Informed-principal problems in environments with generalized private values

Tymofiy Mylovanov
Department of Economics, University of Pennsylvania

Thomas Tröger
Department of Economics, University of Mannheim

We provide a solution to the problem of mechanism selection by a privately informed principal in generalized-private-value environments. In a broad class of these environments, the mechanism-selection game has a perfect-Bayesian equilibrium that has a strong neologism-proofness property. Equilibrium allocations that satisfy this property are characterized in terms of the players' incentive and participation constraints, and can be computed using standard methods.

Keywords. Informed principal, mechanism design, private values, strong unconstrained Pareto optimum.

JEL classification. D82, D86.

1. Introduction

In many applications of mechanism design, the principal has private information that is not directly payoff-relevant to the agents, but may influence her design: a seller's opportunity cost influences the design of her sales mechanism, a supplier's valuation influences her design of a collusive agreement, a speculator's prior beliefs, in environments with heterogeneous priors, influence the design of the bet she offers, and the weight a regulator puts on consumer surplus influences the design of her regulation scheme. In this paper, we solve the problem of mechanism selection by an informed principal in such "generalized-private-value" environments where the agents' payoff functions are independent of the principal's type.

Mechanism selection by an informed principal differs fundamentally from mechanism design by a principal with no private information. The latter can be formulated, due to the revelation principle, as a maximization problem of the principal's payoff function subject to the agents' incentive and participation constraints. If, however, the principal has private information, then upon observing a mechanism proposal, the agents

Tymofiy Mylovanov: mylovanov@gmail.com
Thomas Tröger: troeger@uni-mannheim.de

We would like to thank the co-editor Jeff Ely and three anonymous referees for very useful comments that helped improve the paper. Financial Support by the German Science Foundation (DFG) through SFB/TR 15 "Governance and the Efficiency of Economic Systems" and by the National Science Foundation through Grant 0922365 is gratefully acknowledged.

DOI: 10.3982/TE787

may update their beliefs about the principal's type. Hence, the proposal of a mechanism by an informed principal must be viewed as a move in a game and the maximization approach is not applicable (Myerson 1983).

We focus on the perfect-Bayesian equilibria of the mechanism-selection game that have a strong "neologism-proofness" property. The idea of neologism-proofness was introduced by Farrell (1993).[1] Given an equilibrium, a set of types A is self-signaling if types in this set gain by inducing the belief that the type is in A. Farrell argues that a statement that a type belongs to a self-signaling set is credible and should be believed.

Applying this idea to the mechanism-selection game, we require that any observed deviation from an equilibrium mechanism should be accompanied by a "credible" belief. We call a belief credible if none of the principal-types who would suffer from the deviation is believed to make it, and those types who already enjoy the highest feasible payoff are also believed to not make the deviation. Our concept of credibility differs from Farrell's (1993) classical definition because we do not require that *all* the types who gain from the deviation are believed to make it, and because we require that *none* of the types who already enjoy the highest feasible payoff makes the deviation. An equilibrium such that no credible and profitable deviation exists is called strongly neologism-proof.

Our main result is that a strongly neologism-proof equilibrium exists in a broad class of environments with generalized private values.[2] This is in stark contrast to standard signaling games, where neologism-proof equilibria often do not exist, and weaker equilibrium concepts, such as those based on the intuitive criterion (Cho and Kreps 1987), are considered (Riley 2001).[3]

We allow the outcome space to be any compact metric space. Any finite number of players and types is permitted, as is any continuous payoff function (no single-crossing or any other structural property is required). We assume, however, that private information is stochastically independent across players. Apart from that, we only make one technical assumption ("separability"): there exists an allocation such that all agents' participation and incentive constraints are satisfied with strict inequality. Separability is typically easy to verify in a given environment. For environments with a single agent, we show that separability is satisfied in any environment in which the agent's private information is payoff-relevant for herself.

The strongly neologism-proof equilibrium allocations are attractive because they do not rely on implausible out-of-equilibrium beliefs about the principal. In addition, the strongly neologism-proof equilibrium allocations can be characterized in terms of the players' participation and incentive constraints. This is important because it implies that these allocations can be found via standard mechanism-design methods. By contrast, the existing literature provides no characterization of perfect-Bayesian equilibria of the informed-principal game in general private-value environments.

[1] Ideas similar to Farrell's (1993) were used by Grossman and Perry (1986) to motivate their concept of perfect sequential equilibrium.

[2] In an environment with verifiable types, de Clippel and Minelli (2004) show the existence of an equilibrium that satisfies a related notion of neologism-proofness.

[3] See also Mailath et al. (1993), who argue in favor of another solution concept—undefeated equilibrium—that is also weaker than strong neologism-proofness.

Myerson (1983) and Maskin and Tirole (1990) were the first to consider the problem of mechanism selection by an informed principal. Myerson's elegant analysis is based on an axiomatic approach and applies to environments with a finite outcome space. Maskin and Tirole consider a class of single-agent environments with two possible types of the agent under particular structural assumptions about the outcome space and the players' payoff functions.[4] Part of our contribution can be seen as extending ideas of Maskin and Tirole to a much more general setting. Maskin and Tirole's analysis evolves around the concept of a "strong unconstrained Pareto optimum" (SUPO), which characterizes the entire set of perfect-Bayesian equilibrium allocations under some assumptions. Maskin and Tirole (cf. 1990, footnote 23 and, also, footnote 7 in this paper) state that SUPO can be recast in terms of Farrell's neologism-proofness. In our more general setting, it is more convenient to put a neologism-proofness property in the center right away, rather than trying to generalize the concept of a SUPO. Also, a straight generalization of SUPO would lead to existence problems. Our approach yields the existence of a strongly neologism-proof equilibrium even when no SUPO exists (cf. footnote 9).

The basic reason why a privately informed principal's mechanism may differ from the mechanism that she would offer if her information were public is simple. A privately informed principal may propose a mechanism that is independent of her private information, while she herself is a player in her mechanism. In such a mechanism, the agents' incentive and participation constraints must hold only on average over the principal's types, given the agents' belief. The principal may be able to gain from such a weakening of the constraints.

The crucial implication of the assumption of generalized private values is that the form of the agents' incentive and participation constraints is independent of the principal's type. As observed by Maskin and Tirole (1990), this makes it possible to interpret the different types of the principal as traders in a fictitious economy where each constraint corresponds to a good. The principal-types trade amounts of slack allowed for the various constraints. Any competitive equilibrium in this fictitious economy corresponds to an allocation that is a strongly neologism-proof equilibrium allocation of the mechanism-selection game.

Technically, our main contribution is the result that a competitive equilibrium exists for the fictitious economy. To guarantee existence, we include the possibility that some goods have the price 0 and we allow free disposal (that is, in equilibrium, any constraint may be satisfied with strict inequality on average over the principal's types). A further complication arises from the fact that the traders' "utility functions" in the fictitious economy are determined endogenously in a way that their continuity cannot be guaranteed. Finally, Walras' law may fail. Consequently, while our existence proof builds on Debreu's (1959) classical arguments, some details are substantially different. By comparison, in Maskin and Tirole's setting, it is sufficient to ignore trade in all but two constraints that have positive prices, traders' utility functions in the fictitious economy are differentiable, and Walras' law holds, so that Debreu's arguments extend straightforwardly.

[4] Quesada (2010) provides conditions for equilibrium allocations in Maskin and Tirole (1990) to be deterministic and shows that their characterization continues to hold in a less restrictive environment.

The main result in our paper is a general existence result. By contrast, Maskin and Tirole (1990), Fleckinger (2007), Cella (2008), and Skreta (2011) focus on whether the privacy of the principal's information allows the principal to improve her payoff. (For examples of an environment with private values in which the principal can benefit from the privacy of her information, see Section 4.) Severinov (2008) obtains the full-surplus extraction result for environments with the informed principal and correlated types.

A few papers consider standard private values environments with continuous type spaces and quasilinear preferences. Yilankaya (1999) considers a standard bilateral-trade environment à la Myerson and Satterthwaite (1983), with the seller being the principal. It is shown that in this environment, the privacy of the principal's information does not matter. A similar result is obtained by Tan (1996) for a procurement setting with multiple agents, and by Mylovanov and Tröger (2008) for extensions of Myerson's (1981) optimal-auction environments and for the quasilinear versions of the principal–agent environments of Guesnerie and Laffont (1984) in which the principal is privately informed.

