**Preprints of the**
**Max Planck Institute for**
**Research on Collective Goods**
**Bonn 2010/02**

# Public-Good Provision in a Large Economy

Felix Bierbrauer
Martin Hellwig

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2010/02

# Public-Good Provision in a Large Economy

Felix Bierbrauer / Martin Hellwig

December 2009

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

# Public-Good Provision in a Large Economy*

Felix J. Bierbrauer† and Martin F. Hellwig‡

Max Planck Institute, Bonn

December 2009

**Abstract**

We propose a new approach to the normative analysis of public-good provision in a large economy. Our analysis is based on a mechanism design approach that involves a requirement of coalition-proofness, as well as a requirement of robustness, so that the mechanism must not depend on specific assumptions about individual beliefs. Our main result shows that such a mechanism can condition only on the population shares of people with valuations above and below the per capita provision costs. This suggests an intriguing link between mechanism design for large economies and voting.

*Keywords:* Public-good provision, Mechanism Design, Large Economy

*JEL:* D60, D70, D82, H41

*We are grateful for discussions with and comments from Alia Gizatulina, Mike Golosov, Kristoffel Grechenig, Christian Hellwig, David Martimort, Benny Moldovanu, and Nora Szech.

†Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Email: bierbrauer@coll.mpg.de

‡Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Email: hellwig@coll.mpg.de

# 1 Introduction

In this paper, we propose a new approach to the normative analysis of public-good provision in large economies. By a large economy, we understand an economy with many people in which each individual is too insignificant to have a noticeable effect on variables such as the prices of private goods or the provision levels of public goods. We consider the large-economy paradigm to be appropriate for studying how a society with millions of people can best determine the appropriate levels of resources that are to be devoted to matters such as national defense or the court system, which concern the entire population. We also believe that, when applied to a large economy, the standard mechanism design approach to public-good provision provides unsatisfactory results.

The standard mechanism design approach to public-good provision focusses on issues of *individual incentive compatibility*. Under asymmetric information about individual preferences, the question is whether individuals have proper incentives to provide the system as a whole with the information about preferences that it needs for efficient public-good provision. In a "small" economy, in which each individual has a distinct chance of being "pivotal", i.e., of having a noticeable effect on the provision of a public good, this requires that people's financial contributions must be precisely calibrated to their expressions of preferences. The calibration must be such that people have neither an incentive to overstate their preferences on the assumption that the increase in public-good provision is paid for by somebody else nor an incentive to understate their preferences on the assumption that the money which they can thereby save is worth more than the reduction in public-good provision. The implications of this requirement have been thoroughly explored in the literature. It is well known that an efficient provision rule can be implemented if the calibration of payments to expressed preferences is such that people are induced to take account of the external effects that they impose on others whenever they are "pivotal" for the provision or non-provision of the public good.[1]

In a large economy, these concerns are moot. In such an economy, any notion that a person's payments should be calibrated to the effects that this person's communication about her preferences have on the provision of the public good leads to the simple conclusion that her payment should be independent of what she communicates. If what a person says affects neither the collective decision on public-good provision nor the payments she has to make, individual incentive compatibility is trivial. If what she says is deemed to have no effect whatsoever, she may as well tell the truth. The information that is thus communicated is sufficient to implement an efficient provision rule for the public good. Participation in the system may not be voluntary, but there is no problem of incentive compatibility.[2]

We want to take issue with this view. The following example illustrates our concerns. Sup-

[1] For implementation in dominant strategies, see Clarke (1971), Groves (1973), Green and Laffont (1979), for (interim) Bayes-Nash implementation, see d'Aspremont and Gérard-Varet (1979). More recently, Bergemann and Morris (2005) have studied interim implementation with a requirement of robustness with respect to the specification of agents' beliefs about the other participants.

[2] We do not insist on voluntary participation. Participation constraints are irrelevant if the state has powers of coercion and these powers can be used to make people contribute to financing a public good even when it does not benefit them.

pose that the public good in question comes as a single indivisible unit. The provision cost *per capita* of the population is 4. A fraction $\frac{3}{10}$ of the population assigns a value of 10 to the public good, a fraction $s$ a value of 3, and a fraction $\frac{7}{10} - s$ a value of 0. An efficient provision rule stipulates that the public good should be provided if the average *per capita* valuation exceeds 4, and that it should not be provided if the average *per capita* valuation is less than 4. In other words, the public good should be provided if $s > \frac{1}{3}$ and should not be provided if $s < \frac{1}{3}$. The requisite resources can be obtained by imposing a payment rule under which everybody pays 4 if the public good is provided and 0 if it is not provided. If people believe that, individually, they are too insignificant to affect the provision of the public good, a mechanism involving this provision and payment rule is incentive-compatible.

If $s$ is common knowledge, this reasoning is unproblematic. This is the case, for instance, if we think of the large-economy model as a limit of finite-economy models with independent private values in which the number of participants becomes large. However, if $s$ is common knowledge, the implementation of an efficient provision rule does not require any information from participants because, even before any such information is provided, it is commonly known whether the public good should be provided or not.[3]

By contrast, if $s$ is the realization of a nondegenerate random variable $\tilde{s}$, the problem of whether the public good should be provided or not involves a genuine information problem. In this case, the information whether the public good should be provided or not must be inferred from the participants' reports about their preferences. If the fraction of people reporting a valuation of 3 exceeds $\frac{1}{3}$, one may infer that $s > \frac{1}{3}$ and that the public good should be provided.

At this point, we are bothered by the notion that efficient provision can be implemented with a payment rule under which everybody pays 4 if the public good is provided and 0 if it is not provided. Why should people with a valuation of 3 report this valuation honestly? Reporting a valuation of 3 contributes to making provision of the public good more likely, if only infinitesimally. If the public good is provided, these people enjoy a benefit of 3 and have to pay 4, for a net payoff equal to $-1$. Each one of them would be better off if the public good was not provided. Moreover, the public good would indeed not be provided if each one of these people reported a valuation of 0. Why, then, should they report honestly, rather than claiming that the public good is worth nothing to them?

If individual incentive compatibility is the only requirement for the public-good provision mechanism, the answer to this question is that nobody minds reporting his or her valuation honestly because nobody feels that his or her report will make a difference to anything anyway. We consider this answer to be unconvincing. Therefore, we propose a new approach to the analysis of public-good provision in a large economy.

This new approach involves requirements of coalition-proofness and of robustness in the sense of Ledyard (1978) and Bergemann and Morris (2005) that are imposed in addition to individual incentive compatibility. Coalition-proofness requires that individuals must not have an incentive to coordinate their behavior so as to manipulate jointly the outcomes of the allocation mecha-

[3]Thus, in models with independent private values, the problem of whether to provide the public good or not becomes moot if one takes limits as the number of participants becomes large and the law of large numbers sets in.

nism. Robustness ensures that the working of a mechanism does not depend on the availability of detailed information about individual beliefs.

To see the relevance of coalition-proofness, note that in the given example, the people who value the public good at either 0 or 3 have an incentive to sabotage the efficient provision rule with equal cost sharing by forming a coalition to coordinate reports in such a way that the fraction of people reporting 3 is always below $\frac{1}{3}$. By contrast to cartel formation in industrial economics, the distorted reports that this sabotage action requires would all be individually incentive-compatible. Therefore, the efficient provision rule with equal cost sharing is not coalition-proof.

Indeed, in the given example, it is impossible to have a coalition-proof rule with equal sharing of public-good provision costs that conditions public-good provision on $s$. More generally, we will show that, under our requirements of coalition-proofness and robustness, public-good provision can be conditioned on the sizes of the set of people who are net beneficiaries of public-good provision and of its complement, the set of people who are harmed by public-good provision, but not on any additional information, e.g., information about the intensities of people's likes and dislikes. In the example, the two relevant sets have sizes $\frac{7}{10}$ and $\frac{3}{10}$, regardless of $s$, and the mechanism designer is reduced to a rule that stipulates public-good provision or not, depending on whether the ex ante expectation of people's valuation of the public good is greater or less than the per capita cost 4, or, equivalently, whether he considers the ex ante expectation of $\tilde{s}$ to be greater or less than $\frac{1}{3}$.

A requirement of coalition-proofness has previously been introduced by Laffont and Martimort (1997, 2000). Our approach differs from theirs in that we focus on coalitions consisting of subsets of the entire population, with coalition membership depending on people's types. Thus, in the above example, we considered a coalition of all people who value the public good at 0 or 3. By contrast, Laffont and Martimort focussed on coalitions of all people, regardless of their types. This focus was appropriate for their purpose, which was to eliminate the possibility, established by Crémer and McLean (1985, 1988), that the mechanism designer might exploit the slightest correlations in individual preferences in order to appropriate the entire surplus that is generated.

In our analysis, coalition-proofness is used as a device to articulate the inherent conflict between people who benefit from public-good provision and people who are harmed by it, rather than a device to prevent the mechanism designer from appropriating rents from everybody. Therefore, we focus on coalitions of subsets of people with common interests. For such subsets, it is natural to have coalition membership depend on people's types. Thus, in the given example, the common interests of people who are harmed by public-good provision are put into focus by a concept of coalition-proofness that allows for collective manipulations of individual reports by the coalition of people who value the public good at 0 or at 3.

We do follow Laffont and Martimort, however, in requiring that coalition formation and the behavior of coalition members satisfy the same information and incentive constraints as the underlying incentive mechanism itself. In particular, we require that the decision to join a coalition and the behavior as a coalition member must be individually incentive-compatible. The information problems of coalition formation and behavior are actually more complex in our

setting than in Laffont and Martimort (1997, 2000) because, apart from problems of individual incentive compatibility of stipulated behaviors of coalition members, a coalition that consists of a subset of the population also must deal with the problem that its information about people outside the coalition is incomplete.

In addition to coalition-proofness, we also impose a requirement of *robustness* of incentive compatibility in the sense of Ledyard (1978) and Bergemann and Morris (2005). Outcomes are allowed to depend *only* on those aspects of the participants' types that are relevant for their payoffs, i.e., their public-goods preferences. They are *not* allowed to depend on other aspects of the participants' types such as the beliefs that they have about other people's payoffs or other people's beliefs. Moreover, the outcome function must be individually incentive-compatible regardless of how the non-payoff-relevant aspects of people's types are specified.

As we explain elsewhere (Bierbrauer and Hellwig 2009), robustness of incentive compatibility is a desirable property because it eliminates the dichotomy between specifications with independent values and specifications with correlated values. Without robustness, for models with participation constraints, the Bayesian approach yields impossibility theorems for first-best implementation with independent values and possibility theorems for first-best implementation with correlated values. If robustness of incentive compatibility is imposed, this dichotomy disappears. Regardless of whether values are independent or correlated, for large economies, one finds that, with participation constraints, first-best implementation is possible for private goods and impossible for public goods.

In the context of a large economy, robustness implies that people's payments cannot be made to depend on their types. This is in line with the notion that payments should be calibrated to the effects that this person's communication about her preferences have on the provision of the public good, which in a large economy are zero. Deviations from this principle could be incentive-compatible, if, conditional on their types, people have different beliefs about the state of the economy and the prospects for public-good provision and a type dependence of payments allows them to bet on the differences in beliefs. However, the incentive-compatibility compatibility of such deviations is not robust to changes in the specification of beliefs.

To explain the issue, we return to the above example and consider a type-dependent payment rule that requires people who value the public good at 3 to pay 0 if the public good is provided and to pay 8 if the public good is not provided. People who value the public good at 0 or 10 pay 10 if the public good is provided and receive 2 if the public good is not provided. Under this rule, people who value the public good at 3 are no longer averse to having revealed that $s$ is greater and not less than $\frac{1}{3}$. When the public good is provided, their net payoff is equal to 3, when the public good is not provided, their net payoff is equal to $-8$.

If the random variable $\tilde{s}$ can only take the values $\frac{2}{10}$ and $\frac{6}{10}$, the combination of this type-dependent payment rule with an efficient provision rule, providing for non-provision if $\tilde{s} = \frac{2}{10}$ and for provision if $\tilde{s} = \frac{6}{10}$, is also compatible with budget balance.

The resulting mechanism is incentive-compatible if type-dependent beliefs are derived from a common prior that assigns probability one half to each of the two possible realizations of $\tilde{s}$ and that assigns values $\frac{7}{10} - s, s$, and $\frac{3}{10}$ to any one person's conditional probabilities, given the event $\tilde{s} = s$, of having valuations 0, 3 and 10. Given this common prior, the probability of public-good

provision, i.e., of the event $\tilde{s} = \frac{6}{10}$, is assessed at $\frac{1}{6}$ by a person with valuation 0, at $\frac{3}{4}$ by a person with valuation 3, and at $\frac{1}{2}$by a person with valuation 10. These differences in beliefs allow for an incentive-compatible dependence of payments on types. The resulting payments scheme can be interpreted as a combination of sharing of the cost of efficient public-good provision and a system of bets on the state of the economy.[4]

However, if the common prior were to assign probabilities one third to the event $\tilde{s} = \frac{2}{10}$ and two thirds to the event $\tilde{s} = \frac{6}{10}$, the given scheme would no longer be incentive-compatible. With beliefs determined by the prior $(\frac{1}{3}, \frac{2}{3})$, the people who value the public good at 10 would consider the payment scheme that is meant for people with valuation 3 to be more attractive than the payment scheme that is meant for themselves. The incentive compatibility of the given type-dependent payment scheme is thus not robust to changes in the specification of beliefs.

More generally, robustness implies that payments must be type-independent. Robustness also enables us to establish a revelation principle for coalition-proof mechanisms. Type independence of payments provides a natural basis for assessing coalition-proofness. Under coalition-proofness, the type-independent payments can only depend on whether the public good is provided or not. Thus, payments might be equal to zero if the public good is not provided and to the per-capita cost if it is provided. There is then a sharp distinction between people whose net payoffs are increased and people whose net payoffs are decreased by the provision of the public goods. These two groups define the key coalitions to consider in assessing whether a provision rule for the public good is coalition-proof.

The main result of our analysis shows that, if one imposes coalition-proofness, as well as robust individual incentive compatibility and anonymity, then the sizes of the two groups, the group of people who are harmed by public-good provision and the group of people who benefit from public-good provision, represent the *only* information that can be used in determining whether the public good is to be provided or not. By contrast, information concerning the intensity of likes and dislikes cannot be used. Apart from exceptional circumstances, therefore, it is impossible to implement a first-best provision rule by a coalition-proof, robustly incentive-compatible anonymous mechanism. By contrast to previous impossibility results, this finding does not involve any participation constraints or a multi-dimensional information problem. Instead, it follows from the observation that coalition-proofness and robust incentive compatibility together destroy the possibility of conditioning on intensities of preferences.

Mechanisms that condition the provision of the public good on the numbers of its adherents and its opponents are reminiscent of voting mechanisms. In our analysis, optimal coalition-proof, robustly incentive-compatible, anonymous mechanisms differ from traditional voting mechanisms in that the decision to provide the public good or not is based on an assessment of expected benefits and costs conditional on numbers of adherents and opponents, rather than any (qualified) majority rule. Even so, we find it intriguing that, once coalition-proofness is imposed in addition

[4]The introduction of the system of bets also has the effect of shifting the expected payoff that a person with valuation 0 receives from public-good provision from $\frac{1}{6}(-4) = -\frac{2}{3}$ to $\frac{1}{6}(-10) + \frac{5}{6}2 = 0$. For a person with valuation 3, the expected payoff is shifted from $-\frac{3}{4}$ to $\frac{1}{4}$, for a person with valuation 10, from 3 to 1. By the type-contingent system of bets, the people who value the public good are made to "share" the benefits with the result that the other people's expected payoffs from the system become nonnegative.

to robust incentive compatibility and anonymity, the mechanisms that we are concerned with involve numbers of votes, for and against, rather than any attempt to measure willingness to pay. Implementation can be done by a show of hands, rather than any more complicated procedure.

The remainder is organized as follows. Section 2 contains the description of the environment and the characterization of robust public good mechanisms in a large economy. In Section 3, we define the notion of a collective manipulation mechanism and of a coalition-proof rule for public-good provision. Section 4 characterizes the provision rules that are both robust and coalition-proof. Finally, in Section 6, we solve for the optimal provision rule for public goods. All proofs are in the Appendix.

# 2 Robust Implementation in a Large Economy

## 2.1 Payoffs and Social Choice Functions

We consider an economy with a continuum of agents of measure 1. There is one private good and one public good. The public good comes as a single indivisible unit. Its installation requires aggregate resources (per capita) equal to $k$ units of the private good.

Given a public-good provision level $Q \in \{0, 1\}$ the utility of any agent $i$ is given as $v_i Q - p_i$, where $v_i$ is the agent's valuation of the public good and $p_i$ is his contribution to the cost of public-good provision. The valuation $v_i$ belongs to a set $V$ of possible valuations, which is independent of $i$.

A social choice function determines under what conditions the public good is to be provided and what contributions are to be made by the different individuals. Following Guesnerie (1995), we impose an anonymity requirement by which the level of public-good provision, as well as the payments of individuals with a given valuation $v$ are unchanged under any permutation of individual characteristics that leaves the cross-section distribution of preferences unaffected. Thus, an anonymous social function determines how public-good provision levels and payment rules depend on the cross-section distribution of preferences. We refer to the latter as the state of the economy. Formally, the state of the economy is an element $s$ of the set $\mathcal{M}(V)$ of probability measures on $V$. An anonymous social choice function is a pair $F = (Q_F, p_F)$ of functions $Q_F : s \mapsto Q_F(s)$ and $p_F : (s, v) \mapsto p_F(s, v)$ such that, for any state of the economy $s$, $Q_F(s) \in \{0, 1\}$ is the level of public-good provision in the state $s$, and $p_F(s, \cdot)$ is a function indicating how, in the state $s$, and agent's payment depends on the agent's valuation.

For any $s \in \mathcal{M}(V)$, the payment rule $p_F(s, \cdot)$ is taken to be integrable with respect to $s$. The integral $\int p_F(s, v)ds(v)$ corresponds to the aggregate revenue that is collected in the state $s$. We say that the anonymous social choice function $F = (Q_F, p_F)$ yields feasible outcomes if and only if, in any state of the economy, the aggregate revenue is sufficient to cover the public-good provision cost $kQ_F(s)$, i.e., if and only if the inequality

$$\int p_F(s, v)ds(v) \geq kQ_F(s) \tag{1}$$

is satisfied for all $s \in \mathcal{M}(V)$.

## 2.2 Types and Beliefs

As usual, we model information by means of an abstract type space. Let $(T, \mathcal{T})$ be a measurable space, $\tau$ a measurable map from $T$ into $V$, and $\beta$ a measurable map from $T$ into the space $\mathcal{M}(\mathcal{M}(T))$ of probability distributions over measures on $T$. We interpret $t_i \in T$ as the abstract "type" of agent $i$, $v_i = \tau(t_i)$ as the *payoff type*, i.e., the public-good valuation of agent $i$ and $\beta(t_i)$ as the "belief type" of agent $i$.

The belief type $\beta(t_i)$ indicates the agent's beliefs about the other agents. We specify these beliefs in terms of cross-section distribution of types in the economy. Thus, $\beta(t_i)$ is a probability measure on the space $\mathcal{M}(T)$ of these cross-section distributions. For any event $X \subset \mathcal{M}(T)$, $\beta(X \mid t_i)$ is the probability that agent $i$ assigns to the event $X$. A typical element of $\mathcal{M}(T)$ will be denoted by $\delta$.

