Robustly Coalition-Proof Incentive Mechanisms for Public Good Provision are Voting Mechanisms and Vice Versa[*]

Felix J. Bierbrauer[†]and Martin F. Hellwig[‡]

December 22, 2015

Abstract

We study the relation between mechanism design and voting in public-good provision. If incentive mechanisms must satisfy conditions of robust coalition-proofness as well as robust incentive compatibility, the participants' contributions to public-good provision can only depend on the level of the public good that is provided and that level can only depend on the population shares of people favouring one level over another. For a public good that comes as a single indivisible unit the outcome depends on whether or not the share of votes in favour of provision exceeds a specified threshold. With more provision levels for the public good more complicated mechanisms can be used but they still involve the counting of votes rather than any measurement of the participants' willingness to pay. The paper thus provides a foundation for the use of voting mechanisms.

Keywords: Public-good provision, Mechanism Design, Voting Mechanisms

JEL: D60, D70, D82, H41

[*] We are grateful for the comments of three anonymous referees and the editor Marco Ottaviani. We also benefitted from discussions with Alia Gizatulina, Mike Golosov, Kristoffel Grechenig, Christian Hellwig, David Martimort, Benny Moldovanu, and Nora Szech.

[†] University of Cologne, Center for Macroeconomic Research, Albertus-Magnus-Platz, 50923 Köln, Germany. Email: bierbrauer@wiso.uni-koeln.de

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

1 Introduction

The modern theory of public-good provision focusses on the incentives that individuals have to reveal their preferences for public goods. The key question is how to calibrate people's payments to their expressions of preferences so that they have no wish either to understate their preferences for the public good (so as to reduce their payments) or to overstate their preferences (so as to get a greater provision level at other people's expense).

In this paper, we argue that, in addition to individual incentive compatibility, we should also impose a condition of coalition-proofness. Coalition-proofness requires that there must not be any group of participants that is able to affect the public-goods provision level by coordinating its members' communication of preferences so as to make some of its members better off without making any other member worse off. We also argue that individual incentive compatibility as well as coalition-proofness should hold robustly, i.e., regardless of the specification of participants' beliefs about the other participants.[1]

We will also show that, under these conditions, *any* incentive mechanism for the provision of a public good must take the form of a *voting mechanism*, i.e., a mechanism under which the level of the public good that is provided depends on the *number* of people preferring this level over the alternatives, without regard to the intensities of people's preferences.[2] For a public good that comes as a single indivisible unit, we will show that any incentive mechanism that satisfies our requirements must take the form of a simple voting mechanism in which people vote for or against the provision of the public good and the provision rule conditions on the shares of votes for the two alternatives. For public goods with more than one provision level, admissible mechanisms can be more complicated but still they rely on numbers of people voting between different alternatives, rather than any expressions of preference intensities.

Our results thus provide a link between mechanism design and political economy approaches to public-goods provision. In practice, how much we spend on public-goods such as highways, national defense, or the judicial system – and also how much we spend on publicly provided private goods such as health care or education – is determined in the political system. Most democratic societies base public decision making on voting under the *"one-person-one-vote-principle"*. This principle governs parliamentary elections, in particular in countries with proportional representation, where different parties are assigned seats in parliament in proportion to their shares on the popular vote. This principle also governs decision making by the elected parliament.[3]

Whereas in most countries voting is about the people who will become members of parliaments and governments, some countries have systems of direct democracy that allows people to

[1] In imposing robustness, we follow Ledyard (1978) and Bergemann and Morris (2005), in imposing coalition-proofness, Bernheim et al. (1986).

[2] We use the term "voting" in the narrow sense in which it is used in political science and political economy, rather than the wider sense in which it is used in social choice theory. Whereas Gibbard (1973) refers to a "voting scheme" as "any scheme which makes a community's choice depend entirely on people's preferences among the alternatives" (see also Satterthwaite (1975)), we reserve the term "voting" to those mechanisms that condition on population shares of agents expressing a preference for one alternative over another.

[3] Violations of the one-person-one-vote principle can be found in federal systems. The use of the electoral college in US presidential elections is a prominent example. These procedures are typically very controversial because they violate the one-person-one-vote principle.

vote directly on whether a public good is to be provided or not. In Switzerland, for example, there regularly are referenda on such questions. At the local level, the question might be whether to have a municipal swimming pool. At the national level, in recent decades, such referenda have approved large investments in the railway system, including the building of major new transalpine tunnels. Since 1989, there also have been several referenda on whether to abolish the Swiss army.[4]

Political science and political economy take it for granted that political decisions are taken through voting under the principle of "one-person-one-vote", and they study the implications for political decision making and for the allocation and use of resources in the public sector.[5] Economists have traditionally been critical because voting fails to take account of preference intensities and therefore can lead to inefficient outcomes.[6] If there are many people opposing the provision of the public good and few people promoting it, a voting mechanism will stipulate non-provision, which is sub-optimal if the proponents could draw very large benefits from the public good, and the opponents do not feel very strongly about the matter. Our analysis shows that this criticism is irrelevant if public-good provision mechanisms must be robustly incentive-compatible and coalition-proof as well as anonymous. Mechanisms that take account of preference intensities necessarily violate one of these conditions.[7]

Because voting can lead to inefficient outcomes, our results imply that it may be impossible to implement a first-best public-good provision rule. In contrast to previous impossibility results, this finding does not involve participation constraints or multi-dimensional hidden characteristics but follows directly from the observation that coalition-proofness and robust incentive compatibility together destroy the possibility of conditioning on intensities of preferences.

The remainder of this paper is organized as follows. Section 2 contains a discussion of related work. Section 3 uses an example to explain in more detail why individual incentive compatibility conditions ought to be supplemented by conditions of coalition-proofness and robustness.

[4] The politics of the choice between representative and direct democracy is an important subject of its own. In countries with a representative system, the political establishment tends to be opposed to referenda, which would reduce its power. However, in highly controversial choices, the greater political legitimacy of a referendum may be needed. For example, in 2011, a referendum in the German region of Baden-Württemberg served to overcome a deadlock between the political authorities and a popular movement opposing a project involving massive investments and massive remodeling of the train station in the city of Stuttgart.

[5] Typically the focus is on majority voting and on a setting with two alternatives. Models of direct democracy often consider a sequence of binary voting decisions and study the conditions under which such a sequence converges to a Condorcet winner. Models of political agency study elections in which an incumbent runs against a newcomer. In Downsian models of political competition voters face a choice between two political platforms.

[6] Thus, Buchanan and Tullock (1962) argue that vote-trading would be desirable because it provides a way to overcome this problem. Similarly, Casella (2005) argues that intensities could be taken into account if voters had an endowment of votes and could assign more votes to issues that are of greater importance to them. Goeree and Zhang (2013) propose to replace votes by monetary bids.

[7] When we refer to voting mechanisms, we do not necessarily mean majority rule. A mechanism involving majority voting can be but need not be optimal. If, at the stage of mechanism design, there is prior information that beneficiaries of the public good feel strongly about it and opponents do not, it may be desirable to have a rule by which the public good is already provided if a sufficiently large minority votes in favor. Majority voting is likely to be desirable if there is no such prior information about potential biases in voting.

Subsequently, Section 4 presents our formal model and introduces the requirements of robust implementability and robust coalition-proofness. Section 5 gives our main result, i.e., the characterization of robustly implementable and robustly coalition-proof public-goods provision and discusses its welfare implications. In Section 6 we discuss several extensions of our analysis. The last section contains concluding remarks. All proofs are in the Appendix.

2 Related work

In social choice and mechanism design theory, there is an extensive literature on the implications of having coalitions coordinating their members' behaviors in order to manipulate overall outcomes.[8] In this literature, Bennett and Conn (1977), Green and Laffont (1979) and Crémer (1996) are most closely related to our work. These papers show that Groves mechanisms, which provide for dominant-strategy implementation of first-best public-good provision rules, are vulnerable to certain deviations by coalitions of agents acting together. In Section 3 below, we replicate their finding and show that in fact there are two distinct types of failures of coalition-proofness: First, there is a failure of coalition-proofness that arises whenever incentive concerns require the mechanism to collect more money than is needed in order to pay for the cost of public-good provision. In this case, a collective deviation by all participants may reduce payments without changing the level of public-good provision. In addition to this type of failure, which has been the focus of the earlier literature, we find a second type of failure in situations in which no one individual is pivotal and therefore, under a Groves mechanism, all individuals make the same payments: A set of individuals can benefit from a collective deviation that eliminates public-good provision because their benefits, while positive, are smaller than the payments they have to make. Likewise, all individuals for whom the benefit exceeds the per capita cost may have an incentive to collectively exaggerate their valuations.

The contributions of Moulin (1980) and Peleg and Sudhölter (1999) are also related to our work. Moulin considers a model of social choice with a finite set of alternatives. Under the assumption that, for some linear order over the set of alternatives, agents have single-peaked preferences, he shows that the median-voter mechanism is dominant-strategy incentive-compatible, for coalitions of agents as well as individuals. [9] Peleg and Sudhölter (1999) extend his analysis to allow for single-peaked preferences over multidimensional sets of alternatives and generalized median-voter mechanisms. For the public-good provision problem with quasi-linear preferences, we establish coalition-proofness not just for the median-voter mechanism, i.e., majority voting, but for *all* voting mechanisms, i.e., all mechanisms that condition only on numbers of votes.

The literature contains different formalizations of the notion of coalition-proofness. In the spirit of Gibbard (1973) and Satterthwaite (1975), the first generation of papers all used a

[8] Early work includes Bennett and Conn (1977), Barbera (1979), Dasgupta et al. (1979), Green and Laffont (1979), and Moulin (1980, 1999).

[9] The median-voter mechanism asks people to report their preferred alternatives and selects the median of the reported "peaks". As is well known, if participants have single-peaked preferences over a linearly ordered set of alternatives, this mechanism is equivalent to majority voting. Moulin (1980) actually deals with the slightly more complicated case of a median-voter mechanism with the addition of phantom voters with fixed and known peaks.

dominant-strategy direct-mechanism approach, requiring that telling the truth about one's preferences be a dominant strategy for coalitions (with a Pareto criterion for assessing the advantages of deviations for coalition members) as well as individuals.[10] This approach pays no attention to the information and incentive problems that might affect the deviating coalitions themselves.

Subsequent work has begun to allow for such restrictions on coalition formation. Thus, for normal-form games of complete information, Bernheim et al. (1986) introduce a concept of coalition-proof Nash equilibrium, in which coalitions are required to be themselves immune to deviations by sub-coalitions (which in turn must be immune...). This restriction on coalition formation plays an important role in Peleg and Sudhölter (1999); without it, generalized median-voter rules for models with single-peaked preferences over multidimensional sets of alternatives would not generally be coalition-proof.[11]

Laffont and Martimort (1997, 2000) and Che and Kim (2006) study "collusion proofness" in a framework involving Bayes-Nash rather than dominant-strategy implementation. They are particularly concerned about internal incentive compatibility conditions, i.e., the question whether coalition members have proper incentives to divulge their private information to "the coalition" and the question whether they have proper incentives to do what "the coalition" stipulates they do. To study the restrictions that such internal incentive compatibility conditions impose on coalitions, they use an extensive-form representation of communication between a (fictitious) coalition organizer and the coalition members.