In this paper, we analyze environments with generalized private values. We do not know how to extend our approach to the environments with common or interdependent values. The difficulty is that the market clearing condition (4) in the definition of the competitive equilibrium is not independent from the allocation of the slack consumption among different types of the principal if the agents' payoffs depend on the principal's type.

Informed-principal problems in common- and interdependent-value environments are considered in Myerson (1983), Maskin and Tirole (1992), Severinov (2008), Skreta (2011), and Balkenborg and Makris (2010). Maskin and Tirole (1992, Proposition 7) provide necessary and sufficient conditions for the existence of neologism-proof equilibria under the assumption that the agent has no private information.

Finally, there exists a separate literature that studies the informed-principal problem in moral-hazard environments, rather than in adverse-selection environments considered here (see, for example, Beaudry 1994, Chade and Silvers 2002, and Kaya 2010).

The rest of the paper is organized as follows. Section 2 describes the model. In Section 3, we characterize strong neologism-proofness in terms of incentive and participation constraints. In Section 4, we give examples. Section 5 deals with the existence of strongly neologism-proof equilibria. Section 6 relates to other solution concepts. Section 7 concludes. Some proofs are given in the Appendix.

2. Model

We consider the interaction of a principal (player 0) and n agents (players $i \in N = \{1, \dots, n\}$). The players must collectively choose an outcome from a compact metric space of *basic outcomes* Z.[5] Every player $i = 0, \dots, n$ has a *type* t_i that belongs to a finite *type space* T_i. A *type profile* is any $\mathbf{t} \in \mathbf{T} = T_0 \times \cdots \times T_n$. Sometimes we use the notation

[5] Hence, in environments with monetary transfers, the set of feasible transfers is truncated at some (arbitrarily high) point.

$\mathbf{T} = T_i \times \mathbf{T}_{-i}$, $\mathbf{t} = (t_i, \mathbf{t}_{-i})$, or $\mathbf{t} = (t_0, t_i, \mathbf{t}_{-0,i})$. Player i's payoff function,

$$u_i : Z \times \mathbf{T} \to \mathbb{R},$$

is assumed to be continuous (note that the continuity assumption is void if Z is finite).

We focus on environments that are characterized by the property that the agents' payoff functions are independent of the principal's type.

Definition 1. An environment features *generalized private values* if, for all $i \neq 0$,

$$u_i(z, (t_0, \mathbf{t}_{-0})) = u_i(z, (t_0', \mathbf{t}_{-0})) \quad \text{for all } z, t_0, t_0', \mathbf{t}_{-0}.$$

An *outcome* is a probability measure over basic outcomes; let $\mathcal{Z}$ denote the set of outcomes. We identify any $z \in Z$ with the point distribution that puts probability 1 on the point z; hence, $Z \subseteq \mathcal{Z}$. We extend the definition of u_i to $\mathcal{Z} \times \mathbf{T}$ via the statistical expectation.

Some outcome $z_0 \in \mathcal{Z}$ is designated as the *disagreement outcome*. Every player's payoff from the disagreement outcome is normalized to 0 (for every profile of other players' types).

The interaction is described by the following *mechanism-selection game*. First, for each player i, nature chooses a type. Let $p_i(t_i) > 0$ denote the probability of type $t_i \in T_i$. We assume that types are stochastically independent. Each player privately observes her type t_i. Second, the principal offers a *mechanism* M, which is a finite perfect-recall game form with players $N \cup \{0\}$ and with outcomes in $\mathcal{Z}$. Third, the agents decide simultaneously whether to accept M. If M is accepted unanimously, then each player chooses a message (consisting of an action at each of her information sets) in M and the outcome specified by M is implemented. If at least one agent rejects M, then the disagreement outcome z_0 is implemented.[6]

An *allocation* is a function

$$\rho : \mathbf{T} \to \mathcal{Z}$$

that assigns an outcome $\rho(\mathbf{t})$ to every type profile $\mathbf{t} \in \mathbf{T}$. Thus, an allocation describes the outcome of the mechanism-selection game as a function of the type profile. Alternatively, an allocation ρ can be interpreted as a *direct mechanism*, where the players $i = 0, \ldots, n$ simultaneously announce types $\hat{t}_i$ (= messages) and the outcome $\rho(\hat{t}_0, \ldots, \hat{t}_n)$ is implemented. The definition of the mechanism-selection game includes the possibility of proposing *in*direct mechanisms as well.

[6] Myerson (1983) and Maskin and Tirole (1990, 1992) define similar mechanism-selection games. Both earlier models, however, restrict attention to simultaneous-move mechanisms, and Maskin and Tirole assume that players can use a public randomization device to decide which equilibrium to play in M. Myerson assumes that an agent's decision to accept or reject is taken simultaneously with the choice of a message in M (other private actions beyond acceptance and rejection may also be allowed). The equilibrium allocations that we find are also equilibrium allocations in the games of Myerson (1983) and Maskin and Tirole (1990, 1992). By contrast, the results in this paper do not apply if private actions can be taken *sequentially* as occurs, for example, in a model in which a rejection of the principal's proposal is followed by a play of a status quo mechanism.

A *perfect-Bayesian equilibrium* for the mechanism-selection game specifies (i) for each type of the principal, an optimal (possibly randomized) mechanism proposal, (ii) for each mechanism, a belief about the principal's type that is computed via Bayes rule if the mechanism is proposed by at least one type, and (iii) for each mechanism, a strategy profile that is a sequential equilibrium in the continuation game that follows the proposal of the mechanism.

A well known drawback of the perfect-Bayesian equilibrium concept is that the belief about the principal's type remains unrestricted if the principal proposes an "off-path" mechanism that no type was expected to propose. This may, in principle, allow for rather implausible equilibria. Hence, rather than attempting to find all equilibria, we focus on the existence and characterization of equilibria that have an additional property called *strong neologism-proofness*.[7] A similar approach was pioneered by Farrell (1993) in the context of standard signaling games (related ideas were put forward by Grossman and Perry 1986).

3. Strongly neologism-proof allocations

In this section, we define strongly neologism-proof allocations and show that any such allocation is a perfect-Bayesian equilibrium outcome. Hence, strong neologism-proofness is a refinement of perfect-Bayesian equilibrium.

Additional notation is required. Consider the continuation game that begins after some arbitrary mechanism is proposed. Let $q_0 : T_0 \to [0, 1]$ denote the probability distribution that describes the agents' belief about the principal's type at the beginning of the continuation game. Let ρ denote the allocation resulting from the acceptance or rejection decisions and the subsequent play of the mechanism. The expected payoff of type t_i of player i if she follows the rejection or acceptance and message choice of type $\hat{t}_i$ is

$$U_i^{\rho,q_0}(\hat{t}_i, t_i) = \sum_{\mathbf{t}_{-i}\in\mathbf{T}_{-i}} u_i(\rho(\hat{t}_i, \mathbf{t}_{-i}), (t_i, \mathbf{t}_{-i}))\mathbf{q}_{-i}(\mathbf{t}_{-i}),$$

where $\mathbf{q}_{-i}(\mathbf{t}_{-i}) = q_0(t_0) \cdot p_1(t_1) \cdots p_{i-1}(t_{i-1}) \cdot p_{i+1}(t_{i+1}) \cdots p_n(t_n)$ if $i \neq 0$ and $\mathbf{q}_{-0}(\mathbf{t}_{-0}) = p_1(t_1) \cdot \cdots \cdot p_n(t_n)$.

The expected payoff of type t_i of player i from allocation ρ is

$$U_i^{\rho,q_0}(t_i) = U_i^{\rho,q_0}(t_i, t_i).$$

We use the shortcut $U_0^{\rho}(t_0) = U_0^{\rho,q_0}(t_0)$, which is justified by the fact that the principal's expected payoff is independent of q_0.