We refer to the map $\beta : T \to \mathcal{M}(\mathcal{M}(T))$ as the *belief system* of the economy. The belief system $\beta$ is called a *common-prior belief system* if there is an underlying probability space so that, given the probability distribution on this space, the belief of any one agent can be identified with a regular conditional probability distribution given his information. We think of the cross-section distribution of types $\delta$ as the realization of a random variable $\tilde{\delta}$ and of the type $t_i$ of any agent $i$ as the realization of a random variable $\tilde{t}_i$, both defined on some underlying probability space. We also think of the belief $\beta(t_i) \in \mathcal{M}(\mathcal{M}(T))$ as a conditional distribution for $\tilde{\delta}$, given the event $\tilde{t}_i = t_i$. For consistency with the notion that $\delta$ is the cross-section distribution of types, we also think of $\delta$ as being a conditional distribution for $\tilde{t}_i$ given the event $\tilde{\delta} = \delta$. Formally, the belief system $\beta$ is a common-prior belief system if and only if it is compatible with such a construction, i.e., if and only if there is a measure $P \in \mathcal{M}(T \times \mathcal{M}(T))$, with marginal distributions denoted by $P_1$ and $P_2$, respectively, such that

$$P(B_t \times B_\delta) = \int_{B_T} \beta(B_\delta \mid t) dP_1(t)$$

and

$$P(B_t \times B_\delta) = \int_{B_\delta} \delta(B_T) dP_2(\delta)$$

for all measurable $B_t \subset T$ and $B_\delta \subset \mathcal{M}(T)$.[5] Throughout the paper, we assume that $\beta$ is a common-prior belief system.

We also assume that the measures $\beta(t)$, $t \in T$, are mutually absolutely continuous, i.e., that they all have the same null sets and that these null sets are the same as the null sets of the marginal distribution $P_2$ of the common prior $P$ on the space $\mathcal{M}(T)$.[6] We refer to this property by saying that the belief system is *moderately uninformative*. If the belief system is moderately

[5] We are not saying anything about the underlying stochastic structure. The simplest specification would treat the cross-section distribution of types as a random variable $\tilde{\delta}$ on some underlying probability space and then postulate that, given $\tilde{\delta} = \delta$ the different agents' types all have the conditional probability distribution $\delta$, a conditional law of large numbers holding across agents. Spelling out the underlying stochastic structure would require us to extend the formulation of Sun (2006) so as to allow for conditional, as opposed to overall independence. See also fn. 10 below.

[6] If the beliefs $\beta(t)$, $t \in T$, are mutually absolutely continuous, one can actually show that there is at most one common prior with which the belief system is consistent.

uninformative, there is no realization $t$ of the random variable $\tilde{t}_i$ such that the observation of the event $\tilde{t}_i = t$ would permit agent $i$ to rule out any event that has positive probability under the prior $P$.

### 2.3 Robust Implementability

Information about types is assumed to be private. A social choice function is interim implementable on a given type space if, for this type space, there exists a mechanism, specifying a message set for each agent and a function from message profiles to allocations, and there exists an equilibrium of the strategic game induced by the mechanism such that the equilibrium outcome is equal to $F(\delta \circ \tau^{-1})$, as stipulated by $F$ for the payoff type distribution $\delta \circ \tau^{-1}$. An anonymous social choice function $F$ is said to be *robustly implementable* if, for every $(T, \mathcal{T})$, and $\tau : T \to V$, there exists an anonymous mechanism $f_R$ and an equilibrium of the game induced by $f_R$ that implement $F$ on the type space $[(T, \mathcal{T}), \tau, \beta]$, for every moderately uninformative common-prior belief system $\beta$.

Our notion of robustness is slightly stronger than that of Bergemann and Morris (2005). Like Bergemann and Morris, we require implementability on every type space, but, following Ledyard (1978), we go further than they do and require that the mechanism that is used for implementation should be the same regardless of what the belief system is. In contrast, Bergemann and Morris allow the mechanism to depend on $\beta$. The difference between their notion of robustness and ours (or Ledyard's) is irrelevant if one is only concerned with individual incentive compatibility. It will, however, make a difference when we add the requirement of coalition-proofness.[7] For individual incentive compatibility, we obtain:

**Proposition 1** *An anonymous social choice function $F = (Q_F, p_F)$ is robustly implementable if and only if it satisfies the following ex post incentive compatibility constraints: For all $v$ and and $v'$ in $V$ and all $s \in \mathcal{M}(V)$,*

$$vQ_F(s) - p_F(v, s) \geq vQ_F(s) - p_F(v', s) \ . \tag{2}$$

*Implementation can be achieved by direct mechanisms with truthtelling strategies.*

Proposition 1 adapts a result due to Bergemann and Morris (2005) to the given setup: Robust implementability is equivalent to *ex post incentive compatibility*, i.e., once $s$ has become known, no individual regrets having revealed his type to the mechanism. By inspection of (2), in our setting, *ex post* implementability is equivalent to the requirement that $p_F(v, s) = p_F(v', s)$ for all $v, v'$ and $s$. If the payment of some agent was, for some $s$, smaller than the payment of some other agent, the latter would like to imitate the agent with the small payment. This would contradict ex post implementability. This observation yields the following corollary to Proposition 1.

[7] See the discussion following Proposition 4 below.

**Corollary 1** *An anonymous social choice function $F = (Q_F, p_F)$ is robustly implementable if and only if payments are independent of individual payoff types, i.e., there is a function $\bar{p}_F : \mathcal{M}(V) \to \mathbb{R}$ such that $p_F$ takes the form $p_F(v, s) = \bar{p}_F(s)$ for all $v \in \Theta$ and all $s \in \mathcal{M}(V)$.*

Given Corollary 1, we will represent a robustly implementable social choice function in the following as a pair of functions $(Q_F, \bar{p}_F)$, where $\bar{p}_F(s)$ is the lump-sum contribution to the cost of public-good provision if the cross-section distribution of payoff types equals $s \in \mathcal{M}(V)$.

### 2.4 Robust Implementation of First-Best Allocations

An anonymous social choice function $F = (Q_F, p_F)$ is said to yield first-best outcomes if, for all $s \in \mathcal{M}(V)$ the pair $(Q_F(s), p_F(s, \cdot))$ maximizes the aggregate surplus

$$\int_V (vQ_F(s) - p_F(s, v))ds(v)$$

subject to the feasibility condition (1). By standard arguments, this requires that the public good should be provided if the aggregate valuation $\int_V v ds(v)$ exceeds the cost $k$ and should not be provided if $\int_V v ds(v)$ is less than $k$. Moreover, there should be no slack in the feasibility constraint, i.e., aggregate payments should exactly cover the cost of public-good provision. Upon combining these observations with Corollary 1, we obtain:

**Proposition 2** *An anonymous social choice function $F = (Q_F, p_F)$ yields first-best outcomes and is robustly implementable if and only if*

$$Q_F(s) \begin{cases} 0, & \text{if } \bar{\mathbf{v}}(s) < k, \\ 1, & \text{if } \bar{\mathbf{v}}(s) > k, \end{cases}$$

*for all $s \in \mathcal{M}(V)$, where $\bar{\mathbf{v}}(s) := \int_V v\, ds$, and*

$$p_F(v, s) = kQ_F(s)$$

*for all $s \in \mathcal{M}(V)$ and all $v \in V$.*

Proposition 2 provides a general possibility result for robust first-best implementation in a large economy. People are asked for their valuations. The public good is provided if and only if the reported average per-capita valuation exceeds $k$. Required contributions are set so that the cost of public-good provision are equally shared; this ensures feasibility (budget balance), as well as robust implementability. Because people never see themselves as having any influence on public-good provision and because people's payments do not depend on their types, each individual is indifferent as to what message he or she sends. Given this indifference, one may as well tell the truth.

Because people are indifferent about the messages they send, the game induced by the direct mechanism on any type space has many equilibria. One may therefore feel uneasy about our relying on implementation by truthtelling strategies. Would people not prefer to report an element of $\operatorname{argmin}_{t \in T} v(t)$ if $\bar{\mathbf{v}}(s) < k$ and an element of $\operatorname{argmax}_{t \in T} v(t)$ if $\bar{\mathbf{v}}(s) > k$, exaggerating

their dislike if they do not want the public good to be provided and exaggerating their enthusiasm if they do want the public good to be provided? Truthtelling actually is weakly dominated by this exaggeration strategy: In the probability zero event that the agent might be pivotal after all, exaggeration might shift the public outcome in the preferred direction when truthtelling would not. In all other events, the choice between the two strategies does not matter.

We are not bothered by this objection. The fact that people are indifferent about the messages they send is an artefact of the continuum model of a large economy. So is the fact that truthtelling is weakly dominated by an exaggeration strategy. The continuum economy is an idealization of large finite economies. For large finite economies, we know that Clarke-Groves mechanisms can be used for first-best implementation in dominant strategies; the social choice functions that are implemented by these mechanisms approximate the social choice function in Proposition 2. However, in the transition to the continuum model, the dominance property is lost. This should be seen as an example of nonrobustness of weak dominance, rather than an argument against the reliance on truthtelling equilibria in the continuum model.

Robust implementation of first-best allocations is *not* compatible with the imposition of interim participation constraints. Under equal cost sharing, anybody with a payoff type below $k$ would wish to veto the the social choice function if he could: If the public good is provided, his payoff is negative because he has to pay more than the public good is worth to him; if the public good is not provided, his payoff is zero. On average, therefore, he loses from this regime.[8]

However, in this paper, we are not concerned with participation constraints. Participation constraints matter *only* if one adheres to the contractarian view of government and the state that underlay Lindahl's (1919) original treatment of public goods. If one has no qualms about the state's using its power of coercion, Proposition 2 suggests that the implementation of first-best allocations in large economies faces no fundamental difficulties.

We do not share this sanguine view. In our view, Proposition 2 does *not* provide a satisfactory basis for the normative theory of public-good provision in a large economy. The requirements of robust implementation are too weak to do full justice to the information and incentive problems of public-good provision in such an economy. Therefore we now turn to a discussion and analysis of coalition-proofness as an additional restriction on social choice functions and incentive mechanisms.

# 3 Coalition-Proofness

To implement a first-best outcome, one must be able to ascertain the aggregate public-good valuation $\bar{\mathbf{v}}(s)$. The example in the introduction shows that some of the people who are providing this information may be effectively hurt by the use to which the information is put. In such a case, incentive compatibility holds only because any one person alone is unable to affect the social outcome and is therefore indifferent about the message that he or she transmits to the

[8]This observation corresponds to the findings of the literature on Bayesian mechanisms with independent private values, see, e.g., Güth and Hellwig (1986), Rob (1989), Mailath and Postlewaite (1990). The relation between Bayesian implementation with independent private values and robust implementation with possibly correlated values is studied in Bierbrauer and Hellwig (2009).

mechanism implementing the social choice function.

However, people with similar valuations have similar interests. Collectively, they might upset the functioning of the mechanism. Therefore, they would seem to have an incentive to form a coalition in order to collectively manipulate the social outcome. We do not want to leave room for such manipulations. In addition to individual incentive compatibility, we therefore impose a requirement of coalition-proofness.

Like Laffont and Martimort (1997, 2000), we treat coalition formation as a mechanism design problem of its own that is subject to incentive compatibility and participation constraints. However, whereas Laffont and Martimort only pay attention to the grand coalition of all agents, we will focus on coalitions of subsets of agents; moreover, we allow for coalition membership to depend on agents' types.

Let $[(T, \mathcal{T}), \tau, \beta]$ be a given type space, and let $f_R = (Q_{f_R}, p_{f_R})$ be an anonymous mechanism. We consider the possibility that this mechanism is manipulated by a coalition of people with specified types, who collectively deviate from truth-telling. We think of this coalition as being operated by a coalition manager who announces a collective manipulation mechanism and asks people to join in order to manipulate messages to the overall mechanism. Conditional on a profile of messages that he receives from coalition members, the coalition manager will choose a profile of lies that coalition members should transmit to the overall mechanism.

We consider the following structure of events, timing and information:

- First, an overall mechanism $f_R$ is announced.
- Then, a coalition organizer may propose a manipulation mechanism. This proposal is made public.
- If a manipulation mechanism has been proposed, individuals choose whether to subscribe to the manipulation mechanism or not. Any subscriber sends a report to the coalition organizer.
- On the basis of the reports that he has received from his subscribers, the coalition organizer provides each subscriber with a recommendation for a report that is to be submitted to the overall mechanism.
- All individuals choose their reports to the overall mechanism.
- The overall mechanism receives a profile of reports, one for each individual, and implements the corresponding allocation.

The overall mechanism will be said to be *coalition-proof*, if it is not possible to propose a manipulation mechanism that will benefit all individuals that join in.[9]

[9] In Section 5 below, we also consider a weaker notion of coalition-proofness. In this concept, collective manipulations are constrainted by the restriction the requirement that they must not provide room for additional collective deviations by subcoalitions. This corresponds to the notion of Bernheim et al. (1986) that collective manipulations themselves must be coalition-proof.

Before we spell out the details of the formalism, we illustrate our approach by the example in the introduction. The set of possible types in this example is $T = \{t^1, t^2, t^3\}$. The associated payoff types are $\tau(t^1) = 0$, $\tau(t^2) = 3$, and $\tau(t^3) = 10$. There is a common prior, which assigns probability one half to each of the two type distributions $\delta^0 = (0.6, 0.1, 0.3)$ and $\delta^1 = (0.2, 0.5, 0.3)$. With a per-capita provision cost $k = 4$, first-best efficiency requires that the public good should be provided if the cross-section distribution of types is $\delta^1$ and that it should not be provided if the cross-section distribution of types is $\delta^0$; moreover, the cost of public-good provision should be evenly shared.

In this example, consider a manipulation mechanism that tries to attract people with types $t^1$ and $t^2$. If these people report their types honestly, the coalition organizer observes whether the type distribution is $\delta^0$ or whether it is $\delta^1$. If it is $\delta^0$, he does not try to manipulate anything, but has each coalition member report his type honestly. If the type distribution is $\delta^1$, he manipulates messages to the overall mechanism to ensure that the overall mechanism perceives the type distribution as being $\delta^0$ rather than $\delta^1$. For this purpose, $\frac{4}{5}$ of the coalition members with type $t_2$ must report falsely that they have type $t_1$; all other coalition members report honestly. To avoid running afoul of our anonymity requirement, the coalition organizer can provide people with lotteries so that, if the actual cross-section distribution of types is $\delta^1$, then, each coalition member with type $t_2$ will report the type $t_1$ (falsely) with probability $\frac{4}{5}$ and the type $t_2$ (honestly) with probability $\frac{1}{5}$. Through this manipulation, the coalition organizer ensures that the cross-section distribution of reports received by the overall mechanism is $\delta^0$ so that the public good is not provided. Because, with equal cost sharing, the cost $k = 4$ of public-good provision to people with types $t_1$ and $t_2$ is more than the public good is worth to them, they all benefit from the manipulation. The first-best mechanism with equal cost sharing thus is not coalition-proof.

### 3.1 Strategies and Outcomes

Proceeding from heuristics to formal analysis, we allow a coalition organizer to specify any set $X$ and to ask people to send him messages from the set $X^e = X \cup \{\emptyset\}$. If agent $i$ sends the message $x_i = \emptyset$, this means that he does not participate in the manipulation mechanism and that he reports directly to the overall mechanism. If agent $i$ sends a message $x_i \in X$ to the coalition organizer, this means that he does participate in the manipulation mechanism. In response, the coalition organizer tells the agent what message to send to the overall mechanism. We think of this message as being generated by a lottery $\ell_i$ over the set $R$ of reports that are admissible under the overall mechanism; formally, $\ell_i \in \mathcal{M}(R)$. Imposing yet another anonymity requirement, we allow $\ell_i$ to depend on the message $x_i \in X$ that agent $i$ has sent and on the distribution $\chi \in \mathcal{M}(X^e)$ of messages that the coalition organizer has altogether received, and we write

$$\ell_i = \ell(x_i, \chi).$$

For instance, if we set $X = T$ in the above example and if $\frac{1}{5}$ of the population report $t_1$ to the manipulation mechanism, $\frac{1}{2}$ of the population report $t_2$, and $\frac{3}{10}$ of the population do not join, i.e., report $\emptyset$, then, for an agent who has reported $t_2$, the manipulation mechanism stipulates the

lottery $\ell(t_2, \chi)$ so that the agent's report to the overall mechanism will be $t_1$ with probability $\frac{4}{5}$ and $t_2$ with probability $\frac{1}{5}$.

Manipulating reports to the overall mechanism is the only thing manipulation mechanisms do. We neglect the possibility that the manipulation mechanism might involve side payments to facilitate coalition formation. In our model of a large economy, this involves no loss of generality. The reason is that any agent will join the coalition only if this involves no cost relative to not joining and free-riding on others forming the coalition. The expected value of the side payment that any agent who joins makes to the coalition organizer must therefore be nonpositive. Because the coalition organizer himself does not want to lose money, expected values of side payments would have to be zero anyway.

As for overall mechanisms, we do not restrict the analysis to direct mechanisms, i.e., we do not assume that the report set $X$ coincides with the type set $T$. A version of the Revelation Principle showing that there is no loss of generality in restricting the analysis to incentive-compatible direct mechanisms will be established in Subsection 3.4 below. The principle holds for both, overall mechanisms and manipulation mechanisms.

Given an overall mechanism $f_R$ and a manipulation mechanism $(X, \ell)$, a strategy for the game induced by $f_R$, and $(X, \ell)$ consists of (i) a function $\mu : T \to X^e$ that specifies an individual's report to the manipulation mechanism, (ii) a function $\lambda : \mu^{-1}(X) \times \mathcal{M}(R) \to R$ specifying the lottery $\lambda(t, \ell)$ that an agent of type $t$ who has joined the coalition and sent the message $\mu(t) \in X$ will actually use to determine his report to the overall mechanism when $\ell$ is the recommendation received from the manipulation mechanism, and (iii) a function $\nu : \mu^{-1}(\emptyset) \to R$ specifying the messages that those individuals who do not join the coalition send to the overall mechanism.

In describing the effects of the manipulation mechanism on the cross-section distribution of reports received by the overall mechanism, we assume a law of large numbers.[10] Given the manipulation mechanism $(X, \ell)$, the strategy $(\mu, \lambda, \nu)$ for the game induced by $f_R$, and $(X, \ell)$, the cross-section distribution of reports received by the manipulation mechanism is $\chi(\delta, \mu) = \delta \circ \mu^{-1}$ if the cross-section distribution of types equals $\delta$. The probability that an agent's report to the overall mechanism belongs to a measurable set $B \subset R$ is given as $\lambda(B|t, \ell(\mu(t), \chi(\delta, \mu))$ if the agent has type $t \in \mu^{-1}(X)$; if the agent has type $t \in \mu^{-1}(\emptyset)$, this probability is or zero or one, depending on whether $\nu(t)$belongs to $B$ or not. We assume that these expressions for probabilities of agents submitting reports in $B$ also indicate the fractions of the population submitting such reports. The distribution of messages received by the overall mechanism is then given as $g(\delta, X, \ell, \mu, \lambda, \nu)$, where for any measurable set $B \subset R$,

$$\begin{aligned} g(B|\delta, X, \ell, \mu, \lambda, \nu) &= \delta(\{t \in \mu^{-1}(\emptyset) | \nu(t) \in B\}) \\ &\quad + \int_{\mu^{-1}(X)} \lambda(B|t, \ell(\mu(t), \chi(\delta, \mu)) d\delta(t). \end{aligned}$$

To simplify the notation, we use $\pi = (X, \ell, \mu, \lambda, \nu)$ as a shorthand notation for the manipulation mechanism $(X, \ell)$ and the strategy $(\mu, \lambda, \nu)$, and we shall refer to $\pi$ simply as a manipulation.

