = q^*(N(Z_t))$ for all Z_t.

(iii) Let us denote a contingency plan with $S(Z_t) = (q(Z_t), N(Z_t))$ and the associated utlity with $\widetilde{U}(S(Z_t)) = U(q(Z_t), N(Z_t))$. Let us call $S^U(Z_t) = (N^U(Z_t), N^U(Z_t))$ the optimal plan for $\delta = \delta'$. Suppose by contradiction that there is some $\delta'' < \delta'$ such that the optimal plan – call it $S^m(Z_t)$ – entails majority for some Z_t^0. Because $S^U(Z_t)$ is feasible for all δ, we must have $\widetilde{U}(S^m(Z_t)) \geq \widetilde{U}(S^U(Z_t))$ for all Z_t, with strict inequality for $Z_t = Z_t^0$ from our genericity condition. But if $S^m(.)$ is feasible for $\delta = \delta''$, it remains feasible for $\delta = \delta' > \delta''$, a contradiction with $S^U(.)$ being optimal for $\delta = \delta'$.

(iv) Suppose by contradiction that for δ'' the optimal plan $S^M(Z_t)$ entails a majority rule for all Z_t, and for some $\delta' > \delta''$ the optimal plan $S^u(Z_t)$ entails unanimity for some Z_t^0. This implies $\widetilde{U}(S^M(Z_t^0)) > \widetilde{U}(S^u(Z_t^0))$, for otherwise we could improve on $S^M(Z_t)$ by replacing $S^M(Z_t^0)$ with $S^u(Z_t^0)$; this would be feasible and would improve the value of the objective. We will now show that $S^u(Z_t)$ cannot be optimal for $\delta = \delta'$. Consider an alternative plan $S^{'}(.)$ such that $S^{'}(Z_t) = S^u(Z_t)$ for all $Z_t \neq Z_t^0$ and $S^{'}(Z_t^0) = S^M(Z_t^0)$. Clearly, if $\sum_{\tau=1}^{\infty} \delta'^{\tau} E\left[\widetilde{U}(S^u(Z_t))|Z_t^0\right] \geq \sum_{\tau=1}^{\infty} \delta'^{\tau} E\left[\widetilde{U}(S^M(Z_t))|Z_t^0\right]$, $S^{'}(Z_t^0)$ is feasible. Since for $Z_t < Z_t^0$ the incentive constraint has been relaxed, $S^{'}(.)$ is feasible for $\delta = \delta'$, a contradiction. Now, if $\sum_{\tau=1}^{\infty} \delta'^{\tau} E\left[\widetilde{U}(S^u(Z_t))|Z_t^0\right] < \sum_{\tau=1}^{\infty} \delta'^{\tau} E\left[\widetilde{U}(S^M(Z_t))|Z_t^0\right]$, we can improve on $S^u(Z_t)$ with a plan $S^{''}(\cdot)$ such that $S^{''}(Z_t) = S^M(Z_t)$ for $Z_t \geq Z_t^0$ and $S^{''}(Z_t) =$

$S^u(Z_t)$ for $Z_t < Z_t^0$. Since $S^M(\cdot)$ is feasible for $\delta = \delta''$, it is also feasible for $\delta' > \delta''$, and hence $S^{''}(Z_t)$ is feasible for $Z_t \geq Z_t^0$. Moreover, since the value function at Z_t^0 has been increased, $S^{''}(Z_t)$ is feasible also for $Z_t < Z_t^0$. It follows that the value of the whole program has been increased, a contradiction.

(v) Suppose by contradiction that at the optimal plan $(q^o(\cdot), N^o(\cdot))$ there exist Z_t and Z_{t-1} such that $q^o(Z_{t-1}) < N^o(Z_{t-1})$ and $q^o(Z_t) = N^o(Z_t)$. We need to distinguish three cases:

(v_a) $B(N^o(Z_t)) \geq B(N^o(Z_{t-1}))$. In this case we can do better with a plan $(q'(\cdot), N'(\cdot))$ that is identical to the original one except that $q'(Z_t) = q^*(N^o(Z_t))$. This increases the instantaneous payoff at Z_t and satisfies (10). We need to show that it is incentive compatible. The fact that the instantaneous payoff at t is higher under the new plan ensures that the incentive constraint is satisfied for all dates before t. The new path also satisfies the incentive constraint at t. To see this, compare it with the incentive constraint at $t-1$ for the old plan: since $B(N^o(Z_t)) \geq B(N^o(Z_{t-1}))$, the left hand side is weakly lower; and by point (iii) of the lemma above, the right hand side is weakly higher. We can conclude that the new plan dominates the original plan, which contradicts the optimality of the latter.

(v_b) $B(N^o(Z_t)) < B(N^o(Z_{t-1}))$ and $N^o(Z_t) > N^o(Z_{t-1})$. In this case we can do better with a plan $(q'(\cdot), N'(\cdot))$ that is identical to the original one except that $N'(Z_t) = N^o(Z_{t-1})$; that is, no new members are admitted at Z_t. To see this, recall that $U(N, N) = \Pr_N(\theta^1 = ... = \theta^N = \theta_L)(B(N) - \theta_L)$. The first factor is always decreasing in N, therefore the proposed change increases instantaneous payoff at Z_t. The proposed plan clearly also satisfies all constraints, therefore the optimality of the original plan is contradicted.