Definition 2. An allocation ρ is called q_0-*feasible* if, given the belief q_0 and using the direct-mechanism interpretation of ρ, no type of any player has an incentive to reject ρ

[7] Extrapolating a result of Maskin and Tirole (1990, Proposition 7) and further examples, one may conjecture that *all* perfect-Bayesian equilibria are strongly neologism-proof, but showing this generally appears to be beyond reach.

or to deviate from announcing her true type: for all i,

$$U_i^{\rho,q_0}(t_i) \geq U_i^{\rho,q_0}(\hat{t}_i, t_i) \quad \text{for all } t_i, \hat{t}_i \tag{1}$$

$$U_i^{\rho,q_0}(t_i) \geq 0 \quad \text{for all } t_i. \tag{2}$$

By the revelation principle, any perfect-Bayesian equilibrium allocation of the mechanism-selection game is p_0-feasible. Hence, as observed by Myerson (1983), without loss of generality, we may restrict attention to perfect-Bayesian equilibria in which all types of the principal offer the same p_0-feasible allocation as a direct mechanism ("principle of inscrutability"). However, as far as continuation equilibria following off-path mechanism proposals are concerned, we cannot restrict attention to p_0-feasible allocations, but have to consider q_0-feasible allocations for arbitrary beliefs q_0. In general, off-path beliefs q_0 have to be different from the prior belief p_0 so as to make deviating mechanisms unattractive for all types of the principal.[8]

Given any allocations ρ and ρ', the set of principal-types that are strictly better off in ρ is denoted

$$S(\rho, \rho') = \{t_0 \in T_0 \mid U_0^{\rho}(t_0) > U_0^{\rho'}(t_0)\}.$$

The set of types who in ρ obtain the highest feasible payoff is denoted

$$H(\rho) = \left\{ t_0 \in T_0 \,\middle|\, U_0^{\rho}(t_0) = \sum_{\mathbf{t}_{-0} \in \mathbf{T}_{-0}} \max_{z \in Z} u_0(z, t_0, \mathbf{t}_{-0}) \mathbf{q}(\mathbf{t}_{-0}) \right\}.$$

Given any allocations ρ and ρ', we say that a belief q_0 about the principal's type is *credible* for ρ' relative to ρ if it is consistent with Bayesian updating given the following behavior: none of the principal-types who is strictly better off in ρ than in ρ' or who already enjoys the highest feasible payoff in ρ, chooses ρ', that is,

$$\forall t_0 \in S(\rho, \rho') \cup H(\rho) : q_0(t_0) = 0.$$

Let $\text{Supp}(q_0)$ denote the support of q_0.

Definition 3. An allocation ρ is called *strongly neologism-proof* if ρ is p_0-feasible and there exists *no* belief q_0 together with a q_0-feasible allocation ρ' such that (i) q_0 is credible for ρ' relative to ρ and (ii) $S(\rho', \rho) \cap \text{Supp}(q_0) \neq \varnothing$.

Note that the credibility of a belief q_0 does *not* reflect any requirement that the types who are strictly better off in ρ' than in ρ would actually choose ρ'. This aspect of our concept of credibility is more general than Farrell's original definition; it reflects that

[8] An instructive example is provided by Yilankaya (1999). He considers a standard bilateral-trade environment à la Myerson and Satterthwaite (1983), with the seller being the principal. A perfect-Bayesian equilibrium allocation is constructed from optimal fixed-price offers by all types of the seller. Some types would gain by deviating to a double-auction mechanism if the buyer kept her prior belief. If, however, the buyer believes that the lowest cost seller proposes the double auction, this deviation becomes unprofitable for all types of the seller.

we do not want to exclude the possibility that some types, although preferring ρ' over ρ, may not be believed to be among the deviators (possibly because there may exist another deviation that is more attractive than ρ').

A second novel aspect of our concept of credibility is our requirement that no type who already enjoys the highest feasible payoff in ρ may be attracted to ρ'. This restriction is necessary for our general existence result (see Example 1 and the proof of Lemma 3 below). But in many environments, including all environments where sufficiently large monetary transfers are feasible, the restriction is not binding:

Remark 1. Suppose that the outcome space allows such large monetary transfers between the players that in any allocation that specifies the largest feasible transfer to the principal, at least one agent's participation constraint is violated. Then $H(\rho) = \varnothing$ for any p_0-feasible allocation ρ.

This remark applies to most environments considered in the earlier literature. Hence, in all these environments, our concept of neologism-proofness is more demanding than Farrell's (which strengthens our existence result).

It remains to show that strongly neologism-proof allocations actually are perfect-Bayesian equilibrium allocations.

Proposition 1. *Any strongly neologism-proof allocation is a perfect-Bayesian equilibrium allocation of the mechanism-selection game.*

Proof. Consider any strongly neologism-proof allocation ρ. We construct a perfect-Bayesian equilibrium of the mechanism-selection game as follows. All types of the principal propose the direct mechanism ρ. If ρ is proposed, all agents accept and all players announce their true types.

It remains to construct, for any mechanism $M \neq \rho$, the agents' belief q^M about the principal's type and the strategy profile τ^M for the continuation game that begins when M is proposed.

Fix some $M \neq \rho$ and consider the following game $\mathcal{G}(M)$:

> First, nature chooses privately observed types $t_0, \dots, t_n$ exactly as in the mechanism-selection game. Second, if $t_0 \in H(\rho)$, then the game ends (we may specify arbitrary payoffs in this case). However, if $t_0 \notin H(\rho)$, then the principal chooses between two actions. One action ends the game and she obtains the payoff $U_0^{\rho}(t_0)$ (we may specify arbitrary payoffs for the agents in this case); the other action is to offer the mechanism M. Third, the agents decide simultaneously whether to accept M. If M is accepted unanimously, then each player chooses a message in M and the outcome specified by M is implemented. If at least one agent rejects M, then the disagreement outcome z_0 is implemented.

This is a finite game with perfect recall and thus has a sequential equilibrium σ^M. Define q^M as any belief about the principal at the beginning of the third stage of the game $\mathcal{G}(M)$ that is consistent with σ^M. Define τ^M as the strategy profile induced by σ^M in the continuation game that begins at the third stage of $\mathcal{G}(M)$.

Let ρ^M denote the allocation induced by τ^M in the continuation game that begins at the third stage of $\mathcal{G}(M)$. It remains to show that in ρ^M, no type of the principal is strictly better off than in ρ, i.e.,

$$S(\rho^M, \rho) = \varnothing.$$

Suppose the opposite. Then the belief q^M is credible for ρ^M relative to ρ, contradicting the fact that ρ is strongly neologism-proof. □

Due to Proposition 1, the strongly neologism-proof allocations correspond precisely to the strongly neologism-proof perfect-Bayesian equilibrium allocations.

Because the strongly neologism-proof allocations are defined in terms of the players' incentive and participation constraints, in a particular application, the strongly neologism-proof perfect-Bayesian equilibrium allocations can be computed without making any reference to the mechanism-selection game. Rather, it is sufficient to use the relevant incentive and participation constraints (as incorporated in the definition of a strongly neologism-proof allocation), which brings the informed-principal problem back into the realm of standard mechanism-design methods.

To find all strongly neologism-proof allocations, one can proceed as follows. A principal's utility vector specifies a utility for each type of the principal. A principal's utility vector is q_0-feasible if it arises from some q_0-feasible allocation. For any belief q_0, find the q_0-Pareto frontier, which is defined as the set of q_0-feasible principal's utility vectors that are not Pareto dominated, from the viewpoint of the principal-types in the support of q_0, by any q_0-feasible principal's utility vector. The q_0-Pareto frontiers can be found by maximizing weighted sums of principal-types utilities.

In environments where, on the p_0-Pareto frontier, no type obtains the highest feasible payoff, the strongly neologism-proof utility vectors are the points on the p_0-Pareto frontier that are not below any of the q_0-Pareto frontiers. If, however, some types on the p_0-Pareto frontier do obtain the highest feasible payoff ("happy types"), then the strongly neologism-proof utility vectors are the p_0-feasible points that are not below any of the q_0-Pareto frontiers where q_0 puts probability 0 on the happy types.

4. Examples

In this section we provide examples of strongly neologism-proof equilibrium allocations.

Example 1. Suppose that the principal's type space is $T_0 = \{0, 1\}$, in which both types are equally likely. There is only one agent, who has no private information. The space of basic outcomes is the unit interval $Z = [0, 1]$. The players have single-peaked preferences $u_0(z, t_0) = -(z - t_0)^2$ and $u_1(z) = -z^2$. (Hence, the agent's preferences are aligned with type 0 of the principal.) The disagreement outcome is $z_0 = 1/2$.

In this example, any deterministic allocation ρ such that $\rho(0) = 0$ and $z_0 \leq \rho(1) \leq 1/\sqrt{2}$ is strongly neologism-proof. In such an allocation, type 0 obtains her (and the agent's) most preferred outcome, and type 1's outcome is between the disagreement outcome and the closest outcome to 1 that makes ρ acceptable to the agent given the prior belief about the principal. Here, $H(\rho) = \{0\}$. Allocation ρ is strongly neologism-proof, because there is no q_0-feasible allocation ρ' such that q_0 puts probability 1 on type $t_0 = 1$ and such that type 1 is better off in ρ' than in ρ.