[10] Beginning with Judd (1985) and Feldman and Gilles (1985), there is an extensive literature on the law of large numbers for large economies. Sun (2006) provides a formulation in which an assumption of essential pairwise independence yields a law of large numbers on any nonnegligible subset of agents; see also Sun and Zhang (2009) and Podczeck (forthcoming).

### 3.2 Interim Equilibrium

Given an overall mechanism $f_R$, and a manipulation $\pi = (X, \ell, \mu, \lambda, \nu)$, an agent of type $t$ who sends the message $r$ to the overall mechanism can expect to receive the payoff

$$u(\pi, t, r, \delta) := \tau(t) Q_{f_R}(g(\delta, \pi)) - p_{f_R}(r, g(\delta, \pi))$$

if the cross-section of types is $\delta$. However, the agent does not know $\delta$. If he does not join the manipulation mechanism, his expectations about $\delta$ are given by his belief type $\beta(t) \in \mathcal{M}(\mathcal{M}(T))$, his expected payoff from sending the report $r$ to the overall mechanism is

$$U_N(\pi, t, r) := \int_{\mathcal{M}(T)} u(\pi, t, r, \delta) d\beta(\delta|t).$$

If, instead, he joins the manipulation mechanism, sends a message $x \in X$ and receives a recommendation $l \in \mathcal{M}(R)$, he will update his beliefs about $\delta$, using the information that $\ell(x, \chi(\delta, \mu)) = l$. His expected payoff from sending the reports $x$ to the manipulation mechanism and using the lottery $\lambda(t, l)$ to determine his report $r$ to the overall mechanism is then equal to

$$\int_R \int_{\mathcal{M}(T)} u(\pi, t, r, \delta) db(\delta|t, x, l) d\lambda(r|t, l) = \int_{\mathcal{M}(T)} \int_R u(\pi, t, r, \delta) d\lambda(r|t, l) db(\delta|t, x, l) \,,$$

where $b(t, x, l) \in \mathcal{M}(\mathcal{M}(T))$ is the agent's updated belief given his belief $\beta(t)$ and the information that $\ell(x, \chi(\delta, \mu)) = l$. From an ex ante perspective, at the time of his deciding on whether to join the manipulation mechanism or not, his expected utility from sending the message $x$ to the manipulation mechanism and subsequently following the response strategy $\lambda$ is equal to

$$U_J(\pi, t, x, \lambda(t, \cdot)) := \int_{\mathcal{M}(T)} \int_R u(\pi, t, r, \delta) d\lambda(r|t, \ell(x, \chi(\delta, \mu))) d\beta(\delta|t).$$

The triple $(\mu, \lambda, \nu)$ is an interim equilibrium for the game induced by the overall mechanism $f_R$ and the manipulation mechanism $(X, \ell)$ if the following conditions are satisfied:

i) For any $t \in \mu^{-1}(X)$, the pair $(\mu(t), \lambda(t, \cdot))$ is a best response to the strategy $(\mu, \lambda, \nu)$, i.e.,

$$U_J(\pi, t, \mu(t), \lambda(t, \cdot)) \geq U_J(\pi, t, x, \lambda'(\cdot))$$

for all $x \in X$ and all responses $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$ to the manipulation organizer's suggestions, and

$$U_J(\pi, t, \mu(t), \lambda(t, \cdot)) \geq U_N(\pi, t, r)$$

for all $r \in R$.

ii) For any $t \in \mu^{-1}(\emptyset)$, the report $\nu(t)$ is a best response to the strategy $(\mu, \lambda, \nu)$, i.e.,

$$U_N(\pi, t, \nu(t)) \geq U_J(\pi, t, x, \lambda'(\cdot))$$

for all $x \in X$ and all responses $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$ to the manipulation organizer's suggestions, and

$$U_N(\pi, t, \nu(t)) \geq U_N(\pi, t, r),$$

for all $r \in R$.

If the payoff $u(\pi, t, r, \delta) = \tau(t)Q_{f_R}(g(\delta, \pi)) - p_{f_R}(r, g(\delta, \pi))$ is independent of the report $r$, these equilibrium conditions are vacuous. This is always the case if the payments $p_{f_R}$ under the mechanism $f_R$ are type-independent.[11]

**Remark 1** *Let $f_R = (Q_{f_R}, p_{f_R})$ be such that the payment rule takes the form $p_{f_R}(r, \rho) = \bar{p}_{f_R}(\rho)$ for all $r \in R$ and all $\rho \in \mathcal{M}(R)$. Then, for any manipulation $\pi = (X, \ell, \mu, \lambda, \nu)$, the strategy $(\mu, \lambda, \nu)$ is an interim equilibrium for the game induced by the overall mechanism $f_R$ and the manipulation mechanism $(X, \ell)$.*

### 3.3 Coalition-Proofness

At last we turn to the definition of coalition-proofness. For a given common prior type space $[(T, \mathcal{T}), \tau, \beta]$ and mechanism $f_R = (Q_{f_R}, p_{f_R})$, let $\sigma^*$ be an interim equilibrium for the game induced by $f_R$. A manipulation $\pi = (X, \ell, \mu, \lambda, \nu)$ consisting of a manipulation mechanism $(X, \ell)$ and an interim equilibrium $(\mu, \lambda, \nu)$ is said to block $\sigma^*$ if the people who join the manipulation mechanism, i.e., the people choosing $\mu(t) \in X$, are strictly better off than they would be in the equilibrium $\sigma^*$, in the absence of the manipulation mechanism. Formally, $\pi$ blocks $\sigma^*$ if

$$U_J(\pi, t, x, \lambda(t, \cdot)) > U(\sigma^*, \sigma^*(t), t) \tag{3}$$

for all $t \in \mu^{-1}(X)$, where

$$U(\sigma^*, \sigma^*(t), t) := \int_{\mathcal{M}(T)} \{\tau(t)Q_{f_R}(\delta \circ \sigma^{*-1}) - p_{f_R}(\sigma^*(t), \delta \circ \sigma^{*-1})\} d\beta(\delta \mid t)$$

How should we think about the reporting strategy $\nu$ of the people who do not join the mechanism? By definition, $\nu$ is part of an equilibrium for the reporting game that is induced by the overall mechanism $f_R$ and the manipulation mechanism $l$. However, as indicated by Remark 1, this requirement may not impose much of a constraint on the choice of $\nu$. We need an account of how the reporting strategy $\nu$ is selected from the set of strategies that form part of an interim equilibrium. Dealing with this issue involves a certain element of arbitrariness. For specificity, we assume that people who do not join a manipulation mechanism submit the same reports to the overall mechanism as they would if the manipulation mechanism wasn't there.[12]

Thus, let $\sigma^*_{\mu^{-1}(\emptyset)}$ be the restriction of $\sigma^*$ to the non-participating types. An interim equilibrium $\sigma^*$ for the mechanism $f_R$ is said to be coalition-proof if there is no manipulation $\pi := (X, \ell, \mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$, with $(\mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ an interim equilibrium for the game induced by $f_R$ and $(X, \ell)$, such that $\pi$ blocks $\sigma^*$.

[11] If we had allowed for manipulation mechanisms with side payments, the equilibrium conditions would, however, impose the restriction that, unless the coalition organizer puts in some money of his own, expected equilibrium side payments must be zero.

[12] An alternative, stronger concept of coalition-proofness would give the organizer of the manipulation mechanism the power actually to choose the reporting strategy of the non-joiners. Yet another, weaker concept of coalition-proofness would give the non-joiners the ability to coordinate on a strategy $\nu$ that represents a collective best response to $\mu$ or on a strategy $\nu$ that minimizes the coalition members' benefits from the presence of the manipulation mechanism (always subject to individual incentive compatibility). As a by-product of our analysis in Section 4, however, we will find that the choice of a selection principle for $\nu$ does not make much of a difference. See fn. 15 below.

### 3.4 Obedience and the Revelation Principle

In the remainder of this section, we consider the place of the Revelation Principle in the present setting. We first observe that there is no loss of generality in assuming that manipulation mechanisms are direct mechanisms that induce truth-telling as well as obedience by coalition joiners. Given a manipulation $\pi = (X, \ell, \mu, \lambda, \nu)$, let

$$T_\pi := \mu^{-1}(X) = T \backslash \mu^{-1}(\emptyset)$$

be the set of coalition joiners, and let $h_{T_\pi} : T \to T^e = T \cup \{\emptyset\}$ be the strategy that stipulates truthtelling, $h_{T_\pi}(t) = t$, for $t \in T_\pi$ and nonparticipation, $h_{T_\pi}(t) = \emptyset$, for $t \notin T_\pi$, i.e., for $t \in \mu^{-1}(\emptyset)$.

**Proposition 3** *Given a type space $[(T, \mathcal{T}), \tau, \beta]$, an overall mechanism $f_R = (Q_{f_R}, p_{f_R})$, and an interim equilibrium $\sigma^*$, suppose that there is a manipulation $\pi = (X, \ell, \mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ that blocks $\sigma^*$. Define $\ell^* : T_{\mu^{-1}(X)} \times \mathcal{M}(T^e_{\mu^{-1}(X)}) \to \mathcal{M}(R)$ and $\lambda^* : T \times \mathcal{M}(R) \to \mathcal{M}(R)$ so that, for any $t \in \mu^{-1}(X)$, $\chi^* \in \mathcal{M}(T^e_{\mu^{-1}(X)})$, and $l \in \mathcal{M}(R)$,*

$$\ell^*(t, \chi^*) = \lambda(t, \ell(\mu(t), \chi^* \circ \mu^{-1})) \tag{4}$$

*and*

$$\lambda^*(t, l) = l. \tag{5}$$

*Then $\sigma^*$ is also blocked by the manipulation $\pi^* = (T_\pi, \ell^*, h_{T_\pi}, \lambda^*, \sigma^*_{\mu^{-1}(\emptyset)'})$.*

To verify coalition-proofness of an interim equilibrium $\sigma^*$ for the mechanism $f_R$, it is thus sufficient to show that there is no incentive-compatible direct manipulation mechanism that is coalition-proof and blocks $\sigma^*$ when coalition joiners follow the manipulation mechanism's recommendations. Given this characterization, we can use a more concise notation for manipulations. We will simply write $\pi = (T_\pi, \ell)$, with the understanding that the manipulation mechanism is $(T, \ell)$ and the strategy is $(h_{T_\pi}, \lambda^*, \sigma^*_{T \backslash T_\pi})$.

Turning from mechanisms to social choice functions, we say that a social choice function $F$ is robustly implementable andcoalition-proof, if and only if, for every $(T, \mathcal{T})$, and $\tau$, there is an anonymous mechanism $f_R$ that implements $F$ as a coalition-proof interim equilibrium on the type space $[(T, \mathcal{T}), t, \tau, \beta]$, for every common-prior belief system $\beta$.

**Proposition 4** *A social choice function $F = (Q_F, p_F)$ is robustly implementable and coalition-proof if and only if there is an anonymous direct mechanism $f$ so that truth-telling implements $F$ and, moreover, truthtelling is a coalition-proof interim equilibrium on every type space.*

According to Proposition 4 we may without loss of generality focus on truth-telling equilibria of direct mechanisms. The requirement of robustness is necessary for this result, i.e., for a fixed type space, the revelation principle does not hold. The reason is akin to the well-known result that the implementation of a social choice function as the unique equilibrium of some mechanism

may require the use of non-direct mechanisms. To see the analogy, note that, for a given type space, our notion of coalition-proofness requires that the possibility of implementing a certain outcome by an interim equilibrium must not be endangered by the existence of a second interim equilibrium which would be preferred by a subset of types. Ordinarily, non-direct mechanism can be used to get rid of additional equilibria with undesirable outcomes that a direct mechanism might have.[13] Robustness eliminates this possibility.

On the basis of these results, we obtain the following very simple characterization of robustly implementable and coalition-proof social choice functions.

**Corollary 2** *A social choice function $F = (Q_F, \bar{p}_F)$ with type independent payments is robustly implementable and coalition-proof if and only if there is no common prior type space $[(T, \mathcal{T}), \tau, \beta]$ and no manipulation $\pi = (T_\pi, \ell)$ such that*

$$\int [\tau(t) Q_F(\hat{s}(\delta, \pi)) - \bar{p}_F(\hat{s}(\delta, \pi))|\ t]\, d\beta(\delta|t) > \int [\tau(t) Q_F(s(\delta)) - \bar{p}_F(s(\delta))|\ t]\, d\beta(\delta|t) \qquad (6)$$

*for all $t \in T_\pi$, where, for any $\delta \in \mathcal{M}(T)$, $s(\delta) = \delta \circ \tau^{-1}$ is the true cross-section distribution of valuations and $\hat{s}(\delta, \pi) := g(\delta, \pi) \circ \tau^{-1}$ is the cross-section distribution of valuations that is communicated to the overall mechanism if all people with $t \in T_\pi$ join the manipulation mechanism and follow its recommendations.*

# 4 Implications of Coalition-Proofness and Robustness

In this section we consider the implications of coalition-proofness and robustness. In essence, we show that coalition-proofness reduces to the requirement that the decision on public goods provision can condition *only* on the size of the set of people who benefit from public-good provision and the size of the set of people who oppose public-good provision. Intensities of preferences cannot play a role. Moreover, the decision on public-good provision must be monotonic in the sense that it is not possible to have it provided when the set of net beneficiaries is smaller than in some other instance where the public good is not provided. The welfare implications of our analysis will be discussed further below. As indicated by the example in the introduction, the requirements of coalition-proofness and robustness may preclude the achievement of first-best outcomes.

**Theorem 1** *If a social choice function $F = (Q_F, p_F)$ is robustly implementable and coalition-proof, then it satisfies the following two properties:*

[13] See Bassetto and Phelan (2008), Jackson (2001), or Moore (1992). The reason that non-direct mechanisms can be superior is that they may include off-the equilibrium actions and payoffs which are calibrated so as to make deviations from some hypothetical equilibrium unattractive. However, robustness requires that an individual has the same best responses under all circumstances. Hence, it becomes impossible to design actions that are attractive only in some-off-the-equilibrium-circumstances but not in some other equilibrium-circumstances; see Bierbrauer (2009) for further details.

1. *There exist numbers* $p_F^0$ *and* $p_F^1$ *so that, for all* $v \in V$ *and all* $s \in \mathcal{M}(V)$,

$$p_F(v,s) = \begin{cases} p_F^0, & \text{if} \quad Q_F(s) = 0\ , \\ p_F^1, & \text{if} \quad Q_F(s) = 1\ . \end{cases} \tag{7}$$

2. *If*

$$V_1(p_F^1 - p_F^0) := \{v \in V \mid v > p_F^1 - p_F^0\} \ \text{ and } \ V_0(p_F^1 - p_F^0) := \{v \in V \mid v < p_F^1 - p_F^0\}$$

*are the sets of payoff types of net gainers and net losers from public-good provision, then, for all* $s$ *and* $s'$ *in* $\mathcal{M}(V)$,

$$\begin{array}{l} s(V_1(p_F^1 - p_F^0)) \geq s'(V_1(p_F^1 - p_F^0)) \ \text{ and } \ s(V_0(p_F^1 - p_F^0)) \leq s'(V_0(p_F^1 - p_F^0)) \\ \text{imply} \quad Q_F(s) \geq Q_F(s'). \end{array}$$

*In particular,*

$$\begin{array}{l} s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0)) \ \text{ and } \ s(V_0(p_F^1 - p_F^0)) = s'(V_{01}(p_F^1 - p_F^0)) \\ \text{imply} \quad Q_F(s) = Q_F(s'). \end{array}$$

Theorem 1 suggests a remarkable link between mechanism design and voting. Economists have long been critical of the prominent role of voting in political systems, arguing that the neglect of preference intensities in voting was a major source of distortions. According to Theorem 1, neglect of preference intensities is a necessary implication of robust implementability and coalition-proofness. Any social welfare function that has these properties can therefore be implemented by a voting mechanism. Formally, we define a voting mechanism as a mechanism with the following properties:

- The message set $R_V$ is a binary set, $R_V = \{\text{no}, \text{yes}\}$; people can only vote for or against a given proposal.
- People vote on whether the public good is to be provided, in which case each individual has to make a payment $p_V^1 \geq k$. Otherwise, if the public good is not provided, each individual has to make a payment $p_V^0 \geq 0$.
- There is a threshold $m_V \in [0,1]$ so that the public good is provided and people pay $p_V^1$ if share of the population voting *yes* is at least $m_V$; the public good is not provided, and people pay $p_V^0$ if the share of the population voting *yes* is is less than $m_V$.

With this definition of a voting mechanism, we obtain:

**Corollary 3** *If a social choice function* $F = (Q_F, p_F)$ *is robustly implementable and coalition-proof, then it can be implemented by a voting mechanism.*

In the remainder of this section, we explain the logic behind Theorem 1. We focus on the implications of coalition-proofness with respect to three special coalitions, the coalition of all participants, the coalition of people with valuations in the set $V_1(p_F^1 - p_F^0)$, who want the public good to be provided, and the coalition of people with valuations in the set

$$V_0(p_F^1 - p_F^0) := \{v \in V \mid v < p_F^1 - p_F^0\},$$

who do not want the public good to be provided.

We proceed through a sequence of lemmas. The first lemma eliminates the possibility that payments might differ across states that involve the same public-good provision level. If payments were high in some states and low in others when both involve the same level of public-good provision, then the grand coalition coalition of all participants together could implement a manipulation that induces the "cheap" outcome even when the actual state would call for the "expensive" outcome. Formally, we obtain:

**Lemma 1** *If a social choice function $F = (Q_F, p_F)$ is robustly implementable and coalition-proof, then there exist numbers $p_F^0$ and $p_F^1$ such that the payment rule satisfies (7) for all $v \in V$ and all $s \in \mathcal{M}(V)$.*

Given this lemma, we restrict our attention to social choice functions with type independent payments that are the same in all states in which the public good is provided and the same in all states in which the public good is not provided. We find it convenient to denote such a social choice function as $F = (Q_F, p_F^0, p_F^1)$ rather than $F = (Q_F, p_F)$. The social choice function is completely characterized by the public-good provision rule and the values $p_F^0, p_F^1$ of the payment rule.

If the social choice function $F = (Q_F, p_F^0, p_F^1)$ is implemented, an individual with valuation $v$ obtains the payoff $v - p_F^1$ in any state $s$ in which the public good is provided and the payoff $-p_F^0$ in any state in which it is not provided. The individual benefits from public-good provision if $v > p_F^1 - p_F^0$. The individual is harmed by public-good provision if $v < p_F^1 - p_F^0$. Our analysis of coalition-proofness of mechanisms on a given type space $[(T, \mathcal{T}), \tau, \beta]$ will focus on the set

$$C_0 := \{t \mid \tau(t) < p_F^1 - p_F^0\} \tag{8}$$

of all types that are harmed by public-good provision and the set

$$C_1 := \{t \mid \tau(t) > p_F^1 - p_F^0\} \tag{9}$$

of all types that benefit from public-good provision. These are two canonical sets of types with common interests. We denote by $\Pi_0$ the set of all manipulations of the form $\pi = (T_{C_0}, \ell)$ and by $\Pi_1$ the set of all manipulations of the form $\pi = (T_{C_0}, \ell)$.

The analysis of coalition-proofness with respect to these coalitions is encumbered by the fact that, although they have the same interests, yet, different types in $C_0$ or different types in $C_1$ may have different beliefs. If the belief system is moderately uninformative, however, there is a limit to the differences in beliefs that may arise. This is the key to the following lemmas.