(v_c) $B(N^o(Z_t)) < B(N^o(Z_{t-1}))$ and $N^o(Z_t) < N^o(Z_{t-1})$. Then it must be that $(q^o(Z_{t-1}), N^o(Z_{t-1}))$ implies a higher instantaneous payoff than $(q^o(Z_t), N^o(Z_t))$, for otherwise $(q^o(Z_t), N^o(Z_t))$ would have been chosen already at Z_{t-1} (clearly all constraints would still have been satisfied). But this contradicts part (ii) of the lemma above. **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 $N^0 = N - N^1$ 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:

$$\max_{n^0,n^1} J(n^0,n^1) \equiv n^1[b(n^0+n^1)-\theta^L]+n^0[b(n^0+n^1)-\theta^H] \quad (12)$$

$$\text{s.t. } 0 \leq n^0 \leq N-N^1, 0 \leq n^1 \leq N^1 \quad (13)$$

Since $b(n)$ 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 $N^1<q^*$ and $g^1(N^1)>0$ for $N^1\geq q^*$. (Note that 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 $N^1>q^*$ 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.

Finally, it is direct to verify that the first-best outcome we just described can be implemented with the voting rule proposed in remark 2.

Proof of Remark 3: The strategies on the equilibrium path can be written as $\sigma=\{a^i(\theta_t^i,v_t),v^i(\theta_t^i)\}_{i=1}^N$.

Let $U^i(\sigma)$ denote the one-period expected utility of member i given σ. Given that we are restricting to strategies that give all players the same expected payoff,

we have $U^i(\sigma) = U(\sigma)$ for all i. The key is to argue that we can focus on two profiles: (i) the profile that corresponds to the modified-unanimity rule defined in Remark 1 – let σ^u denote such profile; and (ii) the profile that corresponds to the first-best voting rule defined in Remark 2 – let σ^* denote such profile.

Clearly, if σ^* is self-enforcing, then it is optimal, since it maximizes $U(\sigma)$. The self-enforcement condition for σ^* is

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

The left hand side is the one-period gain from cheating, which occurs when a θ_H type is called to action.

An alternative profile σ can be preferred to σ^* 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 profile σ 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 σ^u. It is also clear that σ^u implies no unilateral incentive to deviate, since it is an equilibrium of the stage game. Therefore the only candidates for an optimum are σ^u and σ^*. We can easily conclude that σ^* is optimal if δ is higher than a critical level, and σ^u is optimal otherwise. **QED.**

Proof of Remark 4: The proof of Proposition 7 up to point (iv) applies identically to the case of impure collective action, provided "unanimity" rule is replaced with "modified unanimity" rule, and the majority rule is replaced with the first-best rule described in remark 2. As for point (v) of the proof, notice that the relevant gain from cheating with the first-best rule is now $\theta_H - b(N)$. This is because a high-cost type is called to action only if everyone else is called to action. Next note that, since $b(n)$ is increasing, the LHS of the incentive constraint is decreasing in N. Using the same reasoning as in part (v_a) of the proof, the claim follows. **QED.**

References

[1] AGHION, P. and P. BOLTON (2002), "Incomplete Social Contracts," mimeo.

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

[3] BADGER, W.W. (1972) "Political Individualism, Positional Preferences, and Optimal Decision-Rules," in *Probability Models of Collective Decision Making*, edited by R.G. Niemi and H.F. Weisberg, Merrill Publishing: Columbus Ohio.

[4] BARBERA, S. and M.O. JACKSON (2002), "Choosing how to Choose: Self-Stable Majority Rules," mimeo, Caltech.

[5] BARBERA, S., M. Maschler and J. Shalev (2001), "Voting for Voters: A Model of Electoral Evolution," *Games and Economic Behavior*, **37**, pp. 40-78.

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

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

[8] CURTIS, R.B. (1972), "Decision Rules and Collective Values in Constitutional Choice," in *Probability Models of Collective Decision Making*, edited by R.G. Niemi and H.F. Weisberg, Merrill Publishing: Columbus Ohio.

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

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

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

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

[13] MAY, K.O. (1952), "A set of Independent, Necessary and Sufficient Conditions for Simple Majority Decisions," *Econometrica*, vol. 20, pp. 680-684.

[14] MESSNER, M., and M. POLBORN, (2003), "Voting on Majority Rules," *Review of Economic Studies*, forthcoming.

[15] NIEMI, R.G. and H.F. WEISBERG (1972), "Substantive Applications of Collective Decision-Making," in *Probability Models of Collective Decision Making*, edited by R.G. Niemi and H.F. Weisberg, Merrill Publishing: Columbus Ohio.

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

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

[18] 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.

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

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