Our approach in this paper follows the early literature in imposing coalition-proofness axiomatically, without considering coalition formation and communication between coalition members as part of a strategic game. However, unlike the early literature, we use a robust Bayesian approach, and we use the Bernheim et al. (1986) version of coalition-proofness where deviations by coalitions must themselves be immune to deviations by sub-coalitions that in turn are immune to deviations by sub-sub-coalitions, and so forth. Since sub-coalitions with only one member are included, this approach also takes care of individual incentive compatibility in coalitions.

We prefer the Bayesian approach to the dominant-strategy approach because the Bayesian approach is explicit about the role of beliefs and information in participants' decision making. However, we share the view, expressed by Ledyard (1978) and Bergemann and Morris (2005), that mechanism design should not condition on individual beliefs. This is why we impose the Bergemann and Morris (2005) condition of robustness.[12]

It is well known that, at the level of individual decision making, the implications of robust Bayes-Nash and dominant-strategy implementability for mechanism design are the same, so

[10] The requirement is referred to as "group incentive compatibility" by Bennett and Conn (1977), "group strategy-proofness" by Barberà (1979) and Moulin (1980), "coalition incentive compatibility" by Green and Laffont (1979), "implementability in coalitionally dominant-strategies" by Dasgupta et al. (1979), and "strong coalition-proofness" by Peleg and Sudhölter (1999).

[11] Hence their distinction between coalition-proofness, as in Bernheim et al. (1986), and "strong coalition-proofness", which is Barberà's (1979) or Moulin's (1980) "group strategy-proofness".

[12] Börgers and Smith (2014), and Börgers (2015) suggest that this condition might be too strong. Allowing for indirect mechanisms, they suggest that, by the arguments of Ledyard (1978) and Bergemann and Morris (2005), the strategic game induced by an incentive mechanism should be independent of the belief system, but equilibrium outcomes need not be.

the distinction might seem unimportant. However, the distinction makes a difference for the analysis of coalition-proofness. As shown by Laffont and Martimort (1997, 2000), the Bayes-Nash approach lends itself to the analysis of incentive compatibility inside a coalition.

In this respect, the dominant-strategy approach is equivalent to a requirement of *ex post* coalition-proofness, i.e., a requirement that participants cannot gain from collective manipulations that are conditioned on the *actual* state of the economy; it is also equivalent to Bayes-Nash coalition-proofness for all complete-information belief systems, i.e. all belief systems in which participants "know" the state of the economy.[13] Robust Bayes-Nash coalition-proofness, i.e., Bayes-Nash coalition-proofness for all belief systems is of course more demanding than Bayes-Nash coalition-proofness for all complete-information belief systems, or *ex post* coalition-proofness. We require robust Bayes-Nash coalition-proofness but, as we go along, we will explain how our analysis would have to be modified if we used a concept of *ex post* coalition-proofness (or the dominant-strategy approach) instead. Our main findings are unaffected by these changes.

With *ex post* coalition-proofness, or with Bayes-Nash coalition-proofness for complete-information belief systems, there is a presumption that coalitions know even those aspects of the state of the economy that pertain to people who are not coalition members. This presumption is unreasonable.[14] In a companion paper, Bierbrauer and Hellwig (2015), we therefore weaken the concept of coalition-proofness by requiring that collective deviations by blocking coalitions must not depend on belief systems and that such deviations are attractive to coalition members regardless of what their beliefs may be. Using the simplifications provided by a large economy we find that, in the public-good provision problem with private values, social choice functions satisfying a natural monotonicity condition are immune to robustly blocking coalitions if and only if they can be implemented by voting mechanisms. The conclusion is the same, but the concept and the formal arguments are different as they must take account of the lack of information in the design of collective deviations.

Requirements of coalition-proofness had previously been introduced in a series of related papers by Bierbrauer (2009, 2012, 2014) on optimal taxation and public-goods provision in economies with a continuum of agents. The analysis in those papers focuses on the interaction between different dimensions of incentive compatibility that arises when individuals differ not only in their earning abilities, as in Mirrlees (1971), but also in their public-goods preferences.

3 Why Coalition-Proofness Matters

In this section, we use an example to explain why coalition-proofness matters for public-good provision. In the example, there are n individuals who must decide on a public good that comes as a single indivisible unit. The per-capita cost of providing this unit is equal to 4. The benefit

[13] These "equivalences" are to be considered *ceteris paribus*, i.e., as assumptions about information and beliefs are varied, all other aspects of the coalition-proofness concept are held fixed, such as whether or not the Bernheim et al. (1986) restriction to coalitions that are themselves sub-coalition-proof is imposed, whether the Pareto criterion is applied with a strict inequality for all coalition members or with a weak inequality for all and a strict inequality for some, and whether or not side payments between coalition members are allowed.

[14] In Laffont and Martimort (1997, 2000) and Che and Kim (2006), this problem does not arise because the analysis focuses on the grand coalition of all participants.

an agent draws from the public good if provided is either 0, or 3, or 10. Thus, if S_3 and S_{10} are the numbers of agents with valuations 3 and 10, a first-best provision rule requires that the public good be provided if $3S_3 + 10S_{10} > 4n$ and not be provided if $3S_3 + 10S_{10} < 4n$. For specificity, we consider a rule under which the public good is also provided if $3S_3 + 10S_{10} = 4n$.

The valuations $v_i, i = 1, ..., n$, are the participants' private information. In designing an incentive mechanism to elicit this information, one must take account of the strategic interdependence of the different agents' behaviors. Traditionally this is done by requiring implementation through dominant-strategy equilibria or implementation through Bayes-Nash equilibria. We first consider dominant-strategy implementation.

Dominant-Strategy Implementation. A *Clarke-Groves mechanism* serves to implement the efficient provision rule in dominant strategies without requiring the mechanism designer himself to make a contribution to the public good.[15] Such a mechanism asks each agent to report his valuation for the public good and applies the efficient provision rule to the reported valuations. The mechanism determines the payments people must make so that, for each agent, truly reporting his own valuation is a dominant strategy. Together with the provision rule, the payment rules force agents to take account of the externalities that their choices impose on others; see Clarke (1971) and Groves (1973). Thus, in our example, a Clarke-Groves mechanism specifies the payment of any agent i as a function p_i of the agent's reported valuation $\hat{v}_i$ and the numbers $\hat{S}_3$ and $\hat{S}_{10}$ of other agents reporting valuations 3 and 10, such that

$$p_i(\hat{v}_i, \hat{S}_3, \hat{S}_{10}) = h_i(\hat{S}_3, \hat{S}_{10}) + 4n - 3\hat{S}_3 - 10\hat{S}_{10} \quad \text{if} \quad 4n \leq \hat{v}_i + 3\hat{S}_3 + 10\hat{S}_{10} \ , \tag{1}$$

and

$$p_i(\hat{v}_i, \hat{S}_3, \hat{S}_{10}) = h_i(\hat{S}_3, \hat{S}_{10}) \quad \text{if} \quad 4n > \hat{\theta}_i + 3\hat{S}_3 + 10\hat{S}_{10} \ , \tag{2}$$

where

$$h_i(\hat{S}_3, \hat{S}_{10}) = \max[0, 3\hat{S}_3 + 10\hat{S}_{10} - 4(n-1)]. \tag{3}$$

For illustration, we consider the case $n = 10$. Table 1 shows how the payments that agents make depend on their own valuations v_i and on the aggregate valuation $\bar{V} := 3S_3 + 10S_{10}$.[16]

[15] A Clarke-Groves mechanism is a special mechanism in the more general class of Groves mechanisms. Green and Laffont (1977) show that under very general conditions every dominant-strategy mechanism that implements an efficient outcome must be a Groves mechanism.

[16] The entries of the table are derived from equations (1) - (3) by noting that, under truthtelling, the aggregate reported valuation of the other agents, $3\hat{S}_3 + 10\hat{S}_{10}$, is equal to the difference between the overall aggregate valuation $\bar{V}$ and the agent's own valuation v_i. Thus, for $v_i = 3$ and $\bar{V} = 3S_3 + 10S_{10} = 41$, the aggregate valuation of the other agents, $3\hat{S}_3 + 10\hat{S}_{10}$, is 38, so (3) implies $h_i(\hat{S}_3, \hat{S}_{10}) = 38 - 4 \cdot 9 = 2$, and (1) implies $p_i(\hat{v}_i, \hat{S}_3, \hat{S}_{10}) = 4$, as shown in the table.

Table 1.

	$v_i = 0$	$v_i = 3$	$v_i = 10$
$\bar{V} \leq 36$	$p_i = 0$	$p_i = 0$	$p_i = 0$
$36 < \bar{V} < 40$	$p_i = \bar{V} - 36$	$p_i = 0$	$p_i = 0$
$40 \leq \bar{V} < 46$	$p_i = 4$	$p_i = 4$	$p_i = 50 - \bar{V}$
$46 \leq \bar{V}$	$p_i = 4$	$p_i = 4$	$p_i = 4$

Depending on the aggregate valuation $\bar{V}$, there are four distinct cases: If $\bar{V}$ is sufficiently low (below 36) or if $\bar{V}$ is sufficiently high (above 46), the mechanism provides for equal cost sharing, so everybody pays 0 if $\bar{V}$ is low, and everybody pays 4 if $\bar{V}$ is high. In between, the mechanism deviates from equal cost sharing. For $\bar{V}$ between 40 and 46, the public good is provided, agents with valuations 0 and 3 each pay 4, but agents with valuation 10 pay more.[17] For $\bar{V}$ between 36 and 40, the public good is not provided, but even so, agents with valuation 0 must pay. These payments are imposed to ensure truthtelling by agents with valuation 3 in those instances where they are pivotal. Thus, for $\bar{V} = 41$, which results from having $S_0 = 1, S_3 = 7$, and $S_{10} = 2$, any agent with $v_i = 3$ might avert the provision of the public good by reporting $\hat{v}_i = 0$ instead of $\hat{v}_i = v_i = 3$. By (2) and (3), however, his payment would not fall to 0 but to 2.[18] Because the payment reduction is less than the utility loss from not having the public good provided, the agent prefers to report the true valuation 3. One easily checks that, quite generally, truthtelling dominates all other choices. In those instances where an agent is pivotal for the decision on public-good provision, the dominance is actually strict, i.e. truthtelling provides a strictly higher payoff than the alternatives.

However, the Clarke-Groves mechanism in our example is not coalition-proof. To see this, suppose that $S_0 = 0, S_3 = 7$, and $S_{10} = 3$. Then $\bar{V} = 51$, the public good is provided, and everybody pays 4. The seven agents with valuation 3 get a net payoff equal to -1, whereas they would get 0 if they all jointly reported a valuation of 0. To implement the first-best outcome, the mechanism relies on their information but, from their perspective, the first-best outcome is worse than the outcome they would get if they all misrepresented their information. A coalition of these agents would prevent the public good from being provided if $S_0 = 0, S_3 = 7$, and $S_{10} = 3$.