**Lemma 2** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying ($\mathcal{Y}$) and for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with a moderately uninformative belief system, the following statements are equivalent:*

*($a_0$) There is no manipulation $\pi \in \Pi_0$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$.*

*($b_0$) For every manipulation $\pi \in \Pi_0$,*

$$\int Q_F(s(\delta))dP_2(\delta) \leq \int Q_F(\hat{s}(\delta, \pi))dP_2(\delta) \; . \tag{10}$$

**Lemma 3** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying ($\mathcal{Y}$) and for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with a moderately uninformative belief system, the following statements are equivalent:*

*($a_1$) There is no manipulation $\pi \in \Pi_1$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$.*

*($b_1$) For every manipulation $\pi \in \Pi_1$,*

$$\int Q_F(s(\delta))dP_2(\delta) \geq \int Q_F(\hat{s}(\delta, \pi))dP_2(\delta) \; . \tag{11}$$

Lemmas 2 and 3 translate the requirements of coalition-proofness into conditions on the provision rule $Q_F$. Condition (10) implies that a coalition organizer who observes the message distribution $\chi(\delta, h_{C_0}) = \delta \circ h_{C_0}^{-1}$ and shapes his recommendations to coalition joiners accordingly finds that the probability of public-good provision is minimized if all members of his coalition report their types truthfully to the overall mechanism. The "probability of public-good provision" here is a conditional probability given the information that is available to the coalition organizer. The coalition organizer observes the size of his coalition, as well as the cross-section distribution of types in his coalition. Because his information is finer than that of his coalition members, his probability assessments are likely to differ from theirs. Given the assumption, however, that all agents have the same prior, they all would have the same conditional probability assessments if they had the coalition organizer's information. Therefore they all agree that, if, conditional on his information, the coalition organizer finds a way to lower the probability of public-good provision, this would not be a bad thing. Indeed, if such an event has positive probability, they consider it to be a good thing; moreover, with a common prior and moderately uninformative beliefs, they are agreed as to which events have positive probability.

Thus, if (10) was violated, there would exist a set of cross-section distributions of types, $D \subset \mathcal{M}(T)$, which has positive prior probability (according to the marginal distribution $P_2$ of the common prior $P$), so that, whenever $\delta \in D$, a coalition organizer who observes $\chi(\delta, h_{C_0})$ is able to manipulate the announcements of individuals with types in $C_0$ in such a way, that, conditional on the information available to him, the probability of public-good provision would be reduced. Because, with a moderately uninformative belief system, all individuals with types

in $C_0$ assign positive probability to the set $D$, they all find it strictly advantageous to participate, i.e., this manipulation blocks the truthful equilibrium.

Similarly, condition (11) implies that a coalition organizer who observes the message distribution $\chi(\delta, h_{C_1})$ and shapes his recommendations to coalition joiners accordingly finds that the probability of public-good provision is maximized if all members of his coalition report their types truthfully to the overall mechanism. If this condition was violated, there would be a manipulation $(C_1, \ell)$ that would raise the probability of public-good provision and would be welcomed by all participants with types in $C_1$.

Upon combining these two lemmas and considering what manipulations $\pi_0 \in \Pi_0$ or $\pi_1 \in \Pi_1$ are available to the two relevant coalitions, $C_0$ and $C_1$, we obtain the following further characterization.

**Lemma 4** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying ($\Upsilon$), the following statements are equivalent:*

**(a*)** *If $[(T, \mathcal{T}), \tau, \beta]$ is any common prior type space with a moderately uninformative belief system, there are no manipulations $\pi_0 \in \Pi_0$ or $\pi_1 \in \Pi_1$ that block the truthful equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$.*

**(b*)** *For all $s$ and $s'$ in $\mathcal{M}(V)$, $s(V_0(p_F^1 - p_F^0)) \geq s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) \leq s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) \leq Q_F(s')$. In particular, $s(V_0(p_F^1 - p_F^0)) = s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) = Q_F(s')$.*

This lemma is best understood by observing that conditions (10) and (11) can be interpreted as saying that truthtelling is a Nash equilibrium of a strictly competitive game. In this game, there are two players, 0 and 1. At stage 0, both players simultaneously and independently choose manipulations $\pi_0 \in \Pi_0$ and $\pi_1 \in \Pi_1$. At stage 1, nature draws $\delta \in \mathcal{M}(T)$ according to the distribution $P_2$. Player $i$ observes the distribution $\chi(\delta, h_{C_i}) = \delta \circ h_{C_i}^{-1}$; he thereby learns the population share of the set $C_i$ and the cross-section distribution of types in $C_i$. Given this information, he transmits reports $\ell_i(t, \chi(\delta, h_{C_i})), t \in C_i$, to the overall mechanism. Given the distribution $g(g(\delta, \pi_0), \pi_1)$ of reports that it has actually received, the overall mechanism implements the public-good provision level $Q_F(g(g(\delta, \pi_0), \pi_1) \circ \tau^{-1}) = Q_F(\hat{s}(\hat{s}(\delta, \pi_0), \pi_1))$. This results in payoffs $-Q_F(\hat{s}(\hat{s}(\delta, \pi_0), \pi_1))$ for player 0 and $+Q_F(\hat{s}(\hat{s}(\delta, \pi_0), \pi_1))$ for player 1. This is a zero-sum game. Player 0 seeks to minimize the level of public goods provision; player 1 seeks to maximize it.

In terms of this game, condition (11) asserts that, if player 0 pursues a truthtelling strategy, i.e., if $\pi_0$ is such that $g(\delta, \pi_0) = \delta$ for all $\delta$, then truthtelling is a best response for player 1, i.e., player 1 is willing to choose $\pi_1$ so that $g(\delta, \pi_1) = \delta$ for all $\delta$, and, conversely, if player 1 is choosing $\pi_1$ so that $g(\delta, \pi_1) = \delta$ for all $\delta$, then condition (10) asserts that player 0 is willing to choose $\pi_1$ so that $g(\delta, \pi_1) = \delta$ for all $\delta$.

Given that truthtelling is a Nash equilibrium of this strictly competitive game, the saddle-point theorem for such games (see, e.g., Osborne and Rubinstein (1994)) can be used to obtain

additional insights about the function $Q_F$. For an arbitrary common prior $P$, the application of the saddle-point theorem is somewhat encumbered by the fact that neither player knows the other player's information. However, if the belief system is such that $\beta(\{\delta\}|t) = 1$ for some $\delta$ and all $t$, we must have $P_2(\{\delta\}) = 1$, i.e., the common prior assigns all probability mass to the event that the cross-section distribution of types is $\delta$. In this case, conditions (10) and (11) imply that truthtelling is a Nash equilibrium of a strictly competitive game of complete information. The saddle-point theorem implies that, for each player, truthtelling is in fact a maxminimizer, not merely a best response. From this property, one immediately derives the second statement in (b*), namely that $Q_F(\cdot)$ must be constant over any set of states that differ only with respect to the cross-section distributions of types in $C_0$ and in $C_1$, but not in the population shares of $C_0$ or $C_1$: All these states give rise to the same set of feasible lies for the two coalition organizers. Thus, if $Q_F(s') = 1$ for some $s'$, then the outcome $Q = 1$ must also result for all $s$ for which the coalition $C_1$ can mimic its own truth-telling behavior in situation $s'$. Otherwise, the organizer of $C_1$ would deviate from truth-telling so as to make sure that the outcome is $Q = 1$.

Given the second statement in (b*), the first statement follows by observing that, if we had $s(V_0(p_F^1 - p_F^0)) \geq s'(V_0(p_F^1 - p_F^0))$, $s(V_1(p_F^1 - p_F^0)) \leq s'(V_1(p_F^1 - p_F^0))$ and $Q_F(s) > Q_F(s')$ for some $s$ and $s'$, then either player 0 could benefit by deviating from truthtelling in state state $s$ or player 1 could benefit by deviating from truthtelling in state $s'$.[14]

To conclude this section, we note that, in terms of the social choice function $F = (Q_F, p_F)$, statement (b*) in Lemma 4 is the same as the second statement in Theorem 1. The theorem thus follows from Lemmas 4 and 1.

## 5 Weak Coalition-Proofness and an Equivalence Theorem

Theorem 1 only gives necessary conditions for coalition-proofness. To be sure, Lemma 4 gives necessary and sufficient conditions for immunity against all manipulations by coalitions of types in $C_0$ or $C_1$. One easily sees that these conditions are in fact necessary and sufficient for immunity against all manipulations by coalitions of types in a set $T' \subset C_0$ or in a set $T' \subset C_1$.

However, there remains the possibility that, even with the social choice function characterized in Theorem 1, in some common prior type space, the truthful equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ might be blocked by a manipulation by a coalition of types in $C_0$ and of types in $C_1$. At first sight, this may seem quite unlikely because these types have conflicting interests. However, the blocking manipulation might involve a "trade" of the following sort: in some instances where truthtelling would induce the outcome $Q = 0$, the manipulation would induce the outcome $Q = 1$, and in some other instances where truthtelling would induce the outcome $Q = 1$, the manipulation would induce the outcome $Q = 0$. Such a manipulation is attractive to types in $C_1$ if they consider the first set of instance more likely than the second and to types in $C_0$ if they consider the second set of instances more likely than

[14] Which of the two it is, depends on the population share of the set $\{p_F^1 - p_F^0\}$ of payoff types who are indifferent whether the public good is provided or not. If $s(\{p_F^1 - p_F^0\}) \leq s'(\{p_F^1 - p_F^0\})$, the deviation that dominates truthtelling is available to player 0, who can give reports in $\tau^{-1}(\{p_F^1 - p_F^0\})$ and in $\tau^{-1}(V_1(p_F^1 - p_F^0))$ so that reported population shares of the sets of types are the same as in the state $s'$.

the first.[15] Thus, it seems that, for a particular belief system, such a manipulation can block the truthful equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$.

Does this make sense? If the truthful equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ is blocked by a coalition of types in $C_0$ and of types in $C_1$, the types in $C_0$ are happy about the instances in which the manipulation induces the outcome $Q = 0$, rather than $Q = 1$, but they are unhappy about the instances in which the manipulation induces the outcome $Q = 1$, rather than $Q = 0$. Taking the behavior of all other individuals as given, they would therefore seem to have an incentive to form a *subcoalition* that sabotages the blocking manipulation whenever the manipulation would replace the outcome $Q = 0$ by the outcome $Q = 1$. In such instances, the members of the subcoalition might simply report truthfully to the overall mechanism, rather than follow the recommendations of the blocking manipulation. In instances where the blocking manipulation replaces the outcome $Q = 1$ by the outcome $Q = 0$, the sabotaging subcoalition would be inactive so that, in these instances, the types in $C_0$ would still benefit from the blocking manipulation. Such a subcoalition would be attractive to types in $C_0$, i.e., the blocking manipulation itself would fail to be coalition-proof.

In the following, we formalize this idea by introducing a concept of *weak* coalition-proofness. An interim equilibrium for a given mechanism will be said to be weakly coalition-proof if there is no manipulation that blocks it and is itself subcoalition-proof, i.e., does not provoke a further manipulation by a subcoalition.[16]

### Weakly Coalition-Proof Equilibrium

To simplify the exposition, we focus on direct mechanisms and on truth-telling equilibria, at the level manipulation mechanisms and submanipulations, as well as the overall mechanism.[17] Given a social choice function $F = (Q_F, p_F)$, let $f$ be an incentive-compatible direct mechanism that implements $F$ on a type space $[(T, \mathcal{T}), \tau, \beta]$, and let $\pi = (T_\pi, \ell)$ be a manipulation that blocks the truthtelling equilibrium for the game induced by $f$.

[15] If we limited attention to common prior type spaces satisfying a monotone likelihood ratio property, so that individuals with a higher payoff type do not assign less probability to states having more individuals who benefit from public goods provision, all the result established so far would remain valid, and, in addition, joint manipulations by types in $C_0$ and $C_1$ could never occur. However, we seek to avoid any assumption that restricts the class of admissible common priors.

[16] Our notion of weak coalition-proofness is inspired by the "coalition-proof Nash equilibrium" as defined by Bernheim et al. (1986). However, we do not model a possibly infinite sequence of a successive formation of subcoalitions.

[17] A more general formulation, allowing for indirect as well as direct mechanism, can be given along the lines of Section 3. In this more general formulation, standard arguments, as in the proof of Proposition 3, can be used to show that there is no loss of generality in restricting attention to submanipulations that rely on direct mechanisms. There is also no loss of generality in restricting the analysis of subcoalition-proof manipulations to direct mechanisms. The requirement of subcoalition-proofness just adds one further "equilibrium condition" to the game induced by an overall mechanism and a manipulation mechanism on a given type space. Given that manipulation mechanisms are mechanisms with type independent payments that are equal to zero, the arguments in the proofs of Proposition 3 and 4 are easily adapted to show that if all equilibrium conditions hold with some non-direct manipulation mechanism, then there is an "equivalent" direct manipulation mechanism with the same property. Finally, the arguments in the proof of Proposition 4 can also be adapted to show that there is no loss of generality in restricting the analysis of weakly coalition-proof mechanisms to direct mechanisms.

Given $f$ and $\pi$, we consider a submanipulation $\pi^s$ that works as follows. When people with types in $T_\pi$ receive their recommendations from the manipulation mechanism, a subset of these people with types in a set $T_{\pi^s} \subset T_\pi$ inform a submanipulation mechanism designer about their types, the messages that they sent to the first manipulation mechanism, and the recommendations that they received in return. The submanipulation mechanism then provides these people with a new recommendation for the message that they should send to the overall mechanism. The submanipulation mechanism is characterized by the message space $C^e := C \cup \{\emptyset\}$ and the recommendation function $\ell^s : C \times \mathcal{M}(C^e) \to \mathcal{M}(T)$. Here $C := T \times T \times \mathcal{M}(R)$ is the set of triples consisting of people's types, their messages to the first manipulation mechanism, and the recommendations that they received in return, and, as before, $\emptyset$ is the message of a person who is not joining. The function $\ell^s : C \times \mathcal{M}(C^e) \to \mathcal{M}(T)$ shows how the recommendation that a person receives from the submanipulation mechanism depend on his message and on the distribution of messages that the submanipulation mechanism has received.

We say that the manipulation $\pi$ is subcoalition-proof if and only if there is no submanipulation $\pi^s$ such that (i) the behavior specified by $\pi^s$ is, an interim equilibrium for the game induced by $f$, the manipulation mechanism $(T^e, \ell)$, and the submanipulation mechanism $(C^e, \ell^s)$, and (ii) in this equilibrium, all types in $T_{\pi^s}$ are strictly better off than they are under the manipulation $\pi$. As before, we work on the presumption that types outside $T_{\pi^s}$ stick to the behavior specified by $\pi$. We use the shorthand notation $\pi^s = (T_{\pi^s}, \ell^s)$ for this manipulation.

The truthtelling equilibrium for the overall mechanism $f$ is said to be weakly coalition-proof if and only if there is no subcoalition-proof manipulation $\pi = (T_\pi, \ell)$ that blocks this equilibrium. The social choice function $F$ is robustly implementable and weakly coalition-proof if and only if, for every $(T, \mathcal{T})$, and $\tau$, there is an anonymous mechanism $f_R$ that implements $F$ as a weakly coalition-proof interim equilibrium on the type space $[(T, \mathcal{T}), \tau, \beta]$, for every common prior belief system $\beta$.

**An Equivalence Theorem**

For technical reasons, we restrict ourselves to *regular* social choice functions. A social choice function $F = (Q_F, p_F)$ is said to be *regular* if, for any outcome $Q \in \{0, 1\}$, any type distribution $\delta \in \mathcal{M}(T)$, and any set $T' \subset T$, the following is true: if a coalition of types in $T'$ has a manipulation $\pi = (T', \ell)$ that induces the outcome $Q_F(\hat{s}(\delta, \pi)) = Q$ when the type distribution is $\delta$, then it also has a manipulation $\pi' = (T', \ell')$ that achieves this outcome at minimal cost, i.e., the problem

$$\min_{\ell'} \quad p_F(\hat{s}(\delta, T', \ell')) \quad \text{subject to} \quad Q_F(\hat{s}(\delta, T', \ell')) = Q$$

has a solution. The coalition of types in $T'$ does not have an open set problem of the sort that, on the set of lies $\ell'$ that induce the outcome $Q_F(\hat{s}(\delta, T', \ell')) = Q$, the payment $p_F(\hat{s}(\delta, T', \ell'))$ has an infimum which is not also a minimum. This condition ensures that, for every $\delta$ and $T'$, there is a well-defined most attractive manipulation for types in $T'$. Without regularity, it might be the case that, for each manipulation $\pi = (T', \ell)$, there is another manipulation $\pi' = (T', \ell')$ that achieves the same outcome with lower payment requirements; in this case, no manipulation

$\pi = (T', \ell)$ would be subcoalition-proof and the concept of weak coalition-proofness would not have any bite.

Also, for ease of exposition, we limit attention to moderately uninformative common prior type spaces. This implies that we do not have to worry about manipulations and submanipulations that affect the outcome with probability zero.

**Theorem 2** *A regular social choice function $F = (Q_F, p_F)$ is robustly implementable and weakly coalition-proof if and only if it satisfies:*

1. *There exist numbers $p_F^0$ and $p_F^1$ so that, for all $v \in V$ and all $s \in \mathcal{M}(V)$, $p_F(v,s) = p_F^0$ if $Q_F(s) = 0$ and $p_F(v,s) = p_F^1$ if $Q_F(s) = 1$.*

2. *For all $s$ and $s'$ in $\mathcal{M}(V)$,*

$$s(V_1(p_F^1 - p_F^0)) \geq s'(V_1(p_F^1 - p_F^0)) \text{ and } s(V_0(p_F^1 - p_F^0)) \leq s'(V_0(p_F^1 - p_F^0)) \text{ implies } Q_F(s) \geq Q_F(s'),$$

*where $V_1(p_F^1 - p_F^0)$ and $V_0(p_F^1 - p_F^0)$ are again the sets of net gainers and net losers from public-good provision when the difference in payments is $p_F^1 - p_F^0$.*

# 6 Welfare Implications

### Limits to First-Best Implementation

We now turn to the welfare implications of imposing coalition-proofness, as well as robust implementability. We begin with an example that illustrates some of the issues that arise.

**Example 1** *In this example, there are three possible payoff types $V = \{0, 5, 10\}$. The per-capita cost of public-good provision is $k = 4.5$. There are two possible cross-section distributions $s^j$, $j = 1, 2$ of payoff types. The population shares $s_v^j$ of the different payoff types under these two cross-section distributions are given in the following table.*

| $j$ | $s_0^j$ | $s_5^j$ | $s_{10}^j$ | $\bar{\mathbf{v}}(s^j)$ |
|---|---|---|---|---|
| *1* | *0.3* | *0.7* | *0* | *3.5* |
| *2* | *0.4* | *0.1* | *0.5* | *5.5* |

(12)

*The last column in the table indicates the cross-section average valuation $\bar{\mathbf{v}}(s^j)$ of the public good for each distribution.*

In this example, first-best implementation requires that the public good should not be provided in state 1 and that the public good should be provided in state 2. With equal cost sharing, the associated payment outcomes would be $p_F^0 = 0$ and $p_F^1 = 4.5$. Given these payments, the

set $C^0$ of opponents of public-good provision consists of all types with valuations 0 and the set $C^1$ of net beneficiaries of public-good provision consists of all types with valuations 5 and 10. From Table 1, one immediately sees that the set of net beneficiaries has population share 0.8 in state 1 and 0.7 in state 2. Because the population share of the set of net beneficiaries is larger in state 1 than in state 2, first-best implementation runs afoul of the monotonicity requirement in Theorems 1 and 2. In more concrete terms, any mechanism that would implement a social choice function with first-best outcomes would be vulnerable to a manipulation by a coalition organizer who targets participants with valuations 5 and 10, promising that, in state 1, he will ensure that the public good is provided after all, on the basis of manipulation requiring 3/4 of his clientele to report the valuation 10, 1/8 to report the valuation 0, and only 1/8 to report the true valuation 5, thereby giving the impression that the true state is 2, rather than 1.