The example reveals the crucial role of credibility. If the agent's belief q_0 puts a higher than prior probability on type $t_0 = 0$, then there exists a q_0-feasible allocation ρ' that is more favorable to type 1 than any of the strongly neologism-proof allocations ρ described above. Say $q_0(1) = 1/4$. Then the deterministic allocation ρ' given by $\rho'(0) = 0$ and $\rho'(1) = 1$ is q_0-feasible and type 1 is strictly better off than in ρ.[9] But the belief q_0 is not credible for ρ' relative to ρ. ♢

Example 2. A principal and a single agent would like to dissolve a partnership, as in Cramton et al. (1987). Each owns 50% (= one share). Let $y \in [-1, 1]$ denote the amount of shares transferred from the principal to the agent and let $p \in [-\underline{p}, \underline{p}]$ ($\underline{p}$ is large) denote the payment from the agent to the principal. The parties' preferences are expressed by linear payoff functions over basic outcomes (y, p):

$$u_0(y, p, t_0) = p - yt_0, \qquad u_1(y, p, t_1) = yt_1 - p,$$

where the types $t_0 \in T_0 = \{0, 3\}$ and $t_1 \in T_1 = \{1, 2\}$ are the parties' marginal valuations of the shares. Both agent types are equally likely. The disagreement outcome is "no trade," that is, $z_0 = (0, 0)$.

Consider first the belief q_0 that puts probability 1 on type $t_0 = 0$. The highest possible expected payoff of type $t_0 = 0$ in any q_0-feasible allocation is the solution value to the problem

$$\max_{\rho} U_0^{\rho}(t_0) \quad \text{subject to} \quad (1), (2), q_0(t_0) = 1,$$

which is equal to 1. Hence, to be strongly neologism-proof, an allocation must at least give the expected payoff 1 to type $t_0 = 0$. Similarly, a strongly neologism-proof allocation must give at least the expected payoff 1 to type $t_0 = 3$. We indicate these requirements, respectively, by a vertical line and a horizontal line in Figure 1.

Now let $q_0 = 1/2$ and consider the auxiliary problem

$$(*) \quad \max_{\rho} U_0^{\rho}(0) + U_0^{\rho}(3) \quad \text{subject to} \quad (1), (2).$$

This is a linear problem that can be solved using standard methods. There is a continuum of solutions

$$\rho^*(t_0, t_1) = \begin{cases} (1, p^*) & \text{if } t_0 = 0 \\ (-1, -p^*) & \text{if } t_0 = 3 \end{cases}$$

for each $p^* \in [1, 2]$. In any solution, the entire ownership is assigned to the agent if the type of the principal is low and to the principal otherwise. The seller is compensated by a

[9] This argument also shows that in this example, there exists no strong unconstrained Pareto optimum in the sense of Maskin and Tirole (1990). In particular, none of the perfect-Bayesian equilibria of the mechanism-selection game is a SUPO in this example.

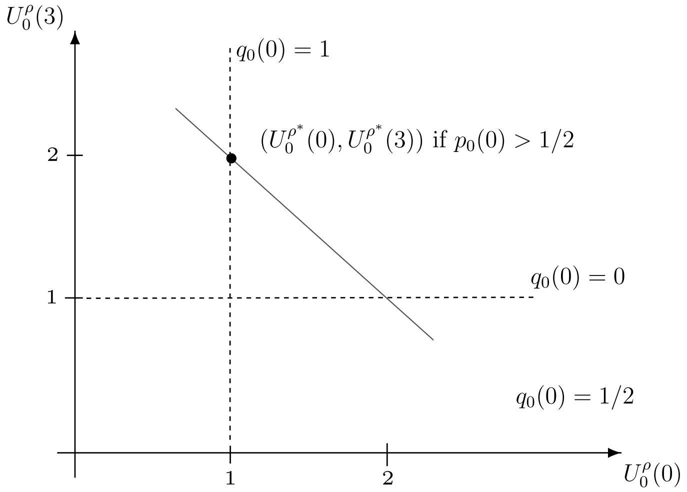

FIGURE 1. Strongly neologism-proof allocations.

payment of p^*. The principal's expected-payoff pairs corresponding to various solutions of problem $(*)$ are indicated in Figure 1.

The set of the principal's payoffs $(U_0^{\rho^*}(0), U_0^{\rho^*}(3))$ that is spanned by $p^* \in [1,2]$ is given by

$$U^* = \{(u, 3-u) \mid 1 \le u \le 2\}.$$

We now argue that any strongly neologism-proof allocation must attain the payoffs in U^*. Let ρ' be an allocation such that

$$U_0^{\rho'}(0) + U_0^{\rho'}(3) < 3, \quad U_0^{\rho'}(0), U_0^{\rho'}(3) \ge 1.$$

Then there exists a pair of payoffs in $(u', u'') \in U^*$ such that $u' > U_0^{\rho'}(0)$, $u'' > U_0^{\rho'}(3)$. By construction, these payoffs are achieved by an allocation that is feasible for $q_0 = 1/2$. Furthermore, $q_0 = 1/2$ is credible relative to ρ'. Hence, ρ' cannot be strongly neologism-proof.

Finally, it can be shown, using standard linear programming methods, that for any q_0, there is no q_0-feasible allocation that attains payoffs in

$$U^{**} = \{(u', u'') \mid u', u'' \ge 1, u' + u'' > 3\}.$$

Hence, a p_0-feasible allocation that attains payoffs in U^* is strongly neologism-proof.

We conclude that for $p_0 = 1/2$, any allocation attaining payoffs in U^* is strongly neologism-proof. For other beliefs, the strongly neologism-proof allocation is given by

$$\rho^*(t_0, t_1) = \begin{cases} (1,1) & \text{if } t_0 = 0,\ p_0(0) > 1/2 \\ (-1,-1) & \text{if } t_0 = 3,\ p_0(0) > 1/2 \\ (1,2) & \text{if } t_0 = 0,\ p_0(0) < 1/2 \\ (-1,-2) & \text{if } t_0 = 3,\ p_0(0) < 1/2. \end{cases}$$

◊

5. Existence of strongly neologism-proof allocations

This section is devoted to our main result, concerning the existence of strongly neologism-proof equilibria. To establish existence, we need one additional property.

Definition 4. An environment is called *separable* if there exists an allocation ρ such that, for any q_0,[10] *all* incentive and participation constraints (1) and (2) are satisfied with strict inequality for all agents i.

For many standard environments, separability can be easily verified.

In environments with a single agent, the following result provides a complete characterization of separability. Roughly speaking, separability requires that (i) the agent's information is payoff-relevant for herself, and that (ii) there is a possibility of an outcome that the agent strictly prefers over disagreement. These conditions are clearly necessary for separability.[11] The nontrivial part is to show sufficiency; a proof can be found in the Appendix.

Remark 2. An environment with a single agent ($n = 1$) is separable if and only if (i) any two types of the agent have nonidentical preferences over Z and (ii) for every type of the agent, there exists at least one outcome such that the agent's payoff is strictly positive.

Below is our main result, which establishes the existence of a strongly neologism-proof allocation and thus provides a solution to the informed-principal problem.

Proposition 2. *A strongly neologism-proof allocation exists in any separable environment with generalized private values.*

The key idea is to obtain existence indirectly by (i) defining a fictitious exchange economy where the different types of the principal trade amounts of slack granted to the incentive and participation constraints of the agents, (ii) establishing a connection between the competitive equilibria in the fictitious exchange economy and strongly neologism-proof allocations, and (iii) establishing existence of a competitive equilibrium in the fictitious exchange economy.

The fictitious economy

We begin by defining the "goods" for the fictitious economy. Let

$$\mathcal{G} = \bigcup_{i \neq 0} \{i\} \times \big(\{(\hat{t}_i, t_i) \mid \hat{t}_i, t_i \in T_i, \hat{t}_i \neq t_i\} \cup T_i\big),$$

where any $(i, \hat{t}_i, t_i) \in \mathcal{G}$ parameterizes an incentive constraint and any $(i, t_i) \in \mathcal{G}$ parameterizes a participation constraint. Any real-valued function c on $\mathcal{G}$ may be interpreted

[10] Due to generalized private values, it is irrelevant which q_0 is used here.