Bennet and Conn (1977) and Green and Laffont (1979) have previously shown that quite generally Groves mechanisms are incompatible with coalition-proofness. Their notion of coalition-proofness however is different from ours in that they consider coalitions to be unconditional arrangements that a subset of agents might have entered into before knowing what their valuations actually are. In contrast, we allow for coalition membership to depend on agents' valuations, so in the example just given we have a coalition of the seven agents with valuation 3.[19]

[17] For example, if $\bar{V} = 42$, the aggregate valuation $3\hat{S}_3 + 10\hat{S}_{10} = \bar{V} - 10$ of agents other than one with valuation 10 is only 32, so (1) and (3) imply that agents with valuation 10 must pay 8. This constellation arises, for instance, if $S_0 = 3, S_3 = 4$, and $S_{10} = 3$.

[18] The mechanism would behave as if S_0 were equal to $2, S_3 = 6$, and $S_{10} = 2$, and the aggregate valuation was equal to 38.

[19] Moreover, our notion of coalition-proofness is consistent with the requirement that the use of privately held

More importantly, whereas Bennet and Conn (1977) and Green and Laffont (1979) proved the failure of coalition-proofness by simply showing that incentive schemes cannot satisfy the conditions for individual incentive compatibility and for coalition incentive compatibility at the same time, we focus on the implications of imposing coalition-proofness as well as individual incentive compatibility. In the process, we find that there are actually *two types of failures of coalition-proofness*, the one considered by Bennet and Conn (1977) and Green and Laffont (1979) and the one considered above. For the coalition just studied, coalition-proofness fails because, for the constellation $S_0 = 0, S_3 = 7,\ S_{10} = 3$, the decision to provide the public good requires truthful reporting by a set of people who all would be better off if the public good was not provided. A second type of failure of coalition-proofness appears when $36 < \bar{V} < 40$. In this case, the grand coalition of all participants can benefit by falsely reporting that all valuations are equal to 0. The provision decision is unaffected but total payments are reduced from $(\bar{V} - 36) \cdot S_0$ to 0.[20] The first type of failure of coalition-proofness involves a coalition changing the level of public good provision, the second type a coalition changing the level of aggregate payments while leaving the level of public-good provision unchanged.[21]

Whereas the analysis of Bennet and Conn (1977) and Green and Laffont (1979) focuses on the second type of failure of coalition-proofness, in our analysis, both types are important. As we will formally show below, the second type of failure of coalition-proofness can only be avoided if the participants' payments depend only on the level of public-good provision. If an incentive-compatible mechanism stipulates the same level of public-good provision for two distinct preference profiles, then it must also stipulate the same values of agents' payments for these two profiles; if it fails to do so, it is not coalition-proof. To avoid the first type of failure of coalition-proofness, the mechanism must abstract from preference intensities and consider only the number of participants expressing an ordinal preference for one outcome over another.

Bayes-Nash Implementation and Robustness. In the Bayes-Nash approach to implementation, the impact of coalition-proofness depends on whether participants are deemed to have independent or correlated values. With independent private values, coalition-proofness imposes similar restrictions on Bayes-Nash implementation as on dominant-strategy implementation. With correlated private values, however, apart from exceptional cases, coalition-proofness im-

information within a coalition must respect incentive constraints. In this respect, we follow Laffont and Martimort (1997, 2000), who treated the problem of organizing a coalition whose members would coordinate their reports as an mechanism design problem of its own, with distinct incentive and participation constraints for all participants. Laffont and Martimort, however, focussed on deviations by the grand coalition of all agents.

[20] Similarly, if $40 \leq \bar{V} < 46$, the grand coalition of all participants can benefit by falsely reporting all valuations are equal to 10. The public good continues to be provided but the aggregate payment is reduced from $40 + (46 - \bar{V}) \cdot S_{10}$ to 40.

[21] This second type of failure of coalition-proofness is related to the failure of budget balance. Whereas the Clarke-Groves mechanism never makes a deficit, it makes a surplus when $36 < \bar{V} \leq 39$ or $40 \leq \bar{V} < 46$. In the argument given, the grand coalition's manipulation of reports eliminates this surplus. Note, however, that a mechanism requiring payments equal to 1 if $\bar{V} < 40$ and 4 if $\bar{V} \geq 40$, regardless of individual valuations, is coalition-proof as well as incentive-compatible but fails to satisfy budget balance if $\bar{V} < 40$. Whereas the second type of failure of coalition-proofness must go along with a failure of budget balance, there is no such link between the first type of failure of coalition-proofness and budget balance.

poses no serious restrictions at all. With correlated values, one can use the approach of Crémer and McLean (1985, 1988) or McAfee and Reny (1992) to construct Bayesian incentive schemes that extract almost all the surplus from public-goods provision from the participants, using the proceeds to pay for the public good. In the above example, such incentive schemes would induce agents with a public-good valuation of 10 to make payments that are more in line with their valuations. The additional payments from these agents (relative to the Clarke-Groves mechanism) can be used to turn agents with a public-good valuation of 3 from being victims to being beneficiaries of public-good provision. With correlated values, therefore, neither the requirement of incentive-compatibility nor the requirement of coalition-proofness is an impediment for implementing first-best outcomes.[22]

The Crémer-McLean or McAfee-Reny mechanisms exploit the fact that, with correlated private values, agents' beliefs about the other agents' valuations - and about the probability of public-good provision - vary as their own valuations vary. In our example, with correlated private values, an agent with a valuation of 3 is likely to have different beliefs than an agent with a valuation of 10. These differences in beliefs induce different attitudes to outcome-contingent payment schemes. These differences in attitudes to outcome-contingent payment schemes can be used to induce agents to reveal their valuations without receiving information rents.[23]

However, the Crémer-McLean or McAfee-Reny mechanisms are extremely sensitive to the specification of agents' beliefs. For example, if correlations are small, beliefs do not vary much, and the gambles involved in outcome contingent payment schemes must be huge to have any significant impact on agents' incentives.

We consider this dependence of incentive mechanisms on the specification of agents' beliefs to be problematic. A mechanism designer cannot be presumed to know the participants' beliefs. Following Bergemann and Morris (2005), one may therefore wish to impose a requirement of robustness with respect to the specification of participants' beliefs, i.e., we will restrict our analysis to Bayes-Nash implementation with incentive mechanisms that do not depend on the specification of the participants' beliefs.[24] This robustness requirement eliminates any reliance on Crémer-McLean or McAfee-Reny mechanisms.

For models with private values, it is well known that there is an equivalence between the requirements of robust Bayes-Nash incentive compatiblity and dominant-strategy incentive compatibility. coalition-proofness therefore restricts the scope for robust Bayes-Nash implementation just as it restricts the scope for dominant-strategy implementation.

[22] For details of the construction, see Bierbrauer and Hellwig (2015).

[23] The construction requires the beliefs that an agent has at different types to satisfy a certain independence condition. In models with finitely many types, this condition holds for a generic (open and dense) set of belief functions whenever the number of possible types of one agent is smaller than the number of possible type constellations of the other agents. Gizatulina and Hellwig (2015) show that, for models with a continuum of types, the analogous independence condition of McAfee and Reny (1992) holds for a residual set of belief functions.

[24] Börgers and Smith (2014) argue that Bergemann and Morris (2005) may be going one step too far. In their view, the mechanism designer's ignorance of participants' beliefs justifies a requirement that the mechanism must be belief-independent, but not a requirement that outcomes under such a mechanism must be belief-independent. As yet, unfortunately, we do not have a simple characterization of the restrictions that the Börgers-Smith approach imposes on mechanism design.

Robust Coalition-Proofness. In the following, we study the implications of coalition-proofness for robust Bayes-Nash implementation, rather than dominant-strategy implementation. The Bayes-Nash approach has the advantage of being explicit about the decision problems that individual agents face. Individuals with given characteristics form expectations about the other agents' behaviors and try to choose best responses to these anticipated behaviors. The Bayes-Nash approach also puts the spotlight on the information that coalitions are presumed to have. This information affects the scope of the potential manipulation strategies that coalitions can engage in. The dominant-strategy approach captures only one of the possible specifications.

In the dominant-strategy approach, coalition-proofness requires that there is no constellation of agents' characteristics for which coordinated false reports by a coalition of participants can change the outcome so that all coalition members are (weakly) better off.[25] In this approach, coalitions can condition their manipulations on the constellation of all agents' characteristics, whether they are members of the coalition or not. In the terminology of Bergemann and Morris (2005), this is a requirement of *ex post* coalition-proofness: after the individuals' types have been revealed, there is no subset of agents who all regret that they did not coordinate on a particular deviation.

In a Bayesian context, for a given belief system, interim coalition-proofness requires that there is no scope for coordinated false reports by a coalition of participants to change outcomes so that all coalition members have (weakly) higher interim expected utilities. Robust coalition-proofness requires that the social choice function be coalition-proof no matter what the belief system may be. For consistent, complete-information belief systems, interim coalition-proofness is equivalent to ex post coalition-proofness as it makes no difference whether participants are assumed to "know" or to have probability-one beliefs about the constellation of all agents' characteristics.

Thus *ex post* coalition-proofness is necessary for robust coalition-proofness, but not sufficient.[26] *Ex post* coalition-proofness does not exclude collective deviations *ex interim* in which individual agents expect to gain in some states of the world and to lose in others, and because of differences in beliefs all participants assign more weight to the potential gains than to the potential losses from the deviation.

Our main result establishes an equivalence relation between robustly incentive-compatible and robustly coalition-proof mechanisms for public-goods provision and voting mechanisms. As a corollary, we show that this characterization also applies to mechanisms that are robustly incentive-compatible and ex post coalition-proof.

In Bierbrauer and Hellwig (2015) we get a version of this conclusion with yet another approach. In that paper, we use a weaker concept of coalition-proofness, with collective deviations that coalition members deem to be advantageous regardless of what their beliefs may be. With this weaker concept of coalition-proofness, for large economies with many participants we find that social choice functions satisfying a natural monotonicity condition are robust implementable by coalition-proof mechanisms if and only if they are implementable by voting mechanisms.

The approach chosen in Bierbrauer and Hellwig (2015) is appropriate if whoever organizes a

[25] See Footnotes 8 and 10.

[26] This statement presumes a ceteris paribus condition as discussed in Footnote 13.

coalition is assumed to have no information about the belief system, like the overall mechanism designer in the robust Bayesian approach. The stronger concept studied in this paper, which allows coalitions to condition their behaviors on the belief system, is appropriate if the overall mechanism designer is concerned that participants, who know the belief system, might use this information as they coordinate on collective deviations.