The possibility that robust first-best implementation may run afoul of coalition-proofness is also illustrated by the example in the introduction, with possible valuations 0, 3, and 10, and a per capita provision cost equal to 4. In that example, all cross-section distributions of types involved population shares 0.3 of net beneficiaries and 0.7 of opponents of public-good provision. A robustly implementable and coalition-proof social choice function would have to be insensitive to whatever people report, which is incompatible with the efficiency requirement that the public good be provided if and only if the population share of individuals with valuation 3 is sufficiently large. By contrast to this earlier example, the example here shows that coalition-proofness has bite even if the population share of net beneficiaries differs from state to state.

More generally, we obtain:

**Corollary 4** *If there is a pair of states $s$ and $s'$, such that $s(V_1(k)) \geq s'(V_1(k))$ and $\bar{v}(s) < k < \bar{v}(s')$, then there is no social choice function that yields first best outcomes and is robustly implementable and coalition-proof.*

**Second-Best Considerations**

If condition (b) in Corollary 4 is violated, the overall mechanism designer is faced with a second-best problem. Given the impossibility of achieving efficient outcomes in every state $s$, he must choose between different deviations from efficiency that are compatible with robust implementability and coalition-proofness. For instance, in Example 1, he can decide whether it is better to forego the net benefits from public-good provision in state 2 or to incur the net losses from public-good provision in state 1. He might also want to change the boundary between yes-sayers and no-sayers by imposing a payment scheme that raises more funds than he needs, e.g., by asking for a payment $p_F^1 = 5.1$ if the public good is provided, rather than $p_F^1 = k = 4.5$, in order to turn people with valuations 5 from net beneficiaries into opponents of public-good provision. This would allow him to implement a first-best public-good provision rule, but there would be a waste of resources in state 2, when the public good is provided.

Any assessment of tradeoffs between the different kinds of inefficiency must rely on a system of weights that the mechanism designer assigns to the different states. For specificity, we assume that the mechanism designer has his own prior beliefs and chooses a social choice function in order

to maximize expected aggregate surplus according to these beliefs, subject to the requirements of feasibility, robust implementability and coalition-proofness. Given our characterization robust implementability and coalition-proofness, this is equivalent to the problem of choosing $p_F^0$, $p_F^1$ and $Q_F : \mathcal{M}(V) \to \{0, 1\}$ so as to maximize the expected aggregate surplus

$$E^M[(\bar{\mathbf{v}}(s) - p_F^1)Q_F(s) - p_F^0(1 - Q_F(s))] \tag{13}$$

subject to the feasibility constraints that $p_F^0 \geq 0$, $p_F^1 \geq k$, and the coalition-proofness condition that for every pair $s$ and $s'$, $s(V_1(p_F^1 - p_F^0)) \geq s'(V_1(p_F^1 - p_F^0))$ implies $Q_F(s) \geq Q_F(s')$. The expectations operator $E^M$ in (13) indicates that expectations over $s$ are taken with respect to the mechanism designer's subjective beliefs.

In Example 1, the solution to this second-best problem depend on the probabilities $P_1^M$ and $P_2^M$ that the mechanism designer assigns to the different states. If the benefits of public-good provision are foregone in state 2, then, relative to first-best, there is a net per capita welfare loss of $5.5 - 4.5 = 1.0$ in this state. If the public-good is provided in state 1, when it should not be, the per capita welfare loss is $4.5 - 3.5 = 1.0$. If the mechanism designer deems the two states to be equiprobable, he will be indifferent between excessive provision in state 1 and non-provision in state 2. If he deems state 2 to be more likely than state 1, he will prefer excessive provision in state 1 to non-provision in state 2; the reverse is true if he deems state 1 to be more likely.

In any case, though, non-provision in state 2 is dominated by a scheme involving non-provision in state 1 and provision with a payment $p_F^1 = 5.1 > k$ in state 2. This scheme involves a per capita welfare loss, relative to first-best, that is equal to $5.1 - 4.5 = 0.6$ in state 2 If the mechanism designer deems the two states to be equiprobable, he will prefer this scheme even to an arrangement involving excessive provision of the public good in state 1. Excessive provision of the public good in state 1, i.e., provision of the public good in both states, with non-wasteful payments $p_F^0 = 0$ and $p_F^1 = k = 4.5$ is only preferred if the probability assigned to state 1 is less than 3/8. If the probability assigned to state 1 exceeds 3/8, the second-best social welfare function stipulates (the efficient) non-provision of the public good in state 1 and provision with a wasteful payment requirement in state 2. A wilful waste of resources may thus be part of a second-best solution when first-best solutions are ruled out by robust incentive compatibility and coalition-proofness.

# 7 Concluding Remarks

Our subject in this paper has been the problem of mechanism design for public-good provision in a large economy with prior uncertainty as to whether it is efficient for the public good to be provided or not. In this economy, conditions for individual incentive compatibility are simple because no one individual can affect the aggregate outcome. If there are no participation constraints, therefore, a social choice function that yields first-best outcomes can be implemented simply by asking people about their preferences and having them share the costs evenly if the public good is provided. In some instances, however, such schemes are implausible because they rely on information that (collectively) hurts the people who provide it; and people's willingness to provide this information is based solely on the consideration that, as individuals, they are

unable to affect the outcome anyway. We impose a requirement of (weak) coalition-proofness to eliminate this possibility.

When coalition-proofness is imposed along with robustness, the implementability of a social choice function that yields first-best outcomes can no longer be taken for granted. Social choice functions are robustly implementable and (weakly) coalition-proof if and only if the provision can be characterized by a threshold such that the public good is provided if the population share of the net beneficiaries exceeds the threshold and is not provided if the population share of the net beneficiaries falls short of the threshold. Preference intensities cannot play a role. Net beneficiaries are the people for whom the benefits of the public good exceed the costs of the contribution they have to make; contributions are the same for all people and depend only on whether the public good is provided or not. Generally, such threshold rules cannot be used to implement first-best outcomes, because they are not responsive to the preference intensities of those who benefit and those who are harmed by public-good provision.

Instead of coalition-proofness, we might also have used a weak dominance criterion. In those instances where people are (collectively) hurt by the information that they provide, truthtelling is weakly dominated because in the zero-probability event that one might be pivotal after all, lying would provide for a better outcome and in all other events, it would not make a difference. An analysis that is based on a weak dominance criterion would, however, be limited to the large-economy model, with no scope for extending it to large finite economies. In large finite economies, in the absence of participation constraints, Clarke-Groves mechanisms would provide for dominant-strategy implementation of first-best outcomes through truthtelling. The dominant-strategy property of truthtelling only disappears when one goes to the continuum-economy limit. If one regards the large economy with a continuum of agents as an idealization of finite economies with many participants, this discontinuity in the implications of the weak-dominance criterion must be considered problematic.

By contrast, there does not seem to be any such discontinuity in the application of robustness and coalition-proofness to large finite economies and to the continuum-economy limit. In large finite economies, the criterion of individual incentive compatibility is less simple to apply because, for each agent, there is a small probability that he might be pivotal for public-good provision. Incentive schemes must take account of this possibility. The fact that individual incentive compatibility has bite, however, does not preclude the possibility that coalition-proofness and robustness might have bite as well. Preliminary research on the issue makes us confident that, in contrast to weak dominance, the imposition of coalition-proofness and robustness has similar effects in large finite economies as in the continuum-economy idealization.

## References

Bassetto, M. and Phelan, C. (2008). Tax riots. *Review of Economic Studies*, 75:649–669.

Bierbrauer, F. (2009). Optimal income taxation and public-good provision in a large economy with aggregate uncertainty. Working Paper, Max Planck Institute for Research on Collective Goods.

Bierbrauer, F. and Hellwig, M. (2009). The Samuelson critique of a voluntary exchange theory of public finance: an incentive-theoretic perspective. Working Paper, Max Planck Institute for Research on Collective Goods.

Bergemann, D. and Morris, S. (2005). Robust mechanism design. *Econometrica*, 73:1771–1813.

Bernheim, B., Peleg, B., and Whinston, M. (1986). Coalition-proof Nash equilibria I. concepts. *Journal of Economic Theory*, 42:1–12.

Clarke, E. (1971). Multipart pricing of public goods. *Public Choice*, 11:17–33.

Crémer, J. and McLean, R. (1985). Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53:345–361.

Crémer, J. and McLean, R. (1988). Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica*, 56:1247–1257.

d'Aspremont, C., Crémer, J., and Gérard-Varet, L. (2004). Balanced Bayesian mechanisms. *Journal of Economic Theory*, 115:385–396.

d'Aspremont, C. and Gérard-Varet, L. (1979). Incentives and incomplete information. *Journal of Public Economics*, 11:25–45.

Feldman, M. and Gilles, C. (1985). An expository note on individual risk without aggregate uncertainty. *Journal of Economic Theory*, 35:26–32.

Green, J. and Laffont, J. (1979). *Incentives in Public Decision-Making*. North-Holland Publishing Company.

Groves, T. (1973). Incentives in teams. *Econometrica*, 41:617–663.

Guesnerie, R. (1995). *A Contribution to the Pure Theory of Taxation*. Cambridge University Press.

Güth, W. and Hellwig, M. (1986). The private supply of a public good. *Journal of Economics*, Supplement 5:121–159.

Hellwig, M. (2003). Public-good provision with many participants. *Review of Economic Studies*, 70:589–614.

Hildenbrand, W. (1974). *Core and Equilibria of a Large Economy*. Princeton University Press.

Jackson, M. (2001). A crash course in implementation theory. *Social Choice and Welfare*, 18:655–708.

Johnson, S., Pratt, J., and Zeckhauser, R. (1990). Efficiency despite mutually payoff-relevant private information: The finite type case. *Econometrica*, 58:873–900.

Judd, K. (1985). The law of large numbers with a continuum of i.i.d. random variables. *Journal of Economic Theory*, 35:19–25.

Laffont, J. and Martimort, D. (1997). Collusion under asymmetric information. *Econometrica*, 65:875–911.

Laffont, J. and Martimort, D. (2000). Mechanism design with collusion and correlation. *Econometrica*, 68:309–342.

Ledyard, J. (1978). Incentive compatibility and incomplete information. *Journal of Economic Theory*, 18:171–189.

Mailath, G. and Postlewaite, A. (1990). Asymmetric bargaining procedures with many agents. *Review of Economic Studies*, 57:351–367.

Moore, J. (1992). Implementation, contracts, and renegotiation in environments with complete information. In Laffont, J.-J., editor, *Advances in Economic Theory: Sixth World Congress, vol. I.* Cambridge, UK, Cambridge University Press.

Neeman, Z. (2004). The relevance of private information in mechanism design. *Journal of Economic Theory*, 117:55–77.

Norman, P. (2004). Efficient mechanisms for public goods with use exclusion. *Review of Economic Studies*, 71:1163–1188.

Osborne, M. and Rubinstein, A. (1994). *A course in Game Theory.* MIT Press, Cambridge, MA.

Podczeck, R. (2009). On existence of rich Fubini extensions. *Economic Theory*, forthcoming.

Rob, J. (1989). Pollution claim settlements under private information. *Journal of Economic Theory*, 47:307–333.

Schmitz, P. (1997). Monopolistic provision of excludable public goods under private information. *Public Finance/ Finance Publiques*, 52:89–101.

Sun, Y. (2006). The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126:31–69.

Sun, Y. and Zhang, Y. (2009). Individual risk and Lebesgue extension without aggregate uncertainty. *Journal of Economic Theory*, 144:432–443.

# A Proofs

**Proof of Proposition 1**

"$\Longrightarrow$": Suppose first that $F = (Q_F, p_F)$ is robustly implementable. Fix some arbitrary $\hat{s} \in \mathcal{M}(V)$, and let $(T, \mathcal{T})$ and $\tau : T \to V$ be such that $\tau(T) = V$ and, for some $\hat{\delta} \in \mathcal{M}(T)$, $\hat{s} = \hat{\delta} \circ \tau^{-1}$. Because $F$ is robustly implementable, there exists a mechanism $f$ that implements $F$ on $[(T, \mathcal{T}), \tau, \beta]$ for every common-prior belief system $\beta$. By the revelation principle, there is no loss of generality in assuming that $f$ is an incentive-compatible direct mechanism

Consider a prior $\hat{P}$ such that $\hat{P}(B_t \times B_\delta) = \hat{\delta}(B_t) \cdot \chi_{B_\delta}(\hat{\delta})$ for any measurable sets $B_t \subset T$ and $B_\delta \subset \mathcal{M}(T)$, where $\chi_{B_\delta}$ is the indicator function of $B_\delta$. This prior is obviously compatible with a belief system $\hat{\beta}$ such that $\hat{\beta}(B_\delta|t) = \chi_{B_\delta}(\hat{\delta})$ for all $t \in T$ and all measurable sets $B_\delta \subset \mathcal{M}(T)$. Heuristically, the prior $\hat{P}$ and the posteriors $\hat{\beta}(\cdot|t)$ are such that everybody "knows" that the cross-section distribution of types is $\hat{\delta}$. If $f$ is an incentive-compatible direct mechanism that implements $F$ on $[(T, \mathcal{T}), \tau, \hat{\beta}]$, we must have

$$\int_{\mathcal{M}(T)} [\tau(t)Q_f(\delta) - p_f(t,\delta)]d\hat{\beta}(\delta \mid t) \geq \int_{\mathcal{M}(T)} [\tau(t)Q_f(\delta) - p_f(t',\delta)]d\hat{\beta}(\delta \mid t) \tag{14}$$

and, hence,

$$\tau(t)Q_f(\hat{\delta}) - p_f(t,\hat{\delta}) \geq \tau(t)Q_f(\hat{\delta}) - p_f(t',\hat{\delta})$$

for all $t$ and all $t'$ in $T$. Because $f$ achieves $F$, it follows that

$$\tau(t)Q_F(\hat{s}) - p_F(\tau(t),\hat{s}) \geq \tau(t)Q_F(\hat{s}) - p_F(\tau(t'),\hat{s})$$

for all $t$ and all $t'$. Because $\tau(T) = V$, this is equivalent to the requirement that

$$vQ_F(\hat{s}) - p_F(v,\hat{s}) \geq vQ_F(\hat{s}) - p_F(v',\hat{s}),$$

for all $v$ and all $v' \in V$. Because $\hat{s}$ was chosen arbitrarily, it follows that $F = (Q_F, p_F)$ is ex post incentive-compatible.

"$\Longleftarrow$": Conversely, suppose that the social choice function $F = (Q_F, p_F)$ is ex post incentive-compatible. For any $(T, \mathcal{T})$ and $\tau : T \to V$, consider the direct mechanism $f = (Q_f, p_f)$ such that for any $t \in T$ and any $\delta \in \mathcal{M}(T)$, $Q_f(\delta) = Q_F(\delta \circ \tau^{-1})$ and $p_f(t,\delta) = p_F(\tau(t), \delta \circ \tau^{-1})$. We show that truthtelling is an equilibrium for the game induced by $f$ on the type space $[(T, \mathcal{T}), \tau, \beta]$, regardless of the belief system $\beta$. By the definition of $f$, we have

$$\int_{\mathcal{M}(T)} [\tau(t)Q_f(\delta) - p_f(t,\delta)]d\beta(\delta \mid t) = \int_{\mathcal{M}(T)} [\tau(t)Q_F(\delta \circ \tau^{-1}) - p_F(\tau(t), \delta \circ \tau^{-1})]d\beta(\delta \mid t)$$

for all $t$. Because $F$ is ex post incentive-compatible, it follows that

$$\begin{aligned} \int_{\mathcal{M}(T)} [\tau(t)Q_f(\delta) - p_f(t,\delta)]d\beta(\delta \mid t) &\geq \int_{\mathcal{M}(T)} [\tau(t)Q_F(\delta \circ \tau^{-1}) - p_F(\tau(t'), \delta \circ \tau^{-1})]d\beta(\delta \mid t) \\ &= \int_{\mathcal{M}(T)} [\tau(t)Q_f(\delta) - p_f(t',\delta)]d\beta(\delta \mid t) \end{aligned}$$

for all $t$ and $t'$, which proves that truthtelling is an equilibrium for the game induced by $f$ on $[(T, \mathcal{T}), \tau, \beta]$. By construction also, the truthtelling equilibrium for $f$ achieves $F$.

■

**Proof of Proposition 3**

The proof of Proposition 3 is split into two parts. In the first part, we establish a version of the revelation principle. In the second part, we will show that there is no loss of generality in assuming that coalition joiners use the obedient response strategy $\lambda^*$. Let $[(T, \mathcal{T}), \tau, \beta]$, $f_R$, $\sigma^*$,

and $\pi = (X, \ell, \mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ be as specified in the proposition. Define $\hat{\ell} : T_{\mu^{-1}(X)} \times \mathcal{M}(T^e_{\mu^{-1}(X)}) \to \mathcal{M}(R)$ so that, for any $t \in \mu^{-1}(X)$ and $\chi^* \in \mathcal{M}(T^e_{\mu^{-1}(X)})$,

$$\hat{\ell}(t, \chi^*) = \ell(\mu(t), \chi^* \circ \mu^{-1}), \tag{15}$$

and consider the manipulation $\hat{\pi} = (\mu^{-1}(X), \hat{\ell}, h_{\mu^{-1}(X)}, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$, where $\lambda$ is the same response strategy as in the manipulation $\pi$. We will show that, if the interim equilibrium $\sigma^*$ for the mechanism $f_R$ on the type space $[(T, \mathcal{T}), \tau, \beta]$ is blocked by the manipulation $\pi$, then it is also blocked by the manipulation $\hat{\pi}$. The argument is, by and large, routine. We proceed in a sequence of steps.

- For both manipulations, $\hat{\pi}$ and $\pi$, the set of nonjoiners is the same, i.e., $h^{-1}_{\mu^{-1}(X)}(\emptyset) = \mu^{-1}(\emptyset)$, by the definition of $h_{\mu^{-1}(X)}$.

- For any $\delta \in \mathcal{M}(T)$, if $\chi^*(\delta, h_{\mu^{-1}(X)}) = \delta \circ h^{-1}_{\mu^{-1}(X)}$ is the cross-section distribution of reports received by the organizer of the manipulation $\hat{\pi}$, then $\chi^*(\delta, h_{\mu^{-1}(X)}) \circ \mu^{-1}$ coincides with the cross-section distribution $\chi(\delta, \mu) = \delta \circ \mu^{-1}$ of reports received by the organizer of the original manipulation $\pi$. By the preceding argument, the population shares, $\chi^*(\delta, h_{\mu^{-1}(X)})(\{\emptyset\})$ and $\chi(\delta, \mu)(\{\emptyset\})$, of the sets of nonjoiners are the same for both manipulations. As for the joiners, we compute

$$\begin{aligned}(\chi^*(\delta, h_{\mu^{-1}(X)}) \circ \mu^{-1})(B) &= \left(\delta \circ h^{-1}_{\mu^{-1}(X)}\right)(\{t | \mu(t) \in B \cap X\}) \\ &= \delta(\{t | \mu(t) \in B \cap X\}) \\ &= (\delta \circ \mu^{-1})(B) = \chi(\delta, \mu)(B)\end{aligned}$$

  for any measurable set $B \subset X$.