[11] With multiple agents, separability implies additional restrictions. For example, there must be an outcome that is strictly preferred to the disagreement outcome by all agents.

as a “bundle of goods.” In particular, a bundle may include a negative amount of any good. Observe that the assumption of finite type spaces for the agents implies that the number of goods is finite, which greatly simplifies the analysis.

Each principal-type $t_0 \in T_0$ is a “trader.” A trader “consumes” a bundle c by maximizing her expected payoff, given that the (positive or negative) slacks in the agents' constraints are described by the function c. Hence, the “utility” that any trader $t_0 \in T_0$ derives from any consumption bundle $c : \mathcal{G} \to \mathbb{R}$ is the solution value $V(t_0, c)$ of the problem

$$J(t_0, c): \qquad \max_{\rho:\mathbf{T}\to\mathcal{Z}} \sum_{\mathbf{t}_{-0}} u_0(\rho(\mathbf{t}), \mathbf{t})\mathbf{q}_{-0}(\mathbf{t}_{-0})$$

$$\text{subject to } \sum_{\mathbf{t}_{-0,i}} u_i(\rho(\mathbf{t}), \mathbf{t})\mathbf{q}_{-0,i}(\mathbf{t}_{-0,i}) \geq -c(i, t_i) \quad \text{for all } (i, t_i) \in \mathcal{G}$$

$$\sum_{\mathbf{t}_{-0,i}} \big(u_i(\rho(\mathbf{t}), \mathbf{t}) - u_i(\rho(\hat{t}_i, \mathbf{t}_{-i}), \mathbf{t})\big)\mathbf{q}_{-0,i}(\mathbf{t}_{-0,i}) \geq -c(i, \hat{t}_i, t_i)$$

$$\text{for all } (i, \hat{t}_i, t_i) \in \mathcal{G},$$

where $\mathbf{q}_{-0,i}(\mathbf{t}_{-0,i}) = p_1(t_1) \cdots p_{i-1}(t_{i-1}) \cdot p_{i+1}(t_{i+1}) \cdots p_n(t_n)$.

The feasible region of problem $J(t_0, c)$ is independent of the principal's type t_0 because we are dealing with environments with generalized private values. Let C denote the set of bundles c such that the feasible region of problem $J(t_0, c)$ is nonempty. Hence, C is the “consumption set” in the fictitious economy. The set C is nonempty (the point where c is identically 0 belongs to C because the allocation that implements the disagreement outcome satisfies all constraints). Moreover, C is convex.

The following result shows that the traders' utility functions in the fictitious economy are well defined.

Lemma 1. *Problem $J(t_0, c)$ has a solution for all $t_0 \in T_0$ and all $c \in C$.*

Proof. We endow $\mathcal{Z}$ with the weak topology. By Prohorov's theorem (cf. Billingsley 1999, Theorem 5.1), $\mathcal{Z}$ is a sequentially compact topological space. Moreover, by definition of the weak topology, for any $\mathbf{t} \in \mathbf{T}$, the functions $u_0(\cdot, \mathbf{t})$ and $u_i(\cdot, \mathbf{t})$ are sequentially continuous functions of $\mathcal{Z}$. Hence, with respect to the product topology on $\mathcal{Z}^{|\mathbf{T}|}$, the objective of problem $J(t_0, c)$ is continuous and the feasible region is compact. Hence, a maximizer exists by Weierstrass' theorem. □

The description of the fictitious economy is completed by the stipulation that each trader's endowment of each good is 0.

We are now ready to define competitive equilibrium in the fictitious economy. A “price vector” is any nonnegative function on $\mathcal{G}$ that is not identically zero. Given a price vector γ, the value of any consumption bundle $c \in C$ is denoted

$$\gamma \cdot c = \sum_{g \in \mathcal{G}} \gamma(g)c(g).$$

Definition 5. A *slack-exchange equilibrium* specifies a list of consumption bundles for all traders, $(c^*_{t_0})_{t_0 \in T_0} \in C^{|T_0|}$, and a price vector, γ^*, such that each trader $t_0 \in T_0$ maximizes her utility given her budget constraint

$$c^*_{t_0} \in \operatorname*{argmax}_{c \in C} V(t_0, c) \quad \text{subject to} \quad \gamma^* \cdot c \leq 0, \tag{3}$$

and, for all goods $g \in \mathcal{G}$, the aggregate consumption (with traders weighted by their prior probabilities) does not exceed the aggregate endowment

$$\sum_{t_0 \in T_0} c^*_{t_0}(g) p_0(t_0) \leq 0. \tag{4}$$

Observe that our definition of competitive equilibrium allows zero prices as well as disposal of goods (4). Typically, in equilibrium some prices are 0 and some quantity of the corresponding goods is disposed. This is natural because in principal–agent problems, typically some constraints imply that some other constraints are automatically (strictly) satisfied.

It is instructive to compare our version of slack-exchange equilibrium to that of Maskin and Tirole (1990). They consider a class of private-value environments with one agent who has a "high" or a "low" type. A number of specific assumptions concerning the outcome space and the payoff functions allow them to ignore trade in all but two constraints—the high type's incentive constraint and the low type's participation constraint. Equilibrium prices for these two constraints are strictly positive, and markets are cleared without any disposal of these goods.

Our more general viewpoint clarifies the purpose of Maskin and Tirole's specific assumptions: they guarantee that the equilibrium prices of the ignored constraints (high type's participation and low type's incentive) are equal to 0 and that the trade in these constraints is fully determined by the trade in the two nonignored constraints.

In the generalized private-value environments that we consider, the set of constraints that have nonzero equilibrium prices cannot be identified a priori, but varies with the parameters of the environment. Hence, we must consider trade in all constraints.

Connecting competitive equilibria with strongly neologism-proof allocations

We are interested in the allocations that correspond to slack-exchange equilibria. An allocation ρ is a *slack-exchange-equilibrium allocation* if there exists a slack-exchange equilibrium $((c^*_{t_0})_{t_0 \in T_0}, \gamma^*)$ such that ρ solves problem $J(t_0, c^*_{t_0})$ for all $t_0 \in T_0$.

To establish a connection to strong neologism-proofness, we begin by showing that in a slack-exchange equilibrium Walras' law holds for all traders who are not "satiated," that is, who do not obtain the highest feasible payoff in equilibrium.

Lemma 2. *Let ρ be a slack-exchange equilibrium allocation in an environment with generalized private values. Let $t_0 \in T_0 \setminus H(\rho)$. Then $\gamma^* \cdot c = 0$ for all maximizers c in (3).*

Proof. Suppose that $\gamma^* \cdot c < 0$ for some maximizer c in (3). Let ρ' be a maximizer of problem $J(t_0, c)$. Then

$$U_0^{\rho'}(t_0) = V(t_0, c) = V(t_0, c^*_{t_0}) = U_0^{\rho}(t_0),$$

implying that $t_0 \notin H(\rho')$. Hence, there exists a type profile $\mathbf{t}_{-0}$ such that $\rho'(\mathbf{t})$ puts probability less than 1 on the outcomes in $\arg\max_{z \in Z} u_0(z, \mathbf{t})$.

Define an allocation ρ'' such that $\rho''(\mathbf{t}) \in \arg\max_{z \in Z} u_0(z, \mathbf{t})$ and $\rho''(\mathbf{t}') = \rho'(\mathbf{t}')$ for all type profiles $\mathbf{t}' \neq \mathbf{t}$.

Consider

$$c' : \mathcal{G} \to \mathbb{R},\ g \to c(g) + \epsilon$$

with $\epsilon > 0$ so small that

$$\gamma^* \cdot c' < 0. \tag{5}$$

The allocation ρ' satisfies all constraints of problem $J(t_0, c')$ with strict inequality. Hence, an allocation ρ''' that implements ρ' with probability $\lambda < 1$ and ρ'' with probability $1 - \lambda$ belongs to the feasible region of problem $J(t_0, c')$ if λ is sufficiently close to 1, implying

$$V(t_0, c') \geq U_0^{\rho'''}(t_0) > U_0^{\rho'}(t_0) = V(t_0, c).$$

Moreover, by (5), the point c' satisfies the constraint of the problem in (3). But this contradicts the assumption that c is a maximizer of the problem in (3). □

We are now ready to connect slack-exchange equilibrium and strong neologism-proofness. The result parallels the first welfare theorem for competitive equilibria.

Lemma 3. *Any slack-exchange equilibrium allocation in any environment with generalized private values is strongly neologism-proof.*

Proof. Let ρ be a slack-exchange equilibrium allocation.