Apart from the assumption that coalitions can condition their behavior on belief systems, our formalization of coalition-proofness is more restrictive than the notions that have previously been considered in the literature on incentive mechanisms. If anything, our analysis therefore understates the role of coalition-proofness. Specifically, we will model coalition formation as a non-cooperative activity, which involves its own set of incentive and participation constraints. Further we do not allow side payments to enhance the scope for coalition formation. Finally, we require the coalitions themselves to be sub-coalition-proof. We will see that despite these restrictions, the requirement of coalition-proofness has bite and may preclude the implementation of first-best outcomes. The two kinds of failures of coalition-proofness that we saw in the above example will play a key role in the analysis.

4 The Model

Payoffs and Social Choice Functions. The set of individuals is given by $I = \{1, \ldots, n\}$. There is one private good and one public good. The public good comes as a single indivisible unit.[27] Its installation requires aggregate resources equal to $k\, n$ 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 finite set V of possible valuations, which is the same for all i. We assume that V contains 0 as its smallest element and we write $\bar{v}$ for its largest element. We write $v = (v_1, \ldots, v_n)$ for a typical vector of valuations of all individuals.

A social choice function $F = (Q, P_1, \ldots, P_n)$ consists of a public-good provision rule and payment rules for the different agents. The public-good provision rule $Q : V^n \to \{0, 1\}$ determines whether the public good is to be provided or not, as a function of the vector of public-goods valuations. The payment rules $P_i : V^n \to R$, $i = 1, ..., n$, determine the payment of agent i, also as a function of the vector of public-goods valuations.

We limit attention to *anonymous* social choice functions so that individuals with the same valuation make the same payment and a permutation of the individuals' valuations does not affect the decision on public-goods provision. Formally, we say that a social choice function is *anonymous*, if, for every v, and every pair of individuals $(i, j) \in I^2$, $P_i(v_{-i-j}, v_i, v_j) = P_j(v_{-i-j}, v_j, v_i)$ and $Q(v_{-i-j}, v_i, v_j) = Q(v_{-i-j}, v_j, v_i)$, where v_{-i-j} is a vector that lists the valuations of all individuals, with the exception of individuals i and j.

We say that the anonymous social choice function $F = (Q, P_1, \ldots, P_n)$ yields *feasible* outcomes if and only if, in any state of the economy, the aggregate revenue is sufficient to cover the

[27] In Section 5 we discuss the conditions under which our analysis extends to a setting with more than two public-good provision levels.

public-good provision cost $k\ Q(v)$, i.e., if and only if the inequality

$$\sum_{i=1}^{n} P_i(v) \geq k\ Q(v) \tag{4}$$

is satisfied for all $v \in V^n$.[28]

Types and Beliefs. Information about types is assumed to be private. We model information by means of an abstract type space $[T, \tau, \beta]$. 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. It is determined by the function $\beta : T \to \mathcal{M}(T^{n-1})$, where $\mathcal{M}(T^{n-1})$ is the set of probability measures over the possible types of all individuals, except individual i.

For ease of notation, T is taken to be a finite set. We also assume throughout that τ is surjective. We will occasionally use the shorthand notation $\tau(t) = (\tau(t_i), \ldots \tau(t_n))$ for the vector of payoff types that is induced by a vector of types $t = (t_1, \ldots, t_n)$. Sometimes it will also be convenient to write this vector as $(\tau_{-i}(t_{-i}), \tau(t_i))$.

The belief system β is said to be *degenerate* if, for some $t \in T^n$ and all $i \in I$ the measure $\beta(t_i)$ assigns all probability mass to the singleton t_{-i} i.e., if all agents "know" the type profile to be t.

Implementing a social choice function. We seek to implement a social choice function as a truth-telling equilibrium of a mechanism in which individuals communicate their payoff types.[29] Given such a mechanism, a strategy for player i is a function $\sigma_i : T \to V$. A strategy profile is in the following written as $\sigma = (\sigma_1, \ldots, \sigma_n)$. We denote the expected payoff of type t_i of individual i under a strategy profile σ by

$$U(\sigma, t_i) := \int_{T_{-i}} u(F(\sigma_{-i}(t_{-i}), \sigma_i(t_i), \tau(t_i))\ d\beta(t_{-i} \mid t_i)\ ,$$

where, for any $v = (v_{-i}, v_i) \in V^n$, and any $t_i \in T$,

$$u(F(v_{-i}, v_i), \tau(t_i)) := \tau(t_i) Q(v_{-i}, v_i) - P_i(v_{-i}, v_i)\ .$$

Given a direct mechanism for social choice function F and a type space, truth-telling is a strategy profile σ so that $\sigma = \tau$. We say that a social choice function is implementable on a given type

[28] Smith (2010) defines feasibility so that (4) must hold with equality. We allow for free disposal.

[29] For robust coalition-proof implementation, formally defined below, one can show that this is without loss of generality, i.e., with this solution concept, a social choice function F is implementable if and only if it is implementable as a truth-telling equilibrium of a "direct" mechanism, i.e. of a mechanism so that the individuals' actions sets are equal to V. The requirement of coalition-proofness complicates the proof of the revelation principle if attention is restricted to a narrow set of type spaces, see Boylan (1998) and Bierbrauer (2014). Adding the requirement of robustness, however, makes it possible to prove a revelation principle. For further discussion, see Bierbrauer and Hellwig (2011).

space if truth-telling is an interim Nash equilibrium,[30] i.e. if for all i, all $t_i \in T$, and all $\hat{v}_i \in V$

$$U(\tau, t_i) \quad \geq \quad \int\limits_{T_{-i}} u(F(\tau_{-i}(t_{-i}, \hat{v}_i), \tau(t_i))\ d\beta(t_{-i} \mid t_i)\ . \tag{5}$$

Robust Implementability. Given a set T of types and a function τ, a social choice function F is said to be robustly implementable if (5) holds on every type space $[T, \tau, \beta]$. The following lemma, that we state without proof, is due to Ledyard (1978) and Bergemann and Morris (2005): Robust implementability is equivalent to the requirement that truthtelling is a dominant strategy equilibrium.[31]

Lemma 1 *A social choice function F is robustly implementable if and only if it is payoff-type dominant strategy incentive-compatible: for all i, all v_i, all v_i', and all v_{-i},*

$$v_i Q(v_{-i}, v_i) - P_i(v_{-i}, v_i) \geq v_i Q(v_{-i}, v_i') - P_i(v_{-i}, v_i')\ . \tag{6}$$

The next lemma provides a complete characterization of robustly implementable social choice functions. Again, this result is well known, see e.g. Moulin (1999) or Börgers (2015), so that we omit a proof. According to the Lemma, if an individual is not pivotal for public goods-provision, then her payment cannot depend on her type. If she is pivotal, then there is a payment she has to make whenever she announces a sufficiently high type so that the public good is provided, and another payment that is relevant if she announces a low type and the public good is not provided.

Lemma 2 *A social choice function F is robustly implementable if and only if it has the following properties: For every i, and every v_{-i}, there exists a cutoff type $c_i(v_{-i}) \geq 0$ so that*

$$Q(v_{-i}, v_i) = \begin{cases} 0, & \textit{if} \quad v_i < c_i(v_{-i})\ , \\ 1, & \textit{if} \quad v_i \geq c_i(v_{-i})\ . \end{cases} \tag{7}$$

If $0 < c_i(v_{-i}) \leq \bar{v}$, then there exist numbers $P_i^0(v_{-i})$ and $P_i^1(v_{-i})$ so that

$$P_i(v_{-i}, v_i) = \begin{cases} P_i^0(v_{-i}), & \textit{if} \quad v_i < c_i(v_{-i})\ , \\ P_i^1(v_{-i}), & \textit{if} \quad v_i \geq c_i(v_{-i})\ , \end{cases} \tag{8}$$

where

$$\inf\{v_i \in V \mid v_i > c_i(v_{-i})\} \geq P_i^1(v_{-i}) - P_i^0(v_{-i}) \geq \sup\{v_i \in V \mid v_i \leq c_i(v_{-i})\}\ . \tag{9}$$

If $c_i(v_{-i}) > \bar{v}$, or $c_i(v_{-i}) = 0$ then there exists a number $\bar{P}_i(v_{-i})$ so that, for all v_i,

$$P_i(v_{-i}, v_i) = \bar{P}_i(v_{-i})\ . \tag{10}$$

[30] We avoid the term "interim Bayes-Nash equilibrium" since our analysis does not require the existence of a common prior over the individuals' types.

[31] As mentioned above, despite the equivalence established in Lemma 1, we think of robust implementability as our key requirement so that the consideration of social choice functions that are individually incentive compatible in dominant strategies is only instrumental for a characterization of robust and coalition-proof social choice functions.

This result is important for the scope of coalition formation. Consider a type space where all individuals assign probability 1 to a specific payoff type profile v. Then individuals have multiple best responses. For instance, if $Q(v) = 0$, then every individual is willing to understate his preferences since this has neither an impact on the provision level, nor on the individual's payment. Likewise, if $Q(v) = 1$, then every individual is willing to exaggerate his preferences. In the following, we will show that these multiple best responses generate a degree of freedom for incentive-compatible coalition formation, i.e. there is scope for a misrepresentation of preferences which is in line with the requirement that every individual's behavior is a best response to the behavior of others.

Coalition-proof equilibrium. Fix a type space and consider the game that is induced by the direct mechanism for a social choice function F. We provide a recursive definition of coalition-proofness. We begin by defining an equilibrium that cannot be blocked by a coalition consisting of two individuals, i.e. by a coalition of minimal size. An interim Nash equilibrium σ^* is said to be blocked by a coalition of size 2 if there is a set of individuals $I' \subset I$ with $\#I' = 2$ and a deviation to a strategy profile $\sigma'_{I'} = (\sigma'_i)_{i \in I'}$ such that

i) The strategy $(\sigma^*_{I \setminus I'}, \sigma'_{I'})$ is an interim Nash equilibrium.

ii) For all $i \in I'$ and all t_i so that $\sigma'_i(t_i) \neq \sigma^*(t_i)$, $U((\sigma^*_{I \setminus I'}, \sigma'_{I'}), t_i) \geq U(\sigma^*, t_i)$. Moreover, there is at least one $i \in I'$ with a type t_i so that $U((\sigma^*_{I \setminus I'}, \sigma'_{I'}), t_i) > U_i(\sigma^*, t_i)$.

Our notion of blocking involves incentive and participation constraints. The latter require that the blocking coalitions makes no deviator worse off and at least one deviator strictly better off. The former require that the deviation induces a new equilibrium. I.e. all individuals – the deviators as well as the non-deviators – individually use best responses to the behavior of others.

Following Bernheim et al. (1986), coalitions will be considered effective only if they are not themselves vulnerable to deviations by sub-coalitions. To formalize this idea, we now define "blocking by a coalition of size n" and include the requirement that the outcome induced by this coalition must not be blocked by a coalition of size $n-1$. In this definition, the caveat that the coalition of size n must not be blocked by a coalition of size $n-1$ is itself weakened by the requirement, which is implicit in the recursive definition, that the potentially blocking coalition of size $n-1$ in turn must not be blocked by a coalition of size $n-2$, and so forth.