- For any $\delta \in \mathcal{M}(T)$ and any $\hat{t} \in \mu^{-1}(X)$, the recommendation $\hat{\ell}(\hat{t}, \chi^*(\delta, h_{\mu^{-1}(X)}))$ that the organizer of the manipulation $\pi^*$ provides to a person reporting $\hat{t}$ when the cross-section distribution of reports he receives is $\chi^*(\delta, h_{\mu^{-1}(X)})$ is equal to the recommendation $\ell(\mu(\hat{t}), \chi(\delta, \mu))$ that the organizer of the original manipulation $\pi$ provides to a person reporting $\mu(\hat{t})$ when the cross-section of reports he receives is $\chi(\delta, \mu) = \delta \circ \mu^{-1}$. This follows from (15) and the fact that $\chi^*(\delta, h_{\mu^{-1}(X)}) \circ \mu^{-1} = \chi(\delta, \mu)$.

- For any $\delta \in \mathcal{M}(T)$ and any $t \in T$, when the cross-section distribution of types is $\delta$, the lottery of reports that a of a person of type $t$ submits to the overall mechanism is the same under the manipulation $\hat{\pi}$ as under the manipulation $\pi$. For $t \in \mu^{-1}(X)$, this follows because $\hat{\ell}(t, \chi^*(\delta, h_{\mu^{-1}(X)})) = \ell(\mu(t), \chi(\delta, \mu))$ and because the response strategy $\lambda$ is unchanged. For $t \in T \backslash \mu^{-1}(X)$, its follows from the assumption that the agent's reporting behaviour is given by $\sigma^*(t)$ under both manipulations.

- For any $\delta \in \mathcal{M}(T)$, when the cross-section distribution of types is $\delta$, the cross-section distribution of reports received by the overall mechanism under the manipulation $\hat{\pi}$ is the same as the cross-section distribution of reports received under the manipulation $\pi$, i.e.,

$$g(\delta, \pi) = g(\delta, \hat{\pi}). \tag{16}$$

This follows directly from the preceding observation.

- Given these observations, we also have

$$u(\hat{\pi}, t, r, \delta) = u(\pi, t, r, \delta) \tag{17}$$

for all types $t$, reports $r$ to the overall mechanism, and all type distributions $\delta$, where, as in the text,

$$u(\pi', t, r, \delta) = \tau(t) Q_{f_R}(g(\delta, \pi')) - p_{f_R}(r, g(\delta, \pi'))$$

for $\pi' = \hat{\pi}$ and for $\pi' = \pi$.

- From (17), we obtain

$$U_N(\hat{\pi}, t, r) = U_N(\pi, t, r) \tag{18}$$

for all $t \in T$ and all $r \in R$, and

$$U_J(\hat{\pi}, t, \hat{t}, \lambda'(\cdot)) = U_J(\pi, t, \mu(\hat{t}), \lambda'(\cdot)) \tag{19}$$

for all $t \in T$, all $\hat{t} \in \mu^{-1}(X)$, and all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$.

- Because the triple $(\mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ is an interim equilibrium for the game induced by the overall mechanism $f_R$ and the manipulation mechanism $(X, \ell)$, we have

$$U_J(\pi, t, \mu(t), \lambda(t, \cdot)) \geq U_J(\pi, t, \mu(\hat{t}), \lambda'(\cdot))$$

and

$$U_J(\pi, t, \mu(t), \lambda(t, \cdot)) \geq U_N(\pi, t, r)$$

for all $t$ and $\hat{t}$ in $\mu^{-1}(X)$, all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$, and all $r \in R$. By (19) and (18), it follows that

$$U_J(\hat{\pi}, t, t, \lambda(t, \cdot)) \geq U_J(\hat{\pi}, t, \hat{t}, \lambda'(\cdot))$$

and

$$U_J(\hat{\pi}, t, t, \lambda(t, \cdot)) \geq U_N(\hat{\pi}, t, r)$$

for all $t$ and $\hat{t}$ in $\mu^{-1}(X)$, all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$, and all $r \in R$. The triple $(h_{\mu^{-1}(X)}, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ thus satisfies condition (i) for an interim equilibrium of the game induced by $f_R$ and the manipulation mechanism $(\mu^{-1}(X), \hat{\ell})$.

- Because the triple $(\mu, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ is an interim equilibrium for the game induced by the overall mechanism $f_R$ and the manipulation mechanism $(X, \ell)$, we also have

  $$U_N(\pi, t, \sigma^*(t)) \geq U_J(\pi, t, \mu(\hat{t}), \lambda'(\cdot))$$

  and

  $$U_N(\pi, t, \sigma^*(t)) \geq U_N(\pi, t, r)$$

  for all $t \in \mu^{-1}(\emptyset)$, $\hat{t} \in \mu^{-1}(X)$, all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$, and all $r \in R$. By (19) and (18), it follows that

  $$U_N(\hat{\pi}, t, \sigma^*(t)) \geq U_J(\hat{\pi}, t, \hat{t}, \lambda'(\cdot))$$

  and

  $$U_N(\hat{\pi}, t, \sigma^*(t)) \geq U_N(\hat{\pi}, t, r)$$

  for all $t \in \mu^{-1}(\emptyset)$, $\hat{t} \in \mu^{-1}(X)$, all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$, and all $r \in R$. Thus, the triple $(h_{\mu^{-1}(X)}, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ also satisfies condition (ii) for an interim equilibrium of the game induced by $f_R$ and the manipulation mechanism $(\mu^{-1}(X), \hat{\ell})$. In combination with the preceding step, this shows that $(h_{\mu^{-1}(X)}, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ is indeed an interim equilibrium for this game.

- From (19), we also infer that $U_J(\hat{\pi}, t, t, \lambda(t, \cdot)) = U_J(\pi, t, \mu(t), \lambda(t, \cdot))$. If $U_J(\pi, t, \mu(t), \lambda(t, \cdot)) > U(\sigma^*, \sigma^*(t), t)$ for all $t \in \mu^{-1}(X)$, it follows that $U_J(\hat{\pi}, t, t, \lambda(t, \cdot)) > U(\sigma^*, \sigma^*(t), t)$ for all $t \in \mu^{-1}(X)$, i.e., if the interim equilibrium $\sigma^*$ for the mechanism $f_R$ on the type space $[(T, \mathcal{T}), \tau, \beta]$ is blocked by the manipulation $\pi$, then it is also blocked by the manipulation $\hat{\pi}$. This completes the first part of the proof of Proposition 3.

For the second part of the proof, we now consider the mechanism $\pi^*$ that is specified in the proposition. This is the same as $\hat{\pi}$ except that the recommendation function $\hat{\ell}$ is replaced by the function $\ell^*$ that is specified in the proposition and the response strategy $\lambda$ is replaced by the obedient strategy $\lambda^*$. Given that everything else is unchanged, the set of nonjoiners is obviously the same for $\pi^*$ as for $\hat{\pi}$. Moreover, for any $\delta \in \mathcal{M}(T)$, when the cross-section distribution of types is $\delta$, the cross-section distribution of messages, $\chi^*(\delta, h_{\mu^{-1}(X)})$, that is observed by the manipulation organizer is the same for $\pi^*$ as for $\hat{\pi}$. By the definitions of $\hat{\ell}, \ell^*$, and $\lambda^*$, we also have

$$\lambda^*(t, \ell^*(t, \chi^*)) = \ell^*(t, \chi^*) = \lambda(t, \hat{\ell}(t, \chi^*)) \tag{20}$$

for all $t \in \mu^{-1}(X)$ and all $\chi^* \in \mathcal{M}(T^e_{\mu^{-1}(X)})$. Therefore, for any $\delta \in \mathcal{M}(T)$, when the cross-section distribution of types is $\delta$, the cross-section distribution of messages received by the overall mechanism is the same for $\pi^*$ as for $\hat{\pi}$, i.e.,

$$g(\delta, \pi^*) = g(\delta, \hat{\pi}). \tag{21}$$

By the same arguments as before, we can infer that

$$U_N(\pi^*, t, r) = U_N(\hat{\pi}, t, r) \tag{22}$$

for all $t \in T$ and all $r \in R$, and

$$U_J(\pi^*, t, \hat{t}, \lambda'(\cdot)) = U_J(\hat{\pi}, t, \hat{t}, \lambda' \circ \lambda(t, \cdot)) \tag{23}$$

for all $t \in T$, all $\hat{t} \in \mu^{-1}(X)$, and all response strategies $\lambda' : \mathcal{M}(R) \to \mathcal{M}(R)$. By the same arguments as before, one also finds that, if the triple $(h_{\mu^{-1}(X)}, \lambda, \sigma^*_{\mu^{-1}(\emptyset)})$ is an interim equilibrium of the game induced by $f_R$ and the manipulation mechanism $(\mu^{-1}(X), \hat{\ell})$, then the triple $(h_{\mu^{-1}(X)}, \lambda^*, \sigma^*_{\mu^{-1}(\emptyset)})$ is an interim equilibrium of the game induced by $f_R$ and the manipulation mechanism $(\mu^{-1}(X), \ell^*)$.

From (23), we also have $U_J(\pi^*, t, t, \lambda^*(t, \cdot)) = U_J(\hat{\pi}, t, t, \lambda(t, \cdot))$ for all $t$. If $U_J(\hat{\pi}, t, t, \lambda(t, \cdot)) > U(\sigma^*, \sigma^*(t), t)$ for all $t \in \mu^{-1}(X)$, it follows that $U_J(\pi^*, t, t, \lambda(t, \cdot)) > U(\sigma^*, \sigma^*(t), t)$ for all $t \in \mu^{-1}(X)$, i.e., if the interim equilibrium $\sigma^*$ for the mechanism $f_R$ on the type space $[(T, \mathcal{T}), \tau, \beta]$ is blocked by the manipulation $\hat{\pi}$, then it is also blocked by the manipulation $\pi^*$.

**Proof of Proposition 4**

Let $F = (Q_F, p_F)$ be an anonymous, robustly implementable and coalition-proof social choice function. By Corollary 1, the payment rule $p_F$ takes the form $p_F(v, s) = \bar{p}_F(s)$, so that an agent's payment is independent of his own valuation $v$. For any type set $(T, \mathcal{T})$, and $\tau : T \to V$, let $f = (Q_f, p_f)$ be a direct mechanism such that

$$Q_f(\delta) = Q_F(\delta \circ \tau^{-1}) \quad \text{and} \quad p_f(t, \delta) = \bar{p}_F(\delta \circ \tau^{-1}) \,, \tag{24}$$

for all $\delta \in \mathcal{M}(T)$ and all $t \in T$. Then, trivially, for any belief system $\beta : T \to \mathcal{M}(\mathcal{M}(T))$, truthtelling is an interim equilibrium for the game induced by $f$ on the type space $[(T, \mathcal{T}), \tau, \beta]$; moreover, for any type distribution $\delta$, the equilibrium outcome is $(Q_F(\delta \circ \tau^{-1}), \bar{p}_F(\delta \circ \tau^{-1}))$, as stipulated by the social choice function $F$. Thus, $f$ implements $F$ on $[(T, \mathcal{T}), \tau, \beta]$ for every belief system $\beta : T \to \mathcal{M}(\mathcal{M}(T))$.

We claim that, for any common-prior belief system $\beta$, the truthtelling equilibrium of the game induced by the mechanism $f$ on $[(T, \mathcal{T}), \tau, \beta]$ is also coalition-proof. To establish this claim, we will show that, if the truthtelling equilibrium of the game induced by the mechanism $f$ on $[(T, \mathcal{T}), \tau, \beta]$ is not coalition-proof, then the social choice function $F$ itself is not coalition-proof.

If the truthtelling equilibrium of the game induced by the mechanism $f$ on $[(T, \mathcal{T}), \tau, \beta]$ is not coalition-proof, there exist a common-prior belief system $\hat{\beta}$ and a manipulation $\pi$ such that $\pi$ blocks the truthtelling equilibrium of the game induced by the mechanism $f$ on $[(T, \mathcal{T}), \tau, \hat{\beta}]$. By Proposition 3, we may assume that $\pi$ takes the form $\pi = (T_\pi, \ell, h_{T_\pi}, \lambda^*, h_{T \setminus T_\pi})$, where $T_\pi$ is the set of types joining the manipulating coalition, $h_{T_\pi}$ is truthtelling of joiners towards the manipulation mechanism, $\lambda^*$ is the obedient response to the manipulation mechanism's recommendations, and $h_{T \setminus T_\pi}$ is truthtelling of nonjoiners towards the overall mechanism.

Consider any other mechanism $f_R = (Q_{f_R}, p_{f_R})$ and interim equilibrium $\sigma^*$that implement $F$ on $[(T, \mathcal{T}), \tau, \beta]$ for all $\beta$. Then

$$Q_{f_R}(\delta \circ \sigma^{*-1}) = Q_F(\delta \circ \tau^{-1}) \quad \text{and} \quad p_{f_R}(\sigma^*(t), \delta \circ \sigma^{*-1}) = \bar{p}_F(\delta \circ \tau^{-1}), \tag{25}$$

for all $\delta \in \mathcal{M}(T)$ and all $t \in T$.

To show that $\sigma^*$ is not coalition-proof, we consider a manipulation $\pi_R = (T_\pi, \ell_R, h_{T_\pi}, \lambda^*, \sigma^*_{T \setminus T_\pi})$, where $\ell_R : T_\pi \times \mathcal{M}(T^e_\pi) \to \mathcal{M}(R)$ is specified so that, for any $t \in T_\pi$ and any $\chi \in \mathcal{M}(T^e_\pi)$,

$$\ell_R(t, \chi) = \ell(t, \chi) \circ \sigma^{*-1} \tag{26}$$

The set of nonjoiners is the same for $\pi_R$ as for $\pi$, as is the strategy $h_{T_\pi}$ that determines people's reports to the manipulation mechanism. For any $\delta \in \mathcal{M}(T)$, therefore, when the cross-section distribution of types is $\delta$, the cross-section distribution of messages, $\chi(\delta, h_{T_\pi}) = \delta \circ h^{-1}_{T_\pi}$, that is observed by the manipulation organizer is the same for $\pi_R$ as for $\pi$. By (26) and the obedience of coalition joiners to the recommendations that they receive, it follows that, for any $\delta \in \mathcal{M}(T)$, when the cross-section distribution of types is $\delta$, the distributions of messages received by the mechanism $f_R$ under the manipulation $\pi_R$ and by the mechanism $f$ under the manipulation $\pi$ are related by the equation

$$g_R(\delta, \pi_R) = g(\delta, \pi) \circ \sigma^{*-1}.$$

Using (25) and (24), we now compute

$$Q_{f_R}(g_R(\delta, \pi_R)) = Q_{f_R}(g(\delta, \pi) \circ \sigma^{*-1}) = Q_F(g(\delta, \pi) \circ \tau^{-1}) = Q_f(g(\delta, \pi)) \ ,$$

and

$$p_{f_R}(\sigma^*(\hat{t}), g_R(\delta, \pi_R)) = p_{f_R}(\sigma^*(\hat{t}), g(\delta, \pi) \circ \sigma^{*-1}) = \bar{p}_F(g(\delta, \pi) \circ \tau^{-1}) = p_f(\hat{t}, g(\delta, \pi)) \ ,$$

for any $\delta \in \mathcal{M}(T)$ and $\hat{t} \in T$. For any $t \in T$, we therefore have

$$\begin{aligned} &\int [\tau(t) Q_{f_R}(g_R(\delta, \pi_R)) - p_{f_R}(\sigma^*(\hat{t}), g_R(\delta, \pi_R))] d\beta(\delta|t) \\ &= \int [\tau(t) Q_F(g(\delta, \pi) \circ \tau^{-1}) - \bar{p}_F(g(\delta, \pi) \circ \tau^{-1})] d\beta(\delta|t) \\ &\qquad = \int [\tau(t) Q_f(g(\delta, \pi)) - p_f(\hat{t}, g(\delta, \pi))] d\beta(\delta|t) \ , \end{aligned} \tag{27}$$

for any $\hat{t} \in T$ and any belief system $\beta$. Because the manipulation $\pi$ blocks the truthtelling equilibrium of the game induced by the direct mechanism $f$ on $[(T, \mathcal{T}), \tau, \hat{\beta}]$, it follows that

$$\begin{aligned} &\int [\tau(t) Q_{f_R}(g_R(\delta, \pi_R)) - p_{f_R}(\sigma^*(\hat{t}), g_R(\delta, \pi_R))] d\hat{\beta}(\delta|t) \\ &\qquad > \int [\tau(t) Q_f(\delta) - p_f(\hat{t}, \delta)] d\hat{\beta}(\delta|t) \\ &\qquad = \int [\tau(t) Q_F(\delta \circ \tau^{-1} \delta) - \bar{p}_F(\delta \circ \tau^{-1})] d\hat{\beta}(\delta|t) \\ &= \int [\tau(t) Q_{f_R}(\delta \circ \sigma^{*-1}) - p_{f_R}(\sigma^*(\hat{t}), \delta \circ \sigma^{*-1})] d\hat{\beta}(\delta|t) \ , \end{aligned}$$

for all $t \in T_\pi$, i.e., all coalition joiners expect the manipulation $\pi_R$ to raise their payoffs relative to their payoffs under the direct mechanism $f_R$.

To complete the proof that the manipulation $\pi_R$ blocks the interim equilibrium $\sigma^*$ of the game induced by the mechanism $f_R$ on $[(T, \mathcal{T}), \tau, \hat{\beta}]$, it remains to be shown that the triple $(h_{T_\pi}, \lambda^*, \sigma^*_{T \setminus T_\pi})$ is an interim equilibrium of the game induced by the manipulation mechanism $(T_\pi, \ell_R)$ and the overall mechanism $f_R$. As shown by equation (27), the expected payoff of an agent of type $t$ is independent of his report to the manipulation mechanism. It is also independent of his report to the overall mechanism, at least as long as this report belongs to the range of the interim equilibrium $\sigma^*$. For any other report, i.e., for any $r \in R \setminus \sigma^*(T)$, we claim that his expected payoff cannot be higher. For suppose that it was, i.e., that, for some $t \in T$ and some $r \in R \setminus \sigma^*(T)$, we have

$$\begin{aligned} &\int [\tau(t) Q_{f_R}(g_R(\delta, \pi_R)) - p_{f_R}(r, g_R(\delta, \pi_R))] d\hat{\beta}(\delta|t) \\ > &\int [\tau(t) Q_F(g(\delta, \pi) \circ \tau^{-1}) - \bar{p}_F(g(\delta, \pi) \circ \tau^{-1})] d\hat{\beta}(\delta|t). \end{aligned} \tag{28}$$

Then we must have

$$\begin{aligned} &\tau(t) Q_{f_R}(g_R(\delta, \pi_R)) - p_{f_R}(r, g_R(\delta, \pi_R)) \\ > &\tau(t) Q_F(g(\delta, \pi) \circ \tau^{-1}) - \bar{p}_F(g(\delta, \pi) \circ \tau^{-1}) \\ = &\tau(t) Q_{f_R}(g_R(\delta, \pi_R)) - p_{f_R}(\sigma^*(t), g_R(\delta, \pi_R)) \ , \end{aligned}$$

for some $\delta \in \mathcal{M}(T)$. But then, $\sigma^*$ cannot be an interim equilibrium for the game induced by the mechanism $f_R$ on the type space $[(T, \mathcal{T}), \tau, \beta^*]$ where $\beta^*$ is such that, for all $t \in T$, $\beta^*(t)$ assigns all probability mass to the cross-section distribution $g(\delta, \pi)$. The assumption that (28) holds for some $t \in T$ and some $r \in R \setminus \sigma^*(T)$ thus leads to a contradiction and must be false. It follows that $(h_{T_\pi}, \lambda^*, \sigma^*_{T \setminus T_\pi})$ is indeed an interim equilibrium of the game induced by the manipulation mechanism $(T_\pi, \ell_R)$ and the overall mechanism $f_R$, and that the manipulation $\pi_R$ blocks the interim equilibrium $\sigma^*$ of the game induced by the mechanism $f_R$ on $[(T, \mathcal{T}), \tau, \hat{\beta}]$.