To show that ρ is p_0-feasible, observe first that, by (4), ρ satisfies (1) and (2) for all $i \neq 0$. Because the allocation that implements the disagreement outcome is feasible in problem $J(t_0, 0)$, (2) is satisfied for $i = 0$. Finally, (1) is satisfied for $i = 0$ because the bundle $(r^*_{\hat{t}_0}, c^*_{\hat{t}_0})$ belongs to the feasible region of problem (3).

Suppose that ρ is not strongly neologism-proof. Then there exists a belief q_0 together with a q_0-feasible allocation ρ' such that (i) q_0 is credible for ρ' relative to ρ and (ii) at least one principal-type t_0' is better off in ρ' than in ρ; that is,

$$U_0^{\rho'}(t_0) \geq U_0^{\rho}(t_0) \quad \text{for all } t_0 \in \operatorname{Supp}(q_0) \tag{6}$$

and

$$U_0^{\rho'}(t_0') > U_0^{\rho}(t_0'), \quad t_0' \in \operatorname{Supp}(q_0). \tag{7}$$

For all $t_0 \in \operatorname{Supp}(q_0)$, define c'_{t_0} such that ρ' satisfies all constraints of problem $J(t_0, c'_{t_0})$ with equality. Because ρ' is q_0-feasible,

$$\sum_{t_0 \in T_0} c'_{t_0}(g) q_0(t_0) \leq 0 \quad \text{for all } g \in \mathcal{G}. \tag{8}$$

Using Lemma 2 together with (6), we find

$$\gamma^* \cdot c'_{t_0} \geq 0 \quad \text{for all } t_0 \in \operatorname{Supp}(q_0). \tag{9}$$

Similarly, using (7),

$$\gamma^* \cdot c'_{t'_0} > 0. \tag{10}$$

Building a weighted sum from (9) and (10), we obtain

$$\sum_{t_0 \in T_0} \sum_{g \in \mathcal{G}} \gamma^*(g) c'_{t_0}(g) q_0(t_0) > 0,$$

which yields a contradiction to (8). □

Existence of competitive equilibria in the fictitious economy

The lemma below is the last step toward proving our main result, Proposition 2. Here we use the separability assumption. It guarantees that the budget set of any trader in the slack-exchange economy has an interior point, which is crucial toward showing that her demand correspondence is upper hemicontinuous.

Lemma 4. *A slack-exchange equilibrium exists in any separable environment with generalized private values.*

Our basic line of proof is—like the proof of the corresponding result of Maskin and Tirole (1986, 1990)—inspired by Debreu (1959). The main complication relative to Debreu and Maskin and Tirole arises from the fact that the utility function $V(t_0, \cdot)$ of any trader t_0 is not exogenously given, but is endogenously derived as the solution value of a maximization problem. In particular, the continuity of the objective, which is required in Debreu's arguments, is—in contrast to the situation in Maskin and Tirole's model—not obviously satisfied (the nonobvious assumption of Berge's Maximum Theorem is the lower-hemicontinuity of the feasible region of problem $J(t_0, r, c)$). We circumvent the continuity proof, showing only that $V(t_0, \cdot)$ is upper semicontinuous (15), which captures the absence of downward jumps (see, e.g., Luenberger 1969, p. 40). Hence, by a generalized version of Weierstrass' theorem, an optimal bundle of slacks (i.e., a solution to the problem in (3)) always exists. We use the upper-semicontinuity together with the concavity of $V(t_0, \cdot)$ and the existence of an interior point of the trader's budget set to show that (16), the solution value of problem (3), is lower semicontinuous in the price vector γ. The lower-semicontinuity of the solution value implies the upper-hemicontinuity of the demand correspondence, which is the core step toward applying Kakutani's fixed-point theorem to show equilibrium existence. The complete proof can be found in the Appendix.

6. Relation to other solution concepts

In this section, we explain how strongly neologism-proof equilibrium is related to other solution concepts that have been proposed for informed-principal problems.

The technical concept of a strong unconstrained Pareto optimum (SUPO) plays a major role in the analysis of Maskin and Tirole (1990). A p_0-feasible allocation ρ is a SUPO if there exists no other allocation ρ' together with a belief q_0 about the principal's type such that the agents' (not necessarily the principal's!) incentive and participation constraints are satisfied for ρ' and such that ρ' is weakly preferred to ρ by all types of the principal, and strictly so for at least one type and strictly so for all types if q_0 does not have full support.

Maskin and Tirole (1990, footnote 23) observe that SUPO is equivalent to Farrell's (1993) neologism-proofness, as adapted to their setting. Because in their setting, neologism-proofness is implied by strong neologism-proofness (cf. Remark 1), we can conclude from Maskin and Tirole's observation that SUPO is implied by strong neologism-proofness in their setting.

Myerson (1983) proposes neutral optimum as an axiomatically founded solution concept that always exists in environments with finite outcome spaces and finite type spaces. We do not know the exact relation between neutral optimum and strong neologism-proofness, even in environments with generalized private values. However, a clear relation to another solution concept introduced by Myerson (1983) can be established:

$$\text{strongly neologism-proof allocation} \quad \overset{\not\Longleftarrow}{\Longrightarrow} \quad \text{core allocation.}$$

An allocation ρ is a core allocation if (i) ρ is p_0-feasible and (ii) there exists no allocation ρ' such that ρ' is q_0-feasible for all beliefs q_0 such that $S(\rho', \rho) \neq \varnothing$ and

$$q_0(t_0) = \frac{p(t_0)}{\sum_{t_0' \in S'} p(t_0')} \quad (S(\rho', \rho) \subseteq S' \subseteq T_0).$$

Setting $S' = S(\rho', \rho)$ in this definition, it follows that any strongly neologism-proof equilibrium allocation is a core allocation. Alternatively, in Example 2, any p_0-feasible allocation in which each type of the principal obtains at least the expected payoff 1 is a core allocation, showing that not all core allocations are strongly neologism-proof equilibrium allocations.

Finally, observe that any strongly neologism-proof allocation satisfies the intuitive criterion (as adapted to our setting). To see this, consider an allocation ρ that violates the intuitive criterion. Then there exists a principal-type t_0 and a mechanism M such that, for any belief q_0 that is reasonable when M is proposed, in any sequential equilibrium allocation ρ' of the continuation game when M is proposed, the expected payoff of type t_0 is larger than her payoff in ρ. The belief q_0 that puts probability 1 on type t_0 is reasonable. Hence, q_0 is credible for ρ' relative to ρ, implying that ρ is not strongly neologism-proof.

7. Conclusion

In this paper, we offer a solution to the informed principal problem in the environments with generalized private values. We demonstrate that there exists a perfect-Bayesian equilibrium that is strongly neologism-proof. The equilibrium outcomes can be characterized in terms of the agents' incentive and participation constraints. This makes the problem more tractable. The proof relies on demonstrating the existence of a competitive equilibrium in a fictitious economy in which different types of the principal trade slacks in the agents' incentive and participation constraints.

Appendix

Proof of "if" in Remark 2. In Step 1, we show that the set of agent types can be partitioned into subsets such that to each subset an outcome is assigned that all types in this subset strictly prefer to the disagreement outcome as well as to the outcomes assigned to the other subsets. In Step 2, we show that to each agent type an outcome can be assigned that the agent strictly prefers to the outcomes assigned to the other agents. We then define an allocation such that each agent type obtains a convex combination of the outcome assigned to her subset in Step 1 and the outcome assigned to her in Step 1. If in this convex combination, the Step 1 outcome has a sufficiently large weight, then the allocation satisfies the incentive and participation constraints with strict inequality for all types.

To prove Step 1, one constructs a partition inductively. By (ii), there exists an outcome z_1 that is preferred to the disagreement outcome by at least one agent type, and let P_1 be the set of types that strictly prefer the outcome to the disagreement outcome. If $P_1 \neq T_1$, then, by (ii), some other outcome z' is strictly preferred to disagreement by a subset P_2 of the remaining types. Defining z_2 as a convex combination of the disagreement outcomes z_0 and z', and putting enough weight on z_0, we can guarantee that the types in P_1 prefer z_1 to z_2. The types in P_1 have the reverse preference because they prefer z_2 to z_0 to z_1. This construction can be continued until a partition is obtained.

Step 2 is also proved inductively. Start with any two agent types. By (i), one can assign two outcomes over which the types have opposite strict preferences. Pick a third type. From the outcomes assigned to the first two types, identify one that is most preferred by the third type. Then one perturbs the outcome assignments such that preferences become strict. This can be done by weighting in the most or less preferred outcome with a small probability. This construction can be continued until outcomes are assigned to all types.