An interim Nash equilibrium σ^* is said to be *blocked by a coalition of size n* if there is a set of individuals $I' \subset I$ with $\#I' = n$ and a deviation to a strategy profile $\sigma'_{I'} = (\sigma'_i)_{i \in I'}$ being such that

i) The strategy $(\sigma^*_{I \setminus I'}, \sigma'_{I'})$ is an interim Nash equilibrium that is not blocked by a coalition $I'' \subset I'$ of size $n-1$.

ii) For all $i \in I'$ and all t_i so that $\sigma'_i(t_i) \neq \sigma^*(t_i)$, $U((\sigma^*_{I \setminus I'}, \sigma'_{I'}), t_i) \geq U(\sigma^*, t_i)$. Moreover, there is at least one $i \in I'$ with a type t_i so that $U((\sigma^*_{I \setminus I'}, \sigma'_{I'}), t_i) > U(\sigma^*, t_i)$.

An interim Nash equilibrium is said to be *coalition-proof* if there is no coalition of size $s \in \{2, \dots, n\}$ that blocks it.

Robust coalition-proof social choice functions. Given a set T of types and a function τ, a social choice function F is said to be robust and coalition-proof if a strategy profile τ with truth-telling by all agents is a coalition-proof interim Nash equilibrium on every type space $[T, \tau, \beta]$. The following lemma shows that this requirement of robustness and coalition-proofness implies a condition of ex post coalition-proofness: After the profile of types has become known, there must not exist a coalition of two individuals who regret that they did not block the truth-telling equilibrium. The following lemma states this requirement more formally.

Lemma 3 *Consider a robustly implementable social choice function so that, for some profile of payoff types v, and some subset I' of I with $\#I' = 2$, there exists a deviation $v'_{I'} \neq v_{I'}$ so that*

i) The deviators benefit. For all $i \in I'$,

$$v_i Q(v_{I \setminus I'}, v'_{I'}) - P_i(v_{I \setminus I'}, v'_{I'}) \geq v_i Q(v_{I \setminus I'}, v_{I'}) - P_i(v_{I \setminus I'}, v_{I'}) \ ,$$

with a strict inequality for at least some $i \in I'$.

ii) Each deviator gives a best response. For all $i \in I'$,

$$v_i Q(v_{I \setminus I'}, v'_{I'-i}, v'_i) - P_i(v_{I \setminus I'}, v'_{I'-i}, v'_i) \geq v_i Q(v_{I \setminus I'}, v'_{I'-i}, v_i) - P_i(v_{I \setminus I'}, v'_{I'-i}, v_i) \ .$$

Then the social choice function fails to be robustly coalition-proof.[32]

5 The main result

We now turn to the core of our analysis. We begin with a result showing that the payments of all individuals need to be equal. In addition, they only depend on whether the public good is provided or not. The logic of the argument is as follows: Suppose that there is v with $Q(v) = 0$. By Lemma 2, if we replace the type of individual i, v_i, by 0, this change affects neither the level of public-good provision nor the payment of agent i. The change also must not affect the payment of any other individual j: If it did, there would be type constellations at which i and j could form a coalition that would reduce the payment of agent j while leaving the payment of agent i as well as the level of public-good provision unchanged, i.e. the coalition-proofness condition in Lemma 3 would be violated. Hence, if we successively replace individual valuations by a valuation of zero, all payments remain constant. Ultimately we end up in a situation in which every individual's valuation is zero, and then, by anonymity, the individuals' payments have to be identical. But then, they must have been identical already in the initial situation v with $Q(v) = 0$. A symmetric argument, involving increases of valuations to the maximum $\bar{v}$, applies to type constellations with $Q(v) = 1$.[33]

[32] The reader may wonder why in (ii), we do not have a best-response condition for non-deviators. However, because of the equivalence of robust incentive compatibility and dominant-strategy incentive compatibility, this requirement is trivially satisfied.

[33] The arguments in the proof use only type spaces with a degenerate belief system and only coalitions of two individuals that are sub-coalition-proof simply because there are no sub-coalitions. Thus the proposition

Proposition 1 *An anonymous social choice function F is robustly implementable and robustly coalition-proof only if there exist numbers P^0 and P^1, so that for all v and all i,*

$$P_i(v) = \begin{cases} P^0 & \text{if} \quad Q(v) = 0 \ , \\ P^1 & \text{if} \quad Q(v) = 1 \ . \end{cases} \tag{11}$$

Given this proposition, we restrict our attention to social choice functions stipulating that individual payments are equal and depend only on whether the public good is provided or not. For such social choice functions, we find it convenient to write $F = (Q, P^0, P^1)$ rather than $F = (Q, P_1, \ldots, P_n)$. For ease of exposition, we also limit attention to type spaces where all belief systems have the same null sets. This implies that we do not have to worry about manipulations and submanipulations that affect outcomes at type constellations to which agents assign probability zero at some types and positive probability at others.

Definition 1 (Moderately uninformative belief system) *A belief system β is said to be moderately uninformative if the measures $\beta(t_i)$, $t_i \in T$ are mutually absolutely continuous, i.e., if they all have the same null sets.*

We say that such a belief system is moderately uninformative because learning one's type does not make it possible to rule out any event that would be considered possible with some other type. Given a set T of types and a function τ, a social choice function F is said to be UB-robust and coalition-proof if truth-telling is a coalition-proof interim Nash equilibrium on every type space $[T, \tau, \beta]$ with a moderately uninformative belief system. We are now prepared to state our main result.

Theorem 1 *Consider a social choice function $F = (Q, P^0, P^1)$ and suppose that $P^1 - P^0 \cap V = \emptyset$. Denote by $s_1(v) := \#\{i \mid v_i > P^1 - P^0\}$ the number of individuals who are net gainers from public-good provision, given a payoff profile v. This social choice function is UB-robustly implementable and robustly coalition-proof, if and only if for all v and v',*

$$s_1(v) \geq s_1(v') \quad \textit{implies} \quad Q(v) \geq Q(v'). \tag{12}$$

We refer to social choice functions that satisfy (12) as *voting mechanisms.*[34] This condition implies that it suffices to ask individuals to raise their hand if they are in favor of public-goods provision and to provide the public good if the number of hands in favor is sufficiently high. Thus, the theorem asserts that a social choice function is UB- robust and coalition-proof if and only if it is a voting mechanism.

would continue to hold if we replaced the requirement of robust coalition-proofness by the requirement of ex post coalition-proofness. It would also continue to hold if the scope for coalition formation was not restricted by requirements of sub-coalition-proofness.

[34] As mentioned in Footnote 2, this concept is narrower than the notion of "voting" in Gibbard (1973) and Satterthwaite (1975) and some of the subsequent literature.

To prove the “only if” part of the theorem, we do not need the full strength of the robust coalition-proofness requirement. Showing that every robust and coalition-proof social choice function is a voting mechanism requires only ex post coalition-proofness. Specifically, we show that if the monotonicity condition in (12) was violated then there would be a type space with a degenerate belief system on which either a coalition of individuals who benefit from public-goods provision or a coalition of harmed individuals blocks the truthful equilibrium.

Ex post coalition-proofness is a weaker condition than robust coalition-proofness since coalition-proofness on all type spaces with a degenerate belief system is weaker than coalition-proofness on all type spaces. Therefore the “if-part” of the theorem – which establishes that voting mechanisms are robustly coalition-proof – also implies that voting mechanisms are ex post coalition-proof.

It follows that the equivalence property established in Theorem 1 also holds if the requirement of robust coalition-proofness is replaced by a requirement of ex post coalition-proofness, i.e., a social choice function is UB-robustly implementable and ex post coalition-proof if and only if it can be implemented by a voting mechanism.

As mentioned in the introduction, Moulin (1980) showed that, in models with single-peaked preferences over a linearly ordered set of alternatives, the median-voter mechanism, i.e., majority voting is group strategy-proof as well as individually strategy-proof, or, in the Bayesian terminology, ex post coalition-proof as well as robustly incentive compatible. Our analysis extends this result to allow for arbitrary voting mechanisms. Moreover, we also allow for nondegenerate belief systems. For this purpose however, i.e., to show that every voting mechanism is robustly implementable and robustly coalition-proof, we need the weaker form of coalition-proofness that results from considering only coalitions that are themselves immune against blocking by subcoalitions. The reason is that we also have to deal with belief systems under which a joint deviation by supporters and opponents of public-goods provision might appear attractive.[35] We show that even if such belief systems existed, no such deviation would be subcoalition-proof. Given the monotonicity condition in (12), if the supporters take it for granted that the opponents deviate, they have an incentive to withdraw their contribution to the joint deviation and report their types truthfully.

Our analysis is based on the assumption that individuals cannot use side payments to facilitate coalition formation. In principle, this is a restriction because side-payments can enlarge the set of coalitions that might upset the implementation of a social choice function. This enlargement would not affect the finding that *only* social functions that can be implemented by voting mechanisms are coalition-proof as well as robustly implementable. It seems likely however that allowing side payments in coalition formation would upset the finding that implementability by a voting mechanism is *sufficient* for the coalition-proofness of a social choice function.[36] How-

[35] If we restricted attention to type spaces with degenerate belief systems, or equivalently, to ex post coalition-proofness, this issue would not arise. Given the monotonicity in (12), there is no deviation that is attractive both to supporters and to opponents of public-good provision.

[36] In this context, it is worth noting that Smith (2010) finds that dominant-strategy outcomes of a voting mechanism can be improved upon by an alternative mechanism in which participants are given the option to volunteer to pay for a larger share of the cost of public-good provision. Smith’s result depends on an assumption that, in our notation, $\bar{v} > kn$, i.e. if an agent’s valuation is large enough, he would be willing to pay for the public

ever, we conjecture that, as the number of participants becomes large, there is a sense in which the effectiveness of side payments in facilitating coalition formation becomes small. If there are many participants, small coalitions are not likely to affect the level of public-good provision, and large coalitions are likely to be affected by internal free-rider problems.[37] In a large coalition, any one individual would sense that the own side payment would not have much of an effect on the aggregate outcome and would therefore refrain from any significant payment towards coalition formation.[38] For models with a continuum of agents, we actually show in Bierbrauer and Hellwig (2015) that, under the additional assumption that coalition design itself must be robust, it is not possible to use side payments to enlarge the set of coalitions that block a given social choice function.

Limits to First-Best Implementation. We now turn to the welfare implications of imposing coalition-proofness, as well as robust implementability. The possibility that first-best implementation may run afoul of coalition-proofness can be illustrated by the example in Section 3 with possible valuations 0, 3, and 10, and a per-capita provision cost equal to 4. First-best implemenation requires budget balance and since, by Proposition 1, all individuals make the same payment there has to be equal-cost-sharing. Thus, first-best implementation requires that every individual pays 0 if the public good is not provided and pays 4 otherwise. Hence, individuals with valuations 0 and 3 oppose public-goods provision, whereas individuals with a valuation of 10 benefit from public-goods provision.