■

**Proof of Lemma 1**

Suppose that the Lemma is false. Then there exist $s, \bar{s}$ such that $Q_F(s) = Q_F(\bar{s})$ and $\bar{p}_F(s) > \bar{p}_F(\bar{s})$. Let $[(T, \mathcal{T}), \tau, \beta]$ be such that, for all $t \in T$, $\beta(t)$ assigns probability one to a singleton $\delta$ so that $s = \delta \circ \tau^{-1}$. Then the truthtelling equilibrium for the direct mechanism that implements $F$ on $[(T, \mathcal{T}), \tau, \beta]$ is blocked by a manipulation $\pi = (T, \ell)$ of the grand coalition of all agents, where $\ell$ is specified so that $\ell(t, \delta) = \bar{s} \circ \tau$ for all $t$, and the cross-section distribution of payoff types communicated to the mechanism is $\hat{s}(\delta, \pi) = \bar{s}$. ■

**Proof of Lemma 2**

We first prove that ($a_0$) implies ($b_0$). Suppose that, contrary to ($b_0$), there is a manipulation $\pi_0 \in \Pi_0$ so that

$$\int Q_F(s(\delta)) dP_2(\delta) > \int Q_F(\hat{s}(\delta, \pi_0)) dP_2(\delta).$$

There exists a set $X \subset \mathcal{M}(C_0^e)$ such that $P_2(\{\delta|\chi(\delta, h_{C_0}) \in X\}) > 0$ and, for all $x \in X$,

$$\int Q_F(s(\delta))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x) > \int Q_F(\hat{s}(\delta, \pi_0))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x).$$

If necessary, we may replace $\pi_0 = (C_0, \ell_0, h_{C_0}, \lambda^*, h_{T\setminus C_0})$ by $\hat{\pi}_0 = (C_0, \hat{\ell}_0, h_{C_0}, \lambda^*, h_{T\setminus C_0})$, where, for any $t \in C_0$ and $\chi \in \mathcal{M}(C_0^e)$,

$$\hat{\ell}_0(t, \chi) = \ell_0(t, \chi) \quad \text{if} \quad \chi \in X \quad \text{and} \quad \hat{\ell}_0(t, \chi) = t \quad \text{if} \quad \chi \notin X.$$

Then,

$$\int Q_F(s(\delta))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x) \geq \int Q_F(\hat{s}(\delta, \hat{\pi}_0))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x), \tag{29}$$

for all $x \in \mathcal{M}(C_0^e)$, with

$$\int Q_F(s(\delta))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x) > \int Q_F(\hat{s}(\delta, \hat{\pi}_0))dP_2(\delta|\chi(\tilde{\delta}, h_{C_0}) = x), \tag{30}$$

for all $x \in X$. By the law of iterated expectations, we also have

$$\int Q_F(s(\delta))d\beta(\delta|t) = \int\int Q_F(s(\delta))dP_2(\delta|\chi(\delta', h_{C_0}))d\beta(\delta'|t), \tag{31}$$

and

$$\int Q_F(\hat{s}(\delta, \hat{\pi}_0))d\beta(\delta|t) = \int\int Q_F(\hat{s}(\delta, \hat{\pi}_0))dP_2(\delta|\chi(\delta', h_{C_0}))d\beta(\delta'|t), \tag{32}$$

for all $t \in C_0$. By (29) and (30), it follows that

$$\begin{aligned}
&\int Q_F(s(\delta))d\beta(\delta|t) - \int Q_F(\hat{s}(\delta, \hat{\pi}_0))d\beta(\delta|t) \\
&\geq \int_{\{\delta'|h_{C_0}(\delta') \in X\}} \left[\int Q_F(s(\delta))dP_2(\delta|\chi(\delta', h_{C_0})) \right. \\
&\quad \left. - \int Q_F(\hat{s}(\delta, \hat{\pi}_0))dP_2(\delta|\chi(\delta', h_{C_0}))\right] d\beta(\delta|t)\ ,
\end{aligned}$$

for all $t \in C_0$. Now (30) implies that the integrand on the right-hand side of (33) is strictly positive. Because the belief system is moderately uninformative and $P_2(\{\delta|\chi(\delta, h_{C_0}) \in X\}) > 0$, we also have $\beta(\{\delta|\chi(\delta, h_{C_0}) \in X\}|t) > 0$ for all $t$. For any $t \in C_0$, therefore, the right-hand side of (33) is strictly positive. This shows that the manipulation $\hat{\pi}_0$ blocks the truthtelling equilibrium of the direct mechanism implementing $F$. Thus, if ($b_0$) fails to hold, ($a_0$) fails to hold as well, i.e., ($a_0$) implies ($b_0$).

Conversely, if ($a_0$) fails to hold, there exists a manipulation $\hat{\pi}_0 \in \Pi_0$ such that

$$\int Q_F(s(\delta))d\beta(\delta|t) - \int Q_F(\hat{s}(\delta, \hat{\pi}_0))d\beta(\delta|t) > 0$$

for all $t \in C_0$. Again using (31) and (32), we may infer that there is a set $X$ with $P_2(\{\delta|\chi(\delta, h_{C_0}) \in X\}) > 0$ such that (30) holds for all $x \in X$. If necessary, we may replace the manipulation $\hat{\pi}_0 = (C_0, \hat{\ell}_0, h_{C_0}, \lambda^*, h_{T\setminus C_0})$ by the manipulation $\pi_0^* = (C_0, \ell_0^*, h_{C_0}, \lambda^*, h_{T\setminus C_0})$ where $\ell_0^*(t, \chi) = \hat{\ell}_0(t, \chi)$ if $\chi \in X$ and $\ell_0^*(t, \chi) = t$ if $\chi \notin X$. Then, obviously,

$$\int Q_F(s(\delta))dP_2(\delta) > \int Q_F(\hat{s}(\delta, \pi_0^*))dP_2(\delta),$$

i.e., ($b_0$) fails to hold as well as ($a_0$). ■

**Proof of Lemma 3**

The proof is analogous to the proof of Lemma 2 and is therefore omitted. ■

**Proof of Lemma 4**

We first prove that (a*) implies (b*). The argument proceeds in two steps. We first show that (a*) implies the second statement in (b*), i.e., if (a*) holds, then, for any $s$ and $s'$ in $\mathcal{M}(V)$, $s(V_0(p_F^1 - p_F^0)) = s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) = Q_F(s')$.

For suppose that $s(V_0(\Delta)) = s'(V_0(\Delta))$ and $s(V_1(\Delta)) = s'(V_1(\Delta))$, where $\Delta := p_F^1 - p_F^0$. Let $\bar{s}$ be a further type distribution so that, for any measurable set $B \subset V$, and

$$\bar{s}(B) = s'(B \cap V_0(\Delta)) + s(B \cap (V \backslash V_0(\Delta))). \tag{33}$$

Let $(T, \mathcal{T}), \tau, \delta, \delta', \bar{\delta}$ be such that $s = \delta \circ \tau^{-1}$, $s' = \delta' \circ \tau^{-1}$, and $\bar{s} = \bar{\delta} \circ \tau^{-1}$. Let $P, P', \bar{P}$ be common priors on $T \times \mathcal{M}(T)$ such that the marginal distributions on $\mathcal{M}(T)$ are all degenerate, with

$$P_2(\{\delta\}) = P_2'(\{\delta'\}) = \bar{P}_2(\{\bar{\delta}\}) = 1.$$

The associated belief systems $\beta, \beta', \bar{\beta}$ are also all degenerate so that all types assign all probability mass to $\delta, \delta', \bar{\delta}$, respectively. Because beliefs are type-independent, the belief systems $\beta, \beta', \bar{\beta}$ are also all moderately uninformative.

If (a*) holds, there is no manipulation $\pi_0 \in \Pi_0$ that blocks the truthtelling equlibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$ or $[(T, \mathcal{T}), \tau, \bar{\beta}]$. By Lemma 2, it follows that

$$\int Q_F(s(\delta)) dP_2(\delta) \leq \int Q_F(\hat{s}(\delta, \pi_0)) dP_2(\delta),$$

or, equivalently,

$$Q_F(s) \leq Q_F(\hat{s}(\delta, \pi_0)), \tag{34}$$

for any manipulation $\pi_0 = (C_0, \ell_0)$, and, similarly,

$$Q_F(\bar{s}) \leq Q_F(\hat{s}(\bar{\delta}, \bar{\pi}_0)), \tag{35}$$

for any manipulation $\bar{\pi}_0 = (C_0, \bar{\ell}_0)$. By (33), we have $s(V_0(\Delta)) = \bar{s}(V_0(\Delta))$ and $s(B) = \bar{s}(B)$ for any $B \subset (V \backslash V_0(\Delta))$. Therefore, there exist $\ell_0, \bar{\ell}_0$ such that the associated manipulations satisfy

$$g(\delta, \pi_0) = \bar{\delta} \text{ and } g(\bar{\delta}, \bar{\pi}_0) = \delta,$$

hence,

$$\hat{s}(\delta, \pi_0) = \bar{s} \text{ and } \hat{s}(\bar{\delta}, \bar{\pi}_0) = s. \tag{36}$$

Upon combining (36) with (34) and (35), we obtain

$$Q_F(s) = Q_F(\bar{s}). \tag{37}$$

To prove that $Q_F(s) = Q_F(s')$, it thus suffices to show that $Q_F(\bar{s}) = Q_F(s')$. For this purpose, we use the fact that, if (a*) holds, there is also no manipulation $\pi_1 \in \Pi_1$ that blocks the truthtelling equlibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta']$ or $[(T, \mathcal{T}), \tau, \bar{\beta}]$. By Lemma 3, it follows that

$$Q_F(\bar{s}) \geq Q_F(\hat{s}(\bar{\delta}, \bar{\pi}_1)), \tag{38}$$

and

$$Q_F(s') \geq Q_F(\hat{s}(\delta', \pi_1')), \tag{39}$$

for any manipulations $\bar{\pi}_1 = (C_1, \bar{\ell}_1), \pi_1' = (C_1, \ell_1')$. To assess what lies $\bar{\ell}_1, \ell_1'$ are available to the coalition $C_1$, we rewrite (33) as

$$\bar{s}(B) = s'(B \cap V_0(\Delta)) + s(B \cap \{\Delta\}) + s(B \cap V_1(\Delta))). \tag{40}$$

Because $s(V_0(\Delta)) = s'(V_0(\Delta))$ and $s(V_1(\Delta)) = s'(V_1(\Delta))$, we have

$$s(\{\Delta\}) = 1 - s(V_0(\Delta)) - s(V_1(\Delta)) = 1 - s(V_0(\Delta)) - s(V_1(\Delta)) = s'(\{\Delta\}).$$

Therefore, (40) can be rewritten as

$$\bar{s}(B) = s'(B \cap (V \backslash V_1(\Delta)) + s(B \cap V_1(\Delta))),$$

which yields $s'(V_1(\Delta)) = \bar{s}(V_1(\Delta))$ and $s'(B) = \bar{s}(B)$ for any $B \subset (V \backslash V_1(\Delta))$. Therefore, there exist $\bar{\ell}_1, \ell_1'$ such that the associated manipulations satisfy

$$g(\bar{\delta}, \bar{\pi}_1) = \delta' \text{ and } g(\delta', \pi_1^{'}) = \bar{\delta},$$

hence,

$$\hat{s}(\bar{\delta}, \bar{\pi}_1) = s' \text{ and } \hat{s}(\delta', \pi_1') = \bar{s}. \tag{41}$$

Upon combining (41) with (38) and (39), we obtain $Q_F(\bar{s}) = Q_F(s')$. Thus (a*) implies the validity of the second statement in (b*).

To prove that (a*) also implies the first statement in (b*), suppose that $s$ and $s'$ in $\mathcal{M}(V)$ are such that $s(V_0(\Delta)) \geq s'(V_0(\Delta))$ and $s(V_1(\Delta)) \leq s'(V_1(\Delta))$ where, as before, $\Delta := p_F^1 - p_F^0$. Let $(T, \mathcal{T}), \tau, \delta, \delta'$ be such that $s = \delta \circ \tau^{-1}$ and $s' = \delta' \circ \tau^{-1}$. If (a*) holds, the same argument as before implies that

$$Q_F(s) \leq Q_F(\hat{s}(\delta, \pi_0)), \tag{42}$$

for any manipulation $\pi_0 = (C_0, \ell_0) \in \Pi_0$ and

$$Q_F(s') \geq Q_F(\hat{s}(\delta', \pi_1')), \tag{43}$$

for any manipulation $\pi_1' = (C_1, \ell_1') \in \Pi_1$. If $s(\{\Delta\}) \leq s'(\{\Delta\})$, there exists a manipulation $\pi_0 = (C_0, \ell_0) \in \Pi_0$ such that $\hat{s}(\delta, \pi_0)$ assigns the same mass as $s'$ to each of the three sets $V_0(\Delta), V_1(\Delta)$, and $\{\Delta\}$. By the first part of the argument, this manipulation satisfies $Q_F(\hat{s}(\delta, \pi_0)) = Q_F(s')$, so (42) yields $Q_F(s) \leq Q_F(s')$. If instead, $s(\{\Delta\}) \geq s'(\{\Delta\})$, there exists a manipulation $\pi_1' = (C_1, \ell_1') \in \Pi_1$ such that $\hat{s}(\delta, \pi_1)$ assigns the same mass as $s$ to each of the three sets $V_0(\Delta), V_1(\Delta)$,

and $\{\Delta\}$. By the first part of the argument, this manipulation satisfies $Q_F(\hat{s}(\delta, \pi_1)) = Q_F(s)$, so (43) yields $Q_F(s) \leq Q_F(s')$. In either case, if $s(\{\Delta\}) \leq s'(\{\Delta\})$ and if $s(\{\Delta\}) \geq s'(\{\Delta\})$, we find that $s(V_0(\Delta)) \geq s'(V_0(\Delta))$ and $s(V_1(\Delta)) \leq s'(V_1(\Delta))$ imply $Q_F(s) \leq Q_F(s')$. This completes the proof that (a*) implies (b*).

To prove the converse, suppose that (a*) is not true. Then there exists a common-prior type space $[(T, \mathcal{T}), \tau, \beta]$ with moderately informative beliefs such that the truthtelling equilibrium of the revelation mechanism implementing $F$ on $[(T, \mathcal{T}), \tau, \beta]$ is blocked by some manipulation $\pi_0 \in \Pi_0$ or some manipulation $\pi_1 \in \Pi_1$. Suppose that the blocking manipulation is $\pi_0 \in \Pi_0$. By Lemma 2, we must have

$$\int Q_F(s(\delta))dP_2(\delta) > \int Q_F(\hat{s}(\delta, \pi_0))dP_2(\delta),$$

where $P_2$ is the marginal distribution on $\mathcal{M}(T)$ that is induced by the common prior $P$. It follows that

$$Q_F(s(\delta)) > Q_F(\hat{s}(\delta, \pi_0)), \tag{44}$$

for some $\delta$. For any manipulation $\pi_0 \in \Pi_0$ and any $\delta$, we must have

$$g(C_0|\delta, \pi_0) \leq \delta(C_0)$$

and

$$g(C_1|\delta, \pi_0) \geq \delta(C_1),$$

hence

$$s(V_0(\Delta)|\delta) = s(\tau(C_0)|\delta) = \delta(C_0) \geq g(C_0|\delta, \pi_0) = \hat{s}(\tau(C_0)|\delta, \pi_0) = \hat{s}(V_0(\Delta)|\delta, \pi_0), \tag{45}$$

and

$$s(V_1(\Delta)|\delta) = s(\tau(C_1)|\delta) = \delta(C_1) \leq g(C_1|\delta, \pi_0) = \hat{s}(\tau(C_1)|\delta, \pi_0) = \hat{s}(V_1(\Delta)|\delta, \pi_0). \tag{46}$$

Now (44), (45), and (46) imply that (b*) is not true. If the blocking manipulation is $\pi_1 \in \Pi_1$, Lemma 3 yields

$$Q_F(s(\delta)) < Q_F(\hat{s}(\delta, \pi_1)),$$

for some $\delta$; moreover, $\pi_1 \in \Pi_1$ implies $s(V_1(\Delta)|\delta) \geq \hat{s}(V_1(\Delta)|\delta, \pi_1)$ and $s(V_0(\Delta)|\delta) \leq \hat{s}(V_0(\Delta)|\delta, \pi_1)$, from which one again derives a contradiction to (b*). Thus (b*) fails to hold whenever (a*) fails to hold. ■

# B Proof of Theorem 2

In this second Appendix, we prove Theorem 2. The line of argument is roughly the same as for Theorem 1. We begin by stating an analogue of Corollary 2.

**Corollary 5** *A social choice function $F = (Q_F, \bar{p}_F)$ with type independent payments is robustly implementable and weakly coalition-proof if and only if there is no common prior type space $[(T, \mathcal{T}), \tau, \beta]$ and no subset $T'$ of $T$ with a manipulation $\pi = (T', \ell)$ such that*

$$\int [\tau(t)Q_F(\hat{s}(\delta, \pi)) - \bar{p}_F(\hat{s}(\delta, \pi))]\, d\beta(\delta|t) > \int [\tau(t)Q_F(s(\delta)) - \bar{p}_F(s(\delta))]\, d\beta(\delta|t), \tag{47}$$

*and,moreover, there is no subset $T''$ of $T'$ with a submanipulation $\pi^s$ so that, for all $t \in T''$,*

$$\begin{aligned} &\int [\tau(t)Q_F(\hat{s}(\delta, \pi(\pi^s))) - \bar{p}_F(\hat{s}(\delta, \pi(\pi^s)))]\, d\beta(\delta|t) \\ &> \int [\tau(t)Q_F(\hat{s}(\delta, \pi)) - \bar{p}_F(\hat{s}(\delta, \pi))]\, d\beta(\delta|t), \end{aligned} \tag{48}$$

*where $\pi(\pi^s)$ denotes the combined effect of the manipulation $\pi$ and the submanipulation $\pi^s$ and, as before, $\hat{s}(\delta, \pi(\pi^s)) = g(\delta, \pi(\pi^s)) \circ \tau^{-1}$ denotes the resulting payoff type distribution.*

Corollary 5 follows from the same considerations as Corollary 2. Robustness implies that individual payments do not depend on individual type announcements, but only on the cross-section distribution of announcements. By Remark 1 this implies that any behavior, mediated or not by manipulation and submanipulation mechanisms, constitutes an interim equilibrium. The analysis of coalition-proofness can therefore ignore interim equilibrium conditions and focus exclusively on the conditions for blocking by manipulations and submanipulations. A manipulation $\pi = (T', \ell)$ will block the implementation of $F$ on $[(T, \mathcal{T}), \tau, \beta]$ if it satisfies (47) for all $t \in T'$, but will itself be blocked if there is a subset $T'' \subset T'$ with a submanipulation $\pi^s$ satisfying (48) for all $t \in T''$.