Step 1. There exists a partition $P_1, \dots, P_k$ $(k \geq 1)$ of T_1 and outcomes $z_1, \dots, z_k \in \mathcal{Z}$ such that, for all $j = 1, \dots, k$,

$$\forall t_1 \in P_j\colon \quad u_1(z_j, t_1) > u_1(z_0, t_1)$$

$$\forall t_1 \in P_j, l \in \{1, \dots, k\} \setminus \{j\}\colon \quad u_1(z_j, t_1) > u_1(z_l, t_1).$$

To prove this, consider for any $k = 1, 2, \dots,$ the following statement $(*k)$:

> There exist pairwise-disjoint nonempty sets $P_1, \dots, P_k \subseteq T_1$ and outcomes $z_1, \dots, z_k \in \mathcal{Z}$ such that, for all $j = 1, \dots, k$,
>
> $$\forall t_1 \in P_j\colon \quad u_1(z_j, t_1) > u_1(z_0, t_1)$$
>
> $$\forall t_1 \in P_j, l \in \{1, \dots, k\} \setminus \{j\}\colon \quad u_1(z_j, t_1) > u_1(z_l, t_1)$$
>
> $$\forall t_1 \in T_1 \setminus (P_1 \cup \dots \cup P_k)\colon \quad u_1(z_j, t_1) \le u_1(z_0, t_1).$$

Statement $(*1)$ is true: let $z_1 \in \mathcal{Z}$ be an outcome that is strictly preferred to z_0 by at least one agent-type, and denote by P_1 the set of types that strictly prefer z_1 to z_0.

Let $\hat{k}$ be the largest number such that $(*\hat{k})$ is true (observe that $\hat{k} < \infty$ because $(*k)$ fails for all $k > |T_1|$). It is sufficient to show that $P_1 \cup \dots \cup P_{\hat{k}} = T_1$.

Suppose the opposite. Then there exists an outcome $z' \in \mathcal{Z}$ that is strictly preferred to z_0 by at least one agent-type in $T_1 \setminus (P_1 \cup \dots \cup P_{\hat{k}})$. Define

$$P_{\hat{k}+1} = \{t_1 \in T_1 \setminus (P_1 \cup \dots \cup P_{\hat{k}}) \mid u_1(z', t_1) > u_1(z_0, t_1)\}.$$

Now we can define $z_{\hat{k}+1} = \lambda z_0 + (1 - \lambda) z'$ with $\lambda < 1$ sufficiently close to 1 such that statement $(*(\hat{k} + 1))$ is true. This contradicts the maximality of $\hat{k}$.

Step 2. For every $t_1 \in T_1$, there exists an outcome $\zeta(t_1) \in \mathcal{Z}$ such that

$$\forall t_1, t_1' \in T_1, t_1' \ne t_1\colon \quad u_1(\zeta(t_1), t_1) > u_1(\zeta(t_1'), t_1).$$

To prove this, consider for all $P \subseteq T_1$ the following statement $(*P)$:

> For every $t_1 \in P$, there exists an outcome $\zeta(t_1) \in \mathcal{Z}$ such that
>
> $$\forall t_1, t_1' \in P, t_1' \ne t_1\colon \quad u_1(\zeta(t_1), t_1) > u_1(\zeta(t_1'), t_1).$$

If P is a singleton, then $(*P)$ is clearly true.

Let $\hat{P}$ a set of maximal cardinality with the property that $(*\hat{P})$ is true. It is sufficient to show that $\hat{P} = T_1$.

Suppose the opposite. Choose any $s_1 \in T_1 \setminus \hat{P}$. Define

$$\overline{z} \in \operatorname*{argmax}_{z \in \mathcal{Z}} u_1(z, s_1), \qquad \underline{z} \in \operatorname*{argmin}_{z \in \mathcal{Z}} u_1(z, s_1).$$

Let $S_1 = \arg\max_{t_1 \in \hat{P}} u_1(\zeta(t_1), s_1)$ and choose any $s_1' \in S_1$. Define $w(t_1) = \zeta(t_1)$ for all $t_1 \in \hat{P} \setminus S_1$.

Case 1. $u_1(\zeta(s_1'), s_1) > u_1(\underline{z}, s_1)$. Then define $w(s_1') = \zeta(s_1')$. For all $t_1 \in S_1 \setminus \{s_1'\}$, define $w(t_1) = \lambda \zeta(t_1) + (1 - \lambda) \underline{z}$ with some $\lambda < 1$. With λ chosen sufficiently close to 1, statement $(** \hat{P})$ holds:

> Statement $(*\hat{P})$ holds with $\zeta(\cdot)$ replaced by $w(\cdot)$. Moreover, $u_1(w(t_1), s_1) < u_1(w(s_1'), s_1)$ for all $t_1 \in \hat{P} \setminus \{s_1'\}$.

Case 2: $u_1(\zeta(s_1'), s_1) < u_1(\overline{z}, s_1)$.[12] For all $t_1 \in S_1 \setminus \{s_1'\}$, define $w(t_1) = \zeta(t_1)$. Define $w(s_1') = \lambda \zeta(s_1) + (1 - \lambda) \overline{z}$ with some $\lambda < 1$. With λ chosen sufficiently close to 1, statement $(** \hat{P})$ holds.

[12] One of Cases 1 or 2 always occurs, because otherwise $u_1(\underline{z}, s_1) = u_1(\overline{z}, s_1)$, implying that type s_1 is indifferent between all outcomes, which is impossible because there exists an outcome that she strictly prefers to the disagreement outcome.

Because the types s_1 and s_1' have nonidentical preferences, there exist outcomes $y, y' \in \mathcal{Z}$ such that

$$u_1(y, s_1) > u_1(y', s_1), \qquad u_1(y, s_1') < u_1(y', s_1').$$

Define $w'(s_1) = \lambda w(s_1) + (1-\lambda)y$ and $w'(s_1') = \lambda w(s_1) + (1-\lambda)y'$. Define $w'(t_1) = w(t_1)$ for all $t_1 \in \hat{P} \setminus \{s_1, s_1'\}$. With λ chosen sufficiently close to 1, statement $(*(\hat{P} \cup \{s_1'\}))$ holds with $\zeta(\cdot)$ replaced by $w'(\cdot)$, a contradiction to the maximality of $\hat{P}$, completing Step 2.

Now let $P_1, \ldots, P_k$ $(k \geq 1)$ and $z_1, \ldots, z_k$ be as in Step 1, and let the function $\zeta(\cdot)$ be as in Step 2. For all $t_1 \in T_1$, define $j(t_1)$ such that $t_1 \in P_j$. Then the allocation ρ defined by

$$\rho(t_0, t_1) = \lambda z_{j(t_1)} + (1-\lambda)\zeta(t_1)$$

implies that all incentive and participation constraints of all agent types are satisfied with strict inequality if λ is chosen sufficiently close to 1. □

Proof of Lemma 4. Any consumption bundle in C, as well as any price vector, belongs to the Euclidean space $\mathbb{R}^{|\mathcal{G}|}$. We use standard operators in Euclidean spaces such as $+$, min, or $\leq$, all of which are defined componentwise.

First we show that the set

$$C \text{ is closed.} \tag{11}$$

To see this, consider any sequence $c^m \to c$ such that $c^m \in C$. By assumption, the constraint set of $J(t_0, c^m)$ contains a point ρ^m. For all sufficiently large m, the point ρ^m belongs to the feasible region of problem $J(t_0, c+1)$, where 1 denotes the vector that is identically equal to 1. Because the latter feasible region is compact, ρ^m has a subsequence that converges to some point ρ'. By continuity, ρ' belongs to the feasible region of $J(t_0, c)$. Hence, $c \in C$. This completes the proof of (11).