In this example, a first-best provision rule requires that the public good be provided if $3S_3 + 10S_{10} > 4n$ and not be provided if $3S_3 + 10S_{10} < 4n$. Now, suppose that $n = 10$, and fix the number of individuals with valuation 10 so that $S_{10} = 3$. Then first-best public goods provision requires that the public is provided if and only if at least 4 of the remaining seven individuals have a valuation of 3. By Theorem 1, however, the public goods provision level has to be the same irrespectively of whether the number of people with valuation 3 is above or below 4. More generally, we obtain:

Corollary 1 *If there is a pair of preference profiles $v = (v_1, \dots, v_n)$ and $v' = (v'_1, \dots, v'_n)$, such that $s_1(v) \geq s_1(v')$ and $\frac{1}{n}\sum_{i=1}^n v_i < k < \frac{1}{n}\sum_{i=1}^n v'_i$, then there is no social choice function that yields first best outcomes and is UB-robust and coalition-proof.*

good on his own.

[37] In Bayesian models with independent private values free-rider problems become ever more serious as the number of participants increases; see Mailath and Postlewaite (1990), and Hellwig (2003). Neeman (2004) shows that this result extends to belief systems that do not satisfy the "beliefs-determine-preferences (BDP)"-property. Belief systems with the BDP-property, by contrast, allow for mechanisms à la Crémer and McLean (1985, 1988) that make it possible to overcome free-riding. As we discussed in Section 3, these mechanism are, however, not robust to the specification of the belief system.

[38] Note that we assume that individuals are willing to lie about their preferences only if lying is costless, i.e. only if it is, individually, a best response to everybody else's behavior. A side payment to other coalition members, by contrast, is costly.

Second-Best Considerations. If first-best is out of reach, the mechanism designer is faced with a second-best problem. Given the impossibility of achieving efficient outcomes for every preference profile v, he must choose between different deviations from efficiency that are compatible with robustness and coalition-proofness. For instance, in the above example, he can decide whether it is better to forego the net benefits from public-good provision whenever $S_3 \geq 4$ or to incur the net losses from public-good provision whenever $S_3 < 4$. 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. Thus in our example, he might ask for a payment $P^0 = 1.1$ if the public good is not provided, rather than $P^0 = 0$ in order to turn people with valuations 3 from opponents into supporters 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 whenever the public good is not provided.[39]

6 Extensions

In the following, we discuss several extensions of our analysis. Specifically, we discuss the conditions under which our analysis extends to a setting with more than just two possible provision levels for the public good.

The Equal Payments Property. For the case of a public good that comes as a single indivisible unit Proposition 1 shows that under robustness and coalition-proofness, the payment scheme becomes extremely simple. There is a payment P^1 that every individual has to make whenever the public good is provided and another payment P^0 that is relevant if the public good is not provided. This made it possible to identify the two relevant groups for an analysis of coalition-proof outcomes, the group of individuals who benefit from public-goods provision and the group of individuals who are harmed by public-goods provision.

In the following we will generalize our analysis and assume that there is a finite set $\mathcal{Q}$ of possible provision levels. We will also limit attention to payment rules so that for every provision level $q \in \mathcal{Q}$, there is a payment P^q that all individuals have to make whenever provision level q is implemented. The following Lemma provides a justification for this approach. It follows from

[39] In this paper, we refrain from a characterization of second-best mechanisms that trade off these different kinds of inefficiencies. Such a characterization would require additional assumptions on the mechanism designer's beliefs and objective function. For instance, if we assume that the mechanism designer has his own prior beliefs and seeks to maximize the expected surplus from public-goods provision, then his problem looks as follows: Choose P^0, P^1 and $Q : V^n \to \{0,1\}$ so as to maximize the expected aggregate surplus

$$E^M \left[\left(\frac{1}{n} \sum_{i=1^n} v_i - P_F^1 \right) Q_F(v) - P_F^0 (1 - Q_F(v)) \right] \tag{13}$$

subject to the feasibility constraints $P^0 \geq 0$, and $P^1 \geq k$, and the coalition-proofness condition that for every pair v and v', $s_1(v) \geq s_1(v')$ implies $Q(v) \geq Q(v')$. The expectations operator E^M in (13) indicates that expectations over $v = (v_1, \ldots, v_n)$ are taken with respect to the mechanism designer's subjective beliefs. Smith (2010) develops a belief-free welfare criterion that rests on whether the outcomes of a given social choice function can be weakly improved at all type vectors and strictly improved at some type vectors.

a straightforward generalization of the arguments in the proof of Proposition 1. We therefore state it without proof.

Lemma 4 *Consider a social choice function $F = (Q, P_1, \dots, P_n)$ with $Q : V^n \to \mathcal{Q}$ and $P_i : V^n \to \mathbb{R}$, for all i. Suppose that to every $q \in Q(V^n)$ there is $v' \in V$ so that*

i) $v_i = v'$, for all i, implies $Q(v) = q$.

ii) If $Q(v_{-i}, v_i) = q$, then $Q(v_{-i}, v') = q$.

If this social choice function is robust and coalition-proof, then for every $q \in Q(V^n)$ there exists a number P^q so that $Q(v) = q$ implies $P_i(v) = P^q$, for all i.

The lemma presumes that, for every provision level q, there is an anchor payoff type so that if all individuals are of the anchor type, then q is chosen. Moreover, if q is chosen in a situation in which the payoff type of any one individual i is different from the anchor type, then q is also chosen if i's type is changed into the anchor type.

With just two provision levels, this presumption is actually implied by the requirement of dominant strategy incentive compatibility. For instance, it follows from Lemma 2 that if $Q(v) = 0$, then $Q(v_{-i}, 0) = 0$ and also $Q(0, \dots, 0)$. Hence, 0 is the anchor type for a provision level of 0. Likewise $\bar{v}$ is the anchor type for a provision level of 1. Lemma 4 generalizes the observation in Proposition 1. If there is an anchor type for each provision level, then the payment rule must involve equal payments for all individuals.

In Bierbrauer and Hellwig (2015), we discuss an alternative justification for focusing on social choice functions with equal payments. Following Hammond (1979) and Guesnerie (1995), we use a large-economy approach with a continuum of agents to formalize the notion that each individual is too small to influence the level of public good provision and show that, in this setting, the condition that payments are the same for different agents and depend only on outcomes follows from anonymity and robustness.

Multiple Provision Levels with a Linear Cost Function. Suppose that the per capita cost associated with a provision level of $q \in \mathcal{Q}$ equals $k\ q$. In the following, we will show that our main result extends to this setting if, in addition to robustness and coalition-proofness, we insist on ex post budget balance. A social choice function with equal payments satisfies ex post budget balance if $Q(v) = q$ implies that $P_i(v) = k\ q$, for all i. The utility that an individual with payoff type v_i realizes in state v is therefore equal to $(v_i - k)Q(v)$. We can now define the sets $V_0 := \{v \mid v < k\}$ and $V_1 := \{v \mid v > k\}$. The set V_0 contains all individuals with a payoff type below the marginal cost of provision. When comparing two arbitrary provision levels q and $q' > q$, all individuals in this set prefer the smaller one, q. By contrast, individuals with payoff types in the set V_1 prefer the larger one. Our main result extends to this setup. The following proposition is a straightforward adaptation of Theorem 1, so that we can omit a proof.

Proposition 2 *Suppose that per capita resource requirement associated with a provision level of q equals $k\ q$. Consider a social choice function $F = (Q, P_1, \ldots, P_n)$ so that, for all i and v, $P_i(v) = k\ Q(v)$. This social choice function is robust and coalition-proof if and only if*

$$s_1(v) \geq s_1(v') \quad \textit{implies} \quad Q(v) \geq Q_F(v') \; .$$

Monotonicity in Population Shares. We will now provide a generalization of Theorem 1 that allows for an arbitrary set of provision levels and a nonlinear cost function. We begin with an alternative characterization of dominant strategy incentive compatibility. To this end, it will prove convenient to write the set of possible payoff types V as finite ordered set $V = \{v^1, v^2 \ldots, v^m\}$. Let S^k be the number of individuals with payoff type v^k, and let $S = (S^1, \ldots, S^m)$. Let $\mathcal{S}$ be the set of possible values of S. Let $\bar{P}(S)$ be the individuals' payment in state S. The requirement of dominant strategy incentive compatibility can then be written as: For all k and all $S \in \mathcal{S}$ with $S^k > 0$,

$$v^k Q(S) - \bar{P}(S) \geq v^k Q(\hat{S}) - \bar{P}(\hat{S}) \; , \tag{14}$$

for all $\hat{S} \in \mathcal{S}^{k-1}(S)$, where

$$\mathcal{S}^{k-1}(S) = \{\hat{S} \in \mathcal{S} \mid \hat{S}^k = S^k - 1, \; \hat{S}^l = S^l + 1, \text{ for some } l \neq k, \text{ and } \hat{S}^j = S^j \text{ for all } j \neq k, l\} \; .$$

A repeated application of the inequalities in (14) then yields the following: For all k and all $S \in \mathcal{S}$ with $S^k > 0$,

$$v^k Q(S) - \bar{P}(S) \geq v^k Q(\hat{S}) - \bar{P}(\hat{S}) \; , \tag{15}$$

for all $\hat{S} \in \mathcal{S}^{-k}(S)$, where

$$\mathcal{S}^{-k}(S) = \{\hat{S} \in \mathcal{S} \mid \hat{S}^k \leq S^k, \text{ and } \hat{S}^l \geq S^l, \text{ for all } l \neq k, \} \; .$$

Thus, equal payments and dominant strategy incentive compatibility imply a monotonicity in population shares: An individual with payoff type v^k prefers the outcome for situation S over the outcome for $\hat{S}$ whenever one can move from S to $\hat{S}$ by successively increasing the number of individuals with a payoff different from v^k at the expense of the number of individuals with a payoff type equal to v^k.

Monotonicity in population shares in turn implies coalition-proofness. To see this, note first that there is no type space so that a coalition consisting entirely of individuals with the same payoff type has a profitable deviation. Any such deviation would require that some of these individuals lie about their payoff type. By (15) this can only make these individuals worse off. If we limit attention to moderately uninformative belief systems, then there are also no deviations that are supported by individuals of several payoff types. If individuals with payoff types v^k and v^l deviate jointly, and if all these individuals expect to benefit from that deviation at the ex interim stage, then, given the monotonicity in population shares, individuals with a payoff type v^k have an incentive to withdraw their contribution to the collective deviation and report their types truthfully, so that the joint deviation by types v^k and v^l is not subcoalition-proof.[40] We summarize these observations in the following proposition.