Given the weakening of coalition-proofness by the requirement that blocking coalitions must themselves be subcoalition-proof, the following analogue of Lemma 1 is rather harder to prove than Lemma 1 itself.

**Lemma 5** *If a social choice function $F = (Q_F, \bar{p}_F)$ is regular, as well as robustly implementable and weakly coalition-proof, then there exist numbers $p_F^0$ and $p_F^1$ such that $Q_F(s) = 0$ implies $\bar{p}_F(s) = p_F^0$, and, $Q_F(s) = 1$, implies $\bar{p}_F(s) = p_F^1$.*

**Proof** Suppose that the Lemma is false. Then there exists a regular, robustly implementable and weakly coalition-proof social choice function $F = (Q_F, \bar{p}_F)$ and there exist $s, s' \in \mathcal{M}(V)$ such that $Q_F(s) = Q_F(s')$ and $\bar{p}_F(s) \neq \bar{p}_F(s')$. Because $F$ is regular, we may assume that

$$\bar{p}_F(s') = \min_{s''} \ \bar{p}_F(s'') \quad \text{subject to} \quad Q_F(s'') = Q_F(s). \tag{49}$$

Let $(T, \mathcal{T}), \tau, \delta, \delta'$ be such that $s = \delta \circ \tau^{-1}, s' = \delta' \circ \tau^{-1}$, and let $\beta$ be such that $\beta(\{\delta\}|t) = 1$ for all $t \in T$. Let $\pi = (T, \ell)$ be a manipulation by the grand coalition of all agents so that $g(\delta, \pi) = \delta'$, hence, $\hat{s}(\delta, \pi) = g(\delta, \pi) \circ \tau^{-1} = \delta' \circ \tau^{-1} = s'$ and, therefore,

$$(Q_F(\hat{s}(\delta, \pi)), \bar{p}_F(\hat{s}(\delta, \pi)) = (Q_F(s'), \bar{p}_F(s')).$$

Because $Q_F(s) = Q_F(s')$ and $\bar{p}_F(s) > \bar{p}_F(s')$, this manipulation blocks the truthful equilibrium of the revelation mechanism implementing $F$ on $[(T, \mathcal{T}), \tau, \beta]$.

Because $F$ is weakly coalition-proof, it follows that the manipulation $\pi$ by the grand coalition is not subcoalition-proof. Therefore, there exists a submanipulation $\pi^s = (T^s, \ell^s)$ so that

$$\tau(t)Q_F(\hat{s}(\delta, \pi(\pi^s))) - \bar{p}_F(\hat{s}(\delta, \pi(\pi^s))) > \tau(t)Q_F(\hat{s}(\delta, \pi)) - \bar{p}_F(\hat{s}(\delta, \pi)), \tag{50}$$

for all $t \in T^s$. Because $g(\delta, \pi) = \delta'$, we have $\hat{s}(\delta, \pi(\pi^s)) = \hat{s}(\delta', \pi^s)$, and (50) implies that

$$\tau(t)Q_F(\hat{s}(\delta', \pi^s)) - \bar{p}_F(\hat{s}(\delta', \pi^s)) > \tau(t)Q_F(s') - \bar{p}_F(s'), \tag{51}$$

for all $t \in T^s$. Given that, by (49), the payment $\bar{p}_F(s')$ is minimal over the set of $s''$ yielding the same public-good provision level, the subcoalition $\pi^s$ must induce a different public-good provision level, i.e., we must have $Q_F(\hat{s}(\delta', \pi^s)) \neq Q_F(s')$. Because $F$ is regular, we may also assume that

$$\bar{p}_F(\hat{s}(\delta', \pi^s)) = \min_{\ell} \ \bar{p}_F(\hat{s}(\delta', (T^s, \ell)) \quad \text{subject to} \quad Q_F(\hat{s}(\delta', (T^s, \ell))) = Q_F(\hat{s}(\delta', \pi^s)), \tag{52}$$

i.e., that there is no other manipulation that a coalition of types in $T^s$ could use to get the same outcome at a lower payment.

Condition (51) implies that the submanipulation $\pi^s = (T^s, \ell^s)$ can also be interpreted as a manipulation that blocks the truthful equilibrium of the revelation mechanism implementing $F$ on $[(T, \mathcal{T}), \tau, \beta']$ where $\beta'$ is such that $\beta'(\{\delta'\}|t) = 1$ for all $t \in T$. Because $F$ is weakly coalition-proof, it follows that the manipulation $\pi^s$ itself is not subcoalition-proof. Therefore, there exists a further submanipulation $\bar{\pi}^s = (\bar{T}^s, \bar{\ell}^s)$, with $\bar{T}^s \subset T^s$, so that

$$\tau(t)Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) - \bar{p}_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) > \tau(t)Q_F(\hat{s}(\delta', \pi^s)) - \bar{p}_F(\hat{s}(\delta', \pi^s)), \tag{53}$$

for all $t \in \bar{T}^s$. Because $\bar{T}^s \subset T^s$, we can combine (53) and (51), to obtain

$$\tau(t)Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) - \bar{p}_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) > \tau(t)Q_F(s') - \bar{p}_F(s'),$$

for all $t \in \bar{T}^s$. By (49) again, we infer that $Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) \neq Q_F(s')$, hence $Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) = Q_F(\hat{s}(\delta', \pi^s))$, i.e., the submanipulation $\bar{\pi}^s$ has no effect on the public-good provision level. By (52) and the fact that $\bar{T}^s \subset T^s$, it follows that $\bar{p}_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) \geq \bar{p}_F(\hat{s}(\delta', \pi^s))$. However, $Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) = Q_F(\hat{s}(\delta', \pi^s))$ and $\bar{p}_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) \geq \bar{p}_F(\hat{s}(\delta', \pi^s))$ imply that

$$\tau(t)Q_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) - \bar{p}_F(\hat{s}(\delta', \pi^s(\bar{\pi}^s))) \leq \tau(t)Q_F(\hat{s}(\delta', \pi^s)) - \bar{p}_F(\hat{s}(\delta', \pi^s)),$$

for all $t$, contrary to statement (53). The assumption that the lemma is false has thus led to a contradiction. ■

As in the proof of Lemma 1, the argument rests on the observation that, if a social choice function stipulates the same public-good provision level with different payments in different states, then there is scope for the grand coalition of all agents to block the implementation of this social choice function in states involving the higher payment by manipulating reports so as

to induce the lower payment. With weak coalition-proofness instead of coalition-proofness, one must however deal with the possibility that this manipulation itself may not be subcoalition-proof. The idea then is to show that the subcoalition that blocks this manipulation by the grand coalition can itself be treated as a coalition that blocks the implementation of the social choice function in some other state. Moreover, this (sub)coalition itself is subcoalition-proof.

Given Lemma 5, the following adaptations of Lemmas 2 and 3 follow from the fact that all types in $C_0$ have the same interests and so do all types in $C_1$. Because of this homogeneity of interests, any blocking coalition of types in $C_0$ is automatically subcoalition-proof, and so is any blocking coalition of types in $C_1$.

**Lemma 6** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying (7) and for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with a moderately uninformative belief system, the following statements are equivalent:*

*($a_0$) There is no subcoalition-proof manipulation $\pi \in \Pi_0$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$.*

*($b_0$) For every manipulation $\pi \in \Pi_0$,*

$$\int Q_F(s(\delta))dP_2(\delta) \leq \int Q_F(\hat{s}(\delta, \pi))dP_2(\delta), \tag{54}$$

**Lemma 7** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying (7) and for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with a moderately uninformative belief system, the following statements are equivalent:*

*($a_1$) There is no subcoalition-proof manipulation $\pi \in \Pi_1$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$.*

*($b_1$) For every manipulation $\pi \in \Pi_1$,*

$$\int Q_F(s(\delta))dP_2(\delta) \geq \int Q_F(\hat{s}(\delta, \pi))dP_2(\delta), \tag{55}$$

Given these lemmas, the proof of Lemma 4 goes through without change, to yield

**Lemma 8** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying (7), the following statements are equivalent:*

**(a*)** *If $[(T, \mathcal{T}), \tau, \beta]$ is any common prior type space with a moderately uninformative belief system, there is no subcoalition-proof manipulation $\pi = (T_\pi, \ell)$ with $T_\pi = C_0$ or $T_\pi = C_1$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$.*

**(b*)** *For all $s$ and $s'$ in $\mathcal{M}(V)$, $s(V_0(p_F^1 - p_F^0)) \geq s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) \leq s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) \leq Q_F(s')$. In particular, $s(V_0(p_F^1 - p_F^0)) = s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) = Q_F(s')$.*

From Lemmas 5 and 8, we find that any regular social choice function $F = (Q_F, p_F)$ that is robustly implementable and weakly coalition-proof must satisfy statements 1 and 2 in Theorem 2.

It remains to be shown that, if a social choice function satisfies statements 1 and 2 in Theorem 2, then it is robustly implementable and weakly coalition-proof. Proposition 1 shows that, if a social choice function satisfies statement 1 in Theorem 2, then it is robustly implementable. Lemma 8 shows that, if a social choice function also satisfies statement 2 in Theorem 2, then there is no common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with a moderately uninformative belief system such that the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$ is blocked by a subcoalition-proof manipulation $\pi = (T_\pi, \ell)$ with $T_\pi = C_0$ or $T_\pi = C_1$ To complete the argument, we need to extend this finding to *all* subcoalition-proof manipulations $\pi = (T_\pi, \ell)$.

We first consider manipulations $\pi = (T_\pi, \ell)$ with $T_\pi \subset C_0$ or $T_\pi \subset C_1$. The following lemma shows that the equivalence stated in Lemma 8 remains valid if statement (a*) is extended to allow for coalitions of subsets of $C_0$ or $C_1$. The lemma again exploits the homogeneity of interests of types in $C_0$ and of types in $C_1$. It also exploits the fact that any manipulation by a subset $T_\pi \subset C_0$ or $T_\pi \subset C_1$ can be mimicked by $C_0$ or $C_1$ simply by recommending that types in $C_0 \backslash T_\pi$ or $C_1 \backslash T_\pi$ report the truth to the overall mechanism. Implicitly, therefore, any such manipulation by a subset of $C_0$ or $C_1$ is already covered by the preceding lemmas.

**Lemma 9** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying (7), the following statements are equivalent:*

**(a**)** *If $[(T, \mathcal{T}), \tau, \beta]$ is any common prior type space with a moderately uninformative belief system, there is no subcoalition-proof manipulation $\hat{\pi} = (T_{\hat{\pi}}, \hat{\ell})$ with $T_{\hat{\pi}} \subset C_0$ or $T_{\hat{\pi}} \subset C_1$ that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$.*

**(b*)** *For all $s$ and $s'$ in $\mathcal{M}(V)$, $s(V_0(p_F^1 - p_F^0)) \geq s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) \leq s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) \leq Q_F(s')$. In particular, $s(V_0(p_F^1 - p_F^0)) = s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) = Q_F(s')$.*

**Proof** To see that (a**) implies (b*), it suffices to note that (a**) is stronger than statement (a*) in Lemma 8. To prove that (b*) implies (a**), we observe that, by Lemmas 4 and 2, (b*) implies that, for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with moderately uninformative beliefs, (54) holds for all $\pi \in \Pi_0$. In particular, (b*) implies that (54) holds for all $\pi = (C_0, \ell)$ taking the form

$$\ell(t, \chi) = \hat{\ell}(t, \chi) \text{ for } t \in T_{\hat{\pi}} \text{ and } \ell(t, \chi) = t \text{ for } t \in C_0 \backslash T_{\hat{\pi}}. \tag{56}$$

Because any manipulation $\hat{\pi} = (T_{\hat{\pi}}, \hat{\ell})$ with $T_{\hat{\pi}} \subset C_0$ is equivalent to a manipulation $\pi \in \Pi_0$ taking the form (56), it follows that (54) holds for all $\hat{\pi} = (T_{\hat{\pi}}, \hat{\ell})$ with $T_{\hat{\pi}} \subset C_0$. By the same argument as in the proof of Lemma 2, no such manipulation can block the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T,T), \tau, \beta]$. With a moderately uninformative type space, if all types in a proper subset $T_\pi$ of $C_0$ are made strictly better off by a manipulation, the same will be true for all types in $C_0 \backslash T_\pi$, contrary to the assumption that types in $C_0$ do not possess a blocking manipulation. Similarly, by Lemmas 4 and 3, (b*) implies that, for any common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with moderately uninformative beliefs, (55) holds for all $\pi \in \Pi_1$. In particular, (b*) implies that (55) holds for all $\pi = (C_1, \ell)$ taking the form (56) with $C_0$ replaced by $C_1$. By the same argument as before, it follows that no manipulation $\hat{\pi} = (T_{\hat{\pi}}, \hat{\ell})$ with $T_{\hat{\pi}} \subset C_1$ can block the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$. ■

It remains to be shown that there is also no scope for manipulations that are supported by types in $C_0$ and in $C_1$. This is established by the following Lemma, which shows that, for any type space with a moderately uninformative common prior belief system, a joint manipulation by these types generates incentives to free-ride on the contribution of the others for the functioning of the manipulation and therefore, fails to be subcoalition-proof.

**Lemma 10** *For any social choice function $F = (Q_F, p_F)$ with a payment rule satisfying (7), the following statements are equivalent:*

**(a***)** *If $[(T, \mathcal{T}), \tau, \beta]$ is any common prior type space with a moderately uninformative belief system, there is no subcoalition-proof manipulation $\pi = (T_\pi, \ell)$ at all that blocks the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$.*

**(b*)** *For all $s$ and $s'$ in $\mathcal{M}(V)$, $s(V_0(p_F^1 - p_F^0)) \geq s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) \leq s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) \leq Q_F(s')$. In particular, $s(V_0(p_F^1 - p_F^0)) = s'(V_0(p_F^1 - p_F^0))$ and $s(V_1(p_F^1 - p_F^0)) = s'(V_1(p_F^1 - p_F^0))$ imply $Q_F(s) = Q_F(s')$.*

**Proof** To see that (a***) implies (b*), it suffices to note that (a**) is stronger than statement (a*) in Lemma 8. Suppose therefore, that (b*) holds, but there exists a common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with moderately uninformative beliefs such that the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$ is blocked by some subcoalition-proof manipulation $\pi = (T_\pi, \ell)$. By Lemma 9, $T_\pi$ is not a subset of $C_0$ or $C_1$. By (47), we also have $T_\pi \cap \{\tau^{-1}(p_F^1 - p_F^0)\} = \emptyset$. Since $T = C_0 \cup C_1 \cup \{\tau^{-1}(p_F^1 - p_F^0)\}$, it follows that $T_\pi^0 := T_\pi \cap C_0 \neq \emptyset$ and $T_\pi^1 := T_\pi \cap C_1 \neq \emptyset$. The validity of (47) for $t \in T_\pi^0$ implies that there exists a set $D_0 \in \mathcal{M}(T)$ such that $\beta(D_0|t) > 0$ and, moreover,

$$\delta \in D_0 \text{ implies } Q_F(s(\delta)) = 1 \text{ and } Q_F(\hat{s}(\delta, \pi)) = 0. \tag{57}$$

Similarly, the validity of 47) for $t \in T_\pi^1$ implies that there exists a set $D_1 \in \mathcal{M}(T)$ such that

$\beta(D_1|t) > 0$ and, moreover,

$$\delta \in D_1 \text{ implies } Q_F(s(\delta)) = 0 \text{ and } Q_F(\hat{s}(\delta, \pi)) = 1. \tag{58}$$

Because the belief system is moderately uninformative, we actually have $\beta(D_0|t) > 0$ and $\beta(D_1|t) > 0$ for all $t \in T$.

Consider submanipulation $\pi^s = (T^0_\pi, h_{T^0_\pi})$ which recommends that types in $T^0_\pi$ sabotage the manipulation $\pi$ by reporting truthfully to the overall mechanism. Let $\pi(\pi^s)$ denote the combined effect of the manipulation $\pi$ and the submanipulation $\pi^s$ and and let $\hat{s}(\delta, \pi(\pi^s)) = g(\delta, \pi(\pi^s)) \circ \tau^{-1}$ denote the resulting payoff type distribution. Because the submanipulation $\pi^s$ induces types in $T^0_\pi$ to replace messages in $\tau^{-1}(p^1_F - p^0_F)$ or in $C_1$ by messages in $C_0$, we must have

$$\hat{s}(V_0(p^1_F - p^0_F)|\delta, \pi(\pi^s)) \geq \hat{s}(V_0(p^1_F - p^0_F)|\delta, \pi),$$

and

$$\hat{s}(V_1(p^1_F - p^0_F)|\delta, \pi(\pi^s)) \leq \hat{s}(V_1(p^1_F - p^0_F)|\delta, \pi),$$

for all $\delta$. By (b*), it follows that

$$Q_F(\hat{s}(\delta, \pi(\pi^s)) \leq Q_F(\hat{s}(\delta, \pi)), \tag{59}$$

for all $\delta$. Because the submanipulation $\pi^s$ induces types in $T^0_\pi$ to submit messages in $C_0$, as they do in the absence of any manipulation, and because types in $\{\tau^{-1}(p^1_F - p^0_F)\}$ are not involved in $\pi$ or $\pi^s$, we also have

$$\hat{s}(V_0(p^1_F - p^0_F)|\delta, \pi(\pi^s)) \geq s(V_0(p^1_F - p^0_F)|\delta),$$

and

$$\hat{s}(V_1(p^1_F - p^0_F)|\delta, \pi(\pi^s)) \leq s(V_1(p^1_F - p^0_F)|\delta),$$

for all $\delta$. By (b*), it follows that

$$Q_F(\hat{s}(\delta, \pi(\pi^s)) \leq Q_F(s(\delta)), \tag{60}$$

for all $\delta$. From (60) and (58), we infer that

$$Q_F(\hat{s}(\delta, \pi(\pi^s)) < Q_F(\hat{s}(\delta, \pi)), \tag{61}$$

for all $\delta \in D_1$. Because $\beta(D_1|t) > 0$ for all $t \in T$, (59) and (61) together imply that

$$\begin{aligned} &\int [\tau(t) Q_F(\hat{s}(\delta, \pi(\pi^s))) - \bar{p}_F(\hat{s}(\delta, \pi(\pi^s)))] \, d\beta(\delta|t) \\ &> \int [\tau(t) Q_F(\hat{s}(\delta, \pi)) - \bar{p}_F(\hat{s}(\delta, \pi))] \, d\beta(\delta|t), \end{aligned}$$

for all $t \in T_\pi^0$. Thus, the submanipulation $\pi^s = (T_\pi^0, h_{T_\pi^0})$ blocks the manipulation $\pi$. This contradicts the assumption that $\pi$ is subcoalition-proof. The assumption that (b*) holds, but there exists a common prior type space $[(T, \mathcal{T}), \tau, \beta]$ with moderately uninformative beliefs such that the truthtelling equilibrium of the revelation mechanism implementing $(Q_F, p_F^0, p_F^1)$ on $[(T, \mathcal{T}), \tau, \beta]$ is blocked by a subcoalition-proof manipulation $\pi = (T_\pi, \ell)$ has thus led to a contradiction and must be false. This proves that (b*) implies (a***). ■

Lemma 10 shows that, if a social choice function satisfies statements 1 and 2 in Theorem 2, then it is weakly coalition-proof. This completes the proof of Theorem 2.