Proposition 3 *A social choice function with equal payments is UB-robustly implementable and robustly coalition-proof if and only if it is monotonic in population shares, i.e. if and only if it satisfies (15).*

The large-economy analysis in Bierbrauer and Hellwig (2015) establishes a related result for a model with a continuum of agents, a continuum of possible valuations and an arbitrary finite number of provision levels, with a weaker concept of coalition-proofness, under which collective deviations cannot be conditioned on beliefs, in particular, the complete-information beliefs that played a key role in the proof of Theorem 1. In that setting, a monotonic social choice function is shown to be robustly implementable and immune to robustly blocking coalitions if and only if it can be implemented by a voting mechanism. For an important class of social choice functions, the voting mechanism takes the form of a sequence binary votes on whether to raise the level of public-good provision by another unit. Thus in a setting with three provision levels, 0,1,2, and a convex cost function, participants would be asked to vote on whether they prefer $Q = 1$ over $Q = 0$ and on whether they prefer $Q = 2$ over $Q = 1$. The chosen provision levels are $Q = 0$ if the population share of the set of people preferring $Q = 1$ over $Q = 0$ falls short of a given threshold and $Q = 2$ if the population share of people in favor of $Q = 2$ exceeds another given threshold. In all other cases, the chosen provision level is $Q = 1$. Again only ordinal information on preferences is used. Intensities of preferences must not affect the decision on public-goods provision.

7 Concluding Remarks

Our main subject in this paper has been the problem of mechanism design for public-good provision in an economy with prior uncertainty as to whether it is efficient for the public good to be provided or not. If there are no participation constraints, a social choice function that yields surplus-maximizing outcomes can be implemented as a dominant strategy equilibrium of a properly designed mechanism. We show that, in some instances, however, these equilibria are implausible because they rely on information that (collectively) hurts the people who provide it. We impose a requirement of coalition-proofness to eliminate this possibility.

When coalition-proofness is imposed along with robust incentive compatibility, the implementability of a social choice function that yields surplus-maximizing outcomes can no longer be taken for granted. Social choice functions are robustly implementable and 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

[40] Note that, again by the monotonicity in population shares, there is no sub-sub-coalition that could possibly benefit from deviating from the truthful reports specified for the sub-coalition.

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 surplus-maximizing outcomes, because they are not responsive to the preference intensities of those who benefit and those who are harmed by public-good provision.

The notion of robust incentive compatibility that we use, which goes back to Ledyard (1978) and Bergemann and Morris (2005), has been criticized by Smith (2010), and Börgers and Smith (2014) as being too restrictive. In their view, the ignorance of the mechanism designer about the participants' beliefs requires an incentive mechanism that is defined independently of those beliefs. It does not imply, however, that the actions the participants choose under a given mechanism must be independent of their beliefs. On the basis of this observation, Smith (2010) and Börgers and Smith (2014) provide examples of situations in which it is possible to Pareto-improve, at the ex interim stage, on the outcomes induced by a belief-independent equilibrium. This raises the question whether our theorem would remain true under their weaker notion of "undominatedness" of a social choice function. To make progress on this question, it will probably be necessary to first obtain a characterization of what "undominatedness" implies in the context of the public-good provision problem.

References

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

Austen-Smith, D. and Banks, J. (1996). Information Aggregation, Rationality and the Condorcet Jury Theorem. *American Political Science Review*, 90:34–45.

Barberà, S. (1979). A Note on Group Strategy-Proof Decision Schemes. *Econometrica*, 47:637–640.

Bennett, E. and Conn, D. (1977). The Group Incentive Properties of Mechanisms for the Provision of Public Goods. *Public Choice*, 29: 95-102.

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.

Bierbrauer, F. (2009). Optimal Income Taxation and Public-Good Provision with Endogenous Interest Groups. *Journal of Public Economic Theory*, 11:311–342.

Bierbrauer, F. (2012). Distortionary Taxation and the Free-Rider Problem. *International Tax and Public Finance*, 19:732–752.

Bierbrauer, F. (2014). Optimal Tax and Expenditure Policy with Aggregate Uncertainty. *American Economic Journal: Microeconomics*, 6:205–257.

Bierbrauer, F. and Hellwig, M. (2011). Mechanism Design and Voting for Public-Good Provision. Preprint 2011/31, Max Planck Institute for Research on Collective Goods.

Bierbrauer, F. and Hellwig, M. (2015). Public-Good Provision in Large Economies. Working Paper, Max Planck Institute for Research on Collective Goods.

Buchanan, J. and Tullock, G. (1962). The Calculus of Consent. Univerysity of Michigan Press, Ann Arbor.

Boylan, R. (1998). Coalition-Proof Implementation. *Journal of Economic Theory*, 82:132–143.

Börgers, T. (2015). An Introduction to the Theory of Mechanism Design. Oxford University Press.

Börgers, T. and Smith, D. (2014). Robust Mechanism Design and Dominant Strategy Voting Rules. *Theoretical Economics*, 9: 339–360.

Casella, A. (2005). Storable Votes. *Games and Economic Behavior*, 51: 391–419.

Che, Y. and Kim, J. (2006). Robustly Collusion-Proof Implementation. *Econometrica*, 74:1063–1107.

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.

Crémer, J. (1996). Manipulation by Coalition Under Asymmetric Information: The Case of Groves Mechanisms. *Games and Economic Behavior*, 13:39–73.

Dasgupta, P., Hammond, P., and Maskin, E. (1979). The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility. *The Review of Economic Studies*, 46:185–216.

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica*, 41:587–601.

Gizatulina, A. and Hellwig, M. (2015). The Genericity of the McAfee-Reny Condition for Full Surplus Extraction in Models with a Continuum of Types. Mimeo, Max Planck Institute for Research on Collective Goods.

Goeree, J. and Zhang, J. (2013). *Electoral Engineering: One Man, One Bid.* Discussion Paper, University of Zurich.

Green, J. and Laffont, J.-J. (1977). Characterization of Satisfactory Mechanisms for the Revelation of Preferences for Public Goods. *Econometrica*, 45: 472-487.

Green, J. and Laffont, J.-J. (1979). On Coalition Incentive Compatibility. *Review of Economic Studies*, 46: 243-254.

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

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

Hammond, P. (1979). Straightforward Individual Incentive Compatibility in Large Economies. *Review of Economic Studies*, 46:263–282.

Hellwig, M. (2003). Public-good Provision with Many Participants. *Review of Economic Studies*, 70:589–614.

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 Information Bargaining Problems with Many Agents. *Review of Economic Studies*, 57:351–367.

McAfee, P. and Reny, P. (1992). Correlated Information and Mechanism Design. *Econometrica*, 60: 395–421.

Mirrlees, J. (1971). An Exploration in the Theory of Optimum Income Taxation. *Review of Economic Studies*, 38:175–208.

Moulin, H. (1980). On Strategy-Proofness and Single Peakedness. *Public Choice*, 35:437–455.

Moulin, H. (1999). Incremental Cost Sharing: Characterization by Coalition Strategy-Proofness. *Social Choice and Welfare*, 16:175–208.

Neeman, Z. (2004). The Relevance of Private Information in Mechanism Design. *Journal of Economic Theory*, 117:55–77.

Peleg, B. and Sudhölter, P. (1999). Single-peakedness and coalition-proofness. *Review of Economic Design*, 4: 381–387.

Satterthwaite, M. (1975). Strategy-Proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217.

Smith, D. (2010). A Prior Free Efficiency Comparison of Mechanisms for the Public Good Problem. Mimeo, University of Michigan.

A Proofs

Proof of Proposition 1

We begin by showing that a property that we will term *neutrality* is implied by the requirements of robustness and coalition-proofness. Under this property, the influence of a player i on the payment of any other player j is limited in a particular way. Specifically, neutrality requires that a change in individual i's payoff type, which is inconsequential both for the decision on public-goods provision and for i's payments, does not affect the payment of individual j.

Lemma 5 *If an anonymous social choice function is robustly implementable and robustly coalition-proof, then it is neutral, i.e. for any pair of individuals i and j, any v_{-i-j}, and pair v_i and v_i' and any v_j: $Q(v_{-i-j}, v_i, v_j) = Q(v_{-i-j}, v_i', v_j)$ and $P_i(v_{-i-j}, v_i, v_j) = P_i(v_{-i-j}, v_i', v_j)$ imply $P_j(v_{-i-j}, v_i, v_j) = P_j(v_{-i-j}, v_i', v_j)$.*

Proof Suppose to the contrary that there are individuals i and j, v_{-i-j}, a pair v_i and v_i' and v_j so that $Q(v_{-i-j}, v_i, v_j) = Q(v_{-i-j}, v_i', v_j)$, $P_i(v_{-i-j}, v_i, v_j) = P_i(v_{-i-j}, v_i', v_j)$, but $P_j(v_{-i-j}, v_i, v_j) \neq P_j(v_{-i-j}, v_i', v_j)$. We show that this implies a violation of coalition-proofness.

Suppose the true payoff type profile equals (v_{-i-j}, v_i, v_j). Consider a deviation from truth-telling by individuals i and j, and suppose they report instead (v_i', v_j). By dominant strategy incentive compatibility, individual j, who still reports truthfully, is giving a best response. Individual i's deviation neither affects the decision on public-goods provision, nor his payment. Hence, the deviation yields the same payoff as truthtelling, and therefore is a best response. Since individual i' s payoff is unaffected he is willing to participate in the deviation. Coalition-proofness, therefore requires that individual j is not made strictly better off by the deviation, which requires that $P_j(v_{-i-j}, v_i, v_j) \leq P_j(v_{-i-j}, v_i', v_j)$. Since we hypothesized that $P_j(v_{-i-j}, v_i, v_j) \neq P_j(v_{-i-j}, v_i', v_j)$, it must be the case that $P_j(v_{-i-j}, v_i, v_j) < P_j(v_{-i-j}, v_i', v_j)$. Then, however, if the true payoff type profile equals (v_{-i-j}, v_i', v_j), a deviation by i and j to (v_i, v_j) is such that both are giving a best response, individual i is willing to participate and j is made strictly better off. Hence, a contradiction to coalition-proofness. ■

Lemma 6 *Let F be anonymous, robustly implementable and robustly coalition-proof. Then it has the following properties:*

i) For every pair of individuals i and j, and for every v with $Q(v) = 0$,

$$Q(v_{-i-j}, 0, 0) = 0 \quad \textit{and} \quad P_i(v_{-i-j}, 0, 0) = P_i(v_{-i-j}, v_i, v_j)\ . \tag{16}$$

ii) For every pair of individuals i and j, and 1 for every v with $Q(v) = 1$,

$$Q(v_{-i-j}, \bar{v}, \bar{v}) = 1 \quad \textit{and} \quad P_i(v_{-i-j}, \bar{v}, \bar{v}) = P_i(v_{-i-j}, v_i, v_j). \tag{17}$$

iii) For every pair of individuals i and j, and for every v, $P_i(v_{-i-j}, v_i, v_j) = P_j(v_{-i-j}, v_i, v_j)$.

Proof We only prove part i). The proof of part ii) is analogous. By anonymity, part iii) follows from i) and ii). Let F be anonymous, robustly implementable and robustly coalition-proof. Fix some v so that $Q(v) = 0$ and suppose that two individuals $I' = \{i, j\}$ jointly announce $(0, 0)$ instead of (v_i, v_j). Since $Q(v) = 0$, Lemma 2 implies

$$Q(v_{-i-j}, v_i, v_j) = Q(v_{-i-j}, 0, v_j) = Q(v_{-i-j}, v_i, 0) = Q(v_{-i-j}, 0, 0) = 0 \ ,$$

$P_i(v_{-i-j}, v_i, v_j) = P_i(v_{-i-j}, 0, v_j)$, and $P_j(v_{-i-j}, 0, v_j) = P_j(v_{-i-j}, 0, 0)$. Neutrality then implies $P_i(v_{-i-j}, 0, v_j) = P_i(v_{-i-j}, 0, 0)$. Hence, $P_i(v_{-i-j}, v_i, v_j) = P_i(v_{-i-j}, 0, 0)$. ■

Corollary 2 *If an anonymous social choice function is robustly implementable and robustly coalition-proof, then there exist numbers P^0 and P^1 so that, for all v, and all i,*

$$P_i(v) = \begin{cases} P^0 & \text{if} \quad Q(v) = 0 \ , \\ P^1 & \text{if} \quad Q(v) = 1 \ . \end{cases} \tag{18}$$

Proof We only show that there is a number P^0 so that, for all i, $P_i(v) = P^0$, whenever $Q(v) = 0$. By Lemma 6, it follows that, for all v, all individuals make the same payments, i.e., $P_i(v) = P_j(v) := \bar{P}(v)$, for any pair (i, j). From part i) it follows that, starting from an arbitrary v, with $Q(v) = 0$, if we successively replace valuations v_i, that are possibly different from 0, by 0, the decision on provision and the payment $\bar{P}$ remain unaffected. Hence, if v is such that $Q(v) = Q(0, \ldots, 0) = 0$, then also $\bar{P}(v) = \bar{P}(0, \ldots, 0)$. ■

Proof of Theorem 1

Necessity of the Conditions of Theorem 1. We seek to show the following: Given a social choice function $F = (Q, P^1, P^0)$ and given that $P^1 - P^0 \cap V = \emptyset$, F is robustly implementable and robustly coalition-proof only if, for all v and v' in V^n, $s_1(v) \geq s_1(v')$ implies $Q(v) \geq Q(v')$. Suppose that this condition fails to hold, i.e., that $Q(v) = 0$ and $Q(v') = 1$ for some v and v' such that $s_1(v) \geq s_1(v')$. Our proof has three steps: We first show that, under a direct mechanism for this social choice function, there is a type space with a moderately uninformative belief system so that a coalition has a profitable deviation. As a second step we show that the profitable deviation leads to an interim Nash equilibrium. And finally, as a third step, we argue that this deviation does not provoke further deviations by sub-coalitions.

Step 1. To focus on the essentials of the argument, suppose first that, in fact, $s_1(v) = s_1(v')$. We claim that, for some moderately uninformative belief system, there is a coalition that has a profitable deviation.

Consider the coalition of people with payoff types in $V_0 = \{v_i \in V \mid v_i < P^1 - P^0\}$ and denote $s_0(v) := \#\{i \mid v_i \in V_0\}$. All these people prefer the outcome for the state v to the outcome for the state v'. Thus, if the belief system is degenerate and all beliefs assign all probability mass to a type vector t so that $\tau(t) = v'$, a coalition of people with payoff types in V_0 would have a profitable deviation if they could find a vector of reports $\hat{v}_0 = (\hat{v}_i)_{\{i|v_i \in V_0\}}$ such that $Q(\hat{v}_0, v_1') = 0$, where v_1' is the restriction of the vector v' to $\{i \mid v_i \in V_1\}$, and $V_1 := \{v_i \in V \mid v_i > P^1 - P^0\}$.

For instance, since $Q(v) = 0$, this coalition would prefer to deviate from truth-telling if there was $\hat{v}_0$ so that $(\hat{v}_0, v_1') = v$. Because this coalition does not control v_1', it may not be able to do so. Hence suppose that $(\hat{v}_0, v_1') = v''$ and that $v'' \neq v$. If the coalition of people with payoff types in V_0 is to be prevented from having a profitable deviation if beliefs assign all probability mass to a type profile t with $\tau(t) = v'$, it must be the case that $Q(v'') = 1$.

But now consider the coalition of people with types in V_1. All these people prefer the outcome for v' and, therefore, also the outcome for v'' to the outcome for v. Thus, if the belief system is degenerate and beliefs assign all probability mass to t so that $\tau(t) = v$, a coalition of people with types in V_1 has a profitable deviation if there exists a profile of reports $\hat{v}_1$ so that

$$(v_0, \hat{v}_1) = v'' ,$$

which yields the outcome $Q(v'') = 1$. Because, by the definition of v'', the coalition of people with payoff types in V_1 has the same size under v'' as under v and v' and, moreover, the restrictions of v'' and v to the set V_0 are the same, such a deviation is in fact available.

Thus, if there are two preference profiles v and v' with $s_1(v) = s_1(v')$, under a direct mechanism for a social choice function prescribing $Q(v) = 0$ and $Q(v') = 1$ there is a coalition that possesses a profitable deviation. Either it is a coalition of people with payoff types in V_0 when beliefs put all probability mass on v', or a coalition of people with payoff types in V_1 when beliefs put all probability mass on v.

In the preceding argument, the assumption that $s_1(v) = s_1(v')$, is not really needed. A little reflection shows that, if $s_1(v) \geq s_1(v')$, the coalitions in the preceding argument have even more scope for finding collective deviations so as to generate reports equal to v''.

Step 2. We now show that the collective deviations considered in Step 1 are themselves immune to deviations by sub-coalitions of size one, i.e. that no individual has an incentive to deviate from the behavior stipulated by the coalition. Suppose that individual i belongs to a coalition I' of individuals with payoff types in V_1 that changes the outcome from $Q(v) = 0$ to $Q(v_{I \setminus I'}, \hat{v}_{I'}) = 1$. By the requirement of robust incentive compatibility, truth-telling is individually a best response for all possible reports of the individuals different from i. Thus, we have to verify that, for any such individual i, $v_i Q(v_{I\setminus I'}, \hat{v}_{I'-i}, \hat{v}_i) - P^1$ is as large as the utility realized by i under the payoff type profile $(v_{I\setminus I'}, \hat{v}_{I'-i}, v_i)$. If $Q(v_{I\setminus I'}, \hat{v}_{I'-i}, v_i) = 1$, then this is obviously true. If, by contrast, $Q(v_{I\setminus I'}, \hat{v}_{I'-i}, v_i) = 0$, then truth-telling yields a payoff of $-P^0$. Since, by assumption, $v_i \in V_1$, this is less than the payoff of $v_1 - P^1$ that is realized when reporting $\hat{v}_i$, a contradiction to the assumption that the social choice function is implementable in dominant strategies.

Step 3. We argue that the coalitions that are constructed in Step 1 are themselves sub-coalition-proof. The members of these coalitions have the same beliefs and the same preferences over the two outcomes, $Q = 0$ with payments of P^0, or $Q = 1$ with payments of P^1. Hence, a coalition that benefits from inducing the outcome $Q = 0$ on a type space with a degenerate belief system is sub-coalition-proof: While a sub-coalition might be able to turn the outcome back into $Q = 1$, this would make all members of the sub-coalition worse off.

Sufficiency of the Conditions of Theorem 1. Consider a social choice function $F = (Q, P^0, P^1)$ which satisfies (12). We consider a direct mechanism for this social choice function and show that truth-telling is a coalition-proof equilibrium for every belief system that is moderately uninformative. The truthful strategy is in the following denoted by τ.

Our proof proceeds by contradiction. Hence, suppose that there exists a type space $[T, \tau, \beta]$ with a moderately uninformative belief system such that the equilibrium σ on $[T, \tau, \beta]$ is blocked by a collective deviation $\sigma'_{I'}$. We denote by $\hat{\tau}(\sigma'_{I'}, t)$ the reports that result if the vector of types is $t = (t_1, \ldots, t_n)$, individuals in I' behave according to $\sigma'_{I'}$ and individuals in $I \setminus I'$ behave according to τ. We also denote by $s_0(\tau(t))$ and by $s_0(\hat{\tau}(\sigma'_{I'}, t))$, respectively, the corresponding numbers of individuals who report a payoff type that belongs to V_0.

We may assume that the coalition consists both of individuals with payoff types in V_1 and of individuals with payoff types in V_0. Given the monotonicity property in (12), no other coalition could conceivably block the equilibrium τ. We may also assume that there exist $D_0 \in T^{n-1}$ and $D_1 \in T^{n-1}$ such that, for all $t_i \in T$, the following two conditions hold:

1. $\beta(D_0|t_i) > 0$ and, moreover,

$$t_{-i} \in D_0 \text{ implies } Q(\tau(t)) = 1 \text{ and } Q(\hat{\tau}(\sigma'_{I'}, t)) = 0. \tag{19}$$

2. $\beta(D_1|t) > 0$, for all $t_i \in T$, and, moreover,

$$t_{-i} \in D_1 \text{ implies } Q(\tau(t)) = 0 \text{ and } Q(\hat{\tau}(\sigma'_{I'}, t)) = 1. \tag{20}$$

If conditions (19) and (20) were both violated, then the manipulation would affect the final allocation with probability zero. Hence, a necessary condition for blocking would be violated. If, say, only condition (19) was fulfilled, but condition (20) was violated, then the manipulation would make individuals with payoff types in V_1 worse off. Again, this would violate a necessary condition for blocking.

Now consider a sub-coalition I'' with $I'' = \{i \in I' \mid \tau(t_i) \in V_0\}$ which recommends that the opponents of public-goods provision sabotage the manipulation $\sigma'_{I'}$ by reporting truthfully to the overall mechanism. Consequently, for any type vector t, the mass of individuals who report $\tau(t_i) \in V_0$ is bounded from below by $s_0(\tau(t))$ and by $s_0(\hat{\tau}(\sigma'_{I'}, t))$. Because of (12) this implies, for any individual $i \in I''$,

$$Q(\hat{\tau}(\sigma'_{I' \setminus I''}, t)) = 0 \text{ ,whenever } t_{-i} \in D_1. \tag{21}$$

and

$$Q(\hat{\tau}(\sigma'_{I' \setminus I''}, t)) = 0 \text{ ,whenever } t_{-i} \in D_0. \tag{22}$$

Hence, in all states in which the manipulation $\sigma'_{I'}$ changed the outcome from non-provision into provision, the submanipulation $\sigma''_{I''} = \tau_{I''}$ is undoing this change. In states in which the manipulation $\sigma'_{I'}$ changed the outcome from provision into non-provision, the submanipulation is not undoing this change. Clearly, this makes all individuals with payoff types in V_0 better off. Moreover, because of the monotonicity property in (12), the strategy profile $(\tau_{I''}, \sigma'_{I' \setminus I''}, \tau_{I \setminus I'})$ induced by sub-coalition I'' is not blocked by a subset of I''. Thus, the assumption that $\sigma'_{I'}$ blocks the equilibrium τ has led to a contradiction and must be false. ■