Because Z is compact, there exists a lower bound for the size of the left-hand side of every constraint of $J(t_0, c)$. Hence, there exists $\overline{c} \in C$ such that

$$V(t_0, c) = V(t_0, \min\{c, \overline{c}\}) \quad \text{for all } c \in C. \tag{12}$$

Similarly, there exists $\underline{c} \in \mathbb{R}^{|\mathcal{G}|}$ such that

$$C \subseteq \{c \mid c \geq \underline{c}\}. \tag{13}$$

By (11), (12), and (13), the set

$$D = C \cap \{c \mid c \leq \overline{c}\} \text{ is compact.} \tag{14}$$

Define the unit simplex

$$\Delta = \left\{\gamma \in \mathbb{R}^{|\mathcal{G}|} \,\middle|\, \gamma \geq 0, \sum_{g \in \mathcal{G}} \gamma(g) = 1\right\}.$$

For every $\gamma \in \Delta$, consider the problem

$$E(t_0, \gamma): \quad \max_{c \in D} V(t_0, c) \quad \text{subject to} \quad \gamma \cdot c \leq 0.$$

The objective $V(t_0, \cdot)$ of problem $E(t_0, \gamma)$ is upper semicontinuous: for any convergent sequence (c^m) in D,

$$V\Big(t_0, \lim_m c^m\Big) \geq \limsup_m V(t_0, c^m). \tag{15}$$

To see (15), let $c = \lim c^m$ and let (c^{m_l}) be a subsequence such that $V(t_0, c^{m_l})$ converges. Let ρ^l be a maximizer of problem $J(t_0, c^{m_l})$ and let (ρ^{l_k}) be a subsequence such that ρ^{l_k} converges. Because $\lim_k c^{m_{l_k}} = c$, the limit $\rho' = \lim_k \rho^{l_k}$ belongs to the feasible region of problem $J(t_0, c)$. Hence,

$$V(t_0, c) \geq U_0^{\rho'}(t_0) = \lim_k U_0^{\rho^{l_k}}(t_0) = \lim_k V(t_0, c^{m_{l_k}}) = \lim_l V(t_0, c^{m_l}).$$

By (15) and because, by (14), the feasible region of problem $E(t_0, \gamma)$ is compact, a maximizer to problem $E(t_0, \gamma)$ exists (see, e.g., Luenberger 1969, p. 40); let $e(t_0, \gamma)$ denote the set of maximizers.

The correspondence $e(t_0, \cdot) : \Delta \to D$ is convex-valued because $V(t_0, \cdot)$ is concave. To show that $e(t_0, \cdot)$ is upper hemicontinuous, we begin by showing that for every sequence in Δ,

$$\text{if } \gamma^m \to \gamma, \quad \text{then } \liminf_m v\big(t_0, \gamma^m\big) \geq v(t_0, \gamma), \tag{16}$$

where $v(t_0, x)$ denotes the value reached at the maximum of problem $E(t_0, x)$ for any $x \in \Delta$.

Let $c^* \in e(t_0, \gamma)$. If $c^* \cdot \gamma < 0$, then $c^* \cdot \gamma^m < 0$ if m is sufficiently large, hence c^* belongs to the feasible region of $E(t_0, \gamma^m)$, which shows (16). Now suppose that

$$c^* \cdot \gamma = 0. \tag{17}$$

Using the separability assumption, the set D contains a strictly negative point $c^- < 0$. For all large m, define

$$\alpha^m = \min\left\{1, \frac{-c^- \cdot \gamma^m}{c^* \cdot \gamma^m - c^- \cdot \gamma^m}\right\}. \tag{18}$$

The convex combination $c^m = \alpha^m c^* + (1 - \alpha^m) c^- \in D$. By construction, c^m belongs to the feasible region of problem $E(t_0, \gamma^m)$. Hence, using the concavity of $V(t_0, \cdot)$,

$$\alpha^m V(t_0, c^*) + (1 - \alpha^m) V(t_0, c^-) \leq V(t_0, c^m) \leq v(t_0, \gamma^m). \tag{19}$$

As $m \to \infty$, we have $\alpha^m \to 1$ by (17) and (18). Hence, (19) implies

$$V(t_0, c^*) \leq \liminf_m v(t_0, \gamma^m).$$

Because $V(t_0, c^*) = v(t_0, \gamma)$, we obtain (16).

To show that $e(t_0, \cdot)$ is upper hemicontinuous, suppose that $\gamma^m \to \gamma$, $c^m \in e(t_0, \gamma^m)$, and $c^m \to c$. Then

$$V(t_0, c) \geq \liminf_m V(t_0, c^m) = \liminf_m v(t_0, \gamma^m) \geq v(t_0, \gamma),$$

where the first inequality follows from (15) and the second inequality follows from (16). Hence, $c \in e(t_0, \gamma)$ because c belongs to the feasible region of $E(t_0, \gamma)$.

Define a correspondence $h : \prod_{t_0 \in T_0} D \to \Delta$ by letting $h((c_{t_0})_{t_0 \in T_0})$ be the set of solutions to the problem

$$R((c_{t_0})_{t_0 \in T_0}): \quad \max_{\gamma \in \Delta} \sum_{t_0 \in T_0} p(t_0)\gamma \cdot c_{t_0}.$$

By Berge's maximum theorem, h is upper hemicontinuous. Moreover, h is convex-valued. By Kakutani's theorem, the correspondence

$$\Bigl(\prod_{t_0 \in T_0} D\Bigr) \times \Delta \to \Bigl(\prod_{t_0 \in T_0} D\Bigr) \times \Delta$$
$$(x, \gamma) \mapsto \Bigl(\prod_{t_0 \in T_0} e(t_0, \gamma)\Bigr) \times h(x)$$

has a fixed point $((c^*_{t_0})_{t_0 \in T_0}, \gamma^*)$.

To complete the proof that $((c^*_{t_0})_{t_0 \in T_0}, \gamma^*)$ is a slack-exchange equilibrium, it remains to show (4).

Suppose that (4) fails; i.e., there exists $g \in \mathcal{G}$ such that

$$\sum_{t_0 \in T_0} p(t_0) c^*_{t_0}(g) > 0. \tag{20}$$

Choose $\gamma \in \Delta$ such that $\gamma(g') = 0$ for all $g' \neq g$. Then (20) implies that

$$\sum_{t_0 \in T_0} p(t_0)\gamma \cdot c^*_{t_0} > 0.$$

This contradicts the fact that γ^* solves problem $R((c^*_{t_0})_{t_0 \in T_0})$, because, using the constraint of problem $E(t_0, \gamma^*)$ for all t_0,

$$\sum_{t_0 \in T_0} p(t_0)\gamma^* \cdot c^*_{t_0} \leq 0.$$

□

References

Balkenborg, Dieter and Miltiadis Makris (2010), "An undominated mechanism for a class of informed principal problems with common values." Unpublished paper. [468]

Beaudry, Paul (1994), "Why an informed principal may leave rents to an agent." *International Economic Review*, 35, 821–832. [468]

Billingsley, Patrick (1999), *Convergence of Probability Measures*, second edition. Wiley, New York. [477]

Cella, Michela (2008), "Informed principal with correlation." *Games and Economic Behavior*, 64, 433–456. [468]

Chade, Hector and Randy Silvers (2002), "Informed principal, moral hazard, and the value of a more informative technology." *Economics Letters*, 74, 291–300. [468]

Cho, In-Koo and David M. Kreps (1987), "Signaling games and stable equilibria." *Quarterly Journal of Economics*, 102, 179–221. [466]

Cramton, Peter, Robert Gibbons, and Paul Klemperer (1987), "Dissolving a partnership efficiently." *Econometrica*, 55, 615–632. [474]

de Clippel, Geoffroy and Enrico Minelli (2004), "Two-person bargaining with verifiable information." *Journal of Mathematical Economics*, 40, 799–813. [466]

Debreu, Gerard (1959), *Theory of Value*. Yale University Press, New Haven, Connecticut. [467, 480]

Farrell, Joseph (1993), "Meaning and credibility in cheap-talk games." *Games and Economic Behavior*, 5, 514–531. [466, 470, 481]

Fleckinger, Pierre (2007), "Informed principal and countervailing incentives." *Economics Letters*, 94, 240–244. [468]

Grossman, Sanford J. and Motty Perry (1986), "Perfect sequential equilibrium." *Journal of Economic Theory*, 39, 97–119. [466, 470]

Guesnerie, Roger and Jean-Jacques Laffont (1984), "A complete solution to a class of principal–agent problems with an application to the control of a self-managed firm." *Journal of Public Economics*, 25, 329–369. [468]

Kaya, Ayça (2010), "When does it pay to get informed?" *International Economic Review*, 51, 533–551. [468]

Luenberger, David G. (1969), *Optimization by Vector Space Methods*. Wiley, New York. [480, 485]

Mailath, George J., Masahiro Okuno-Fujiwara, and Andrew Postlewaite (1993), "Belief-based refinements in signalling games." *Journal of Economic Theory*, 60, 241–276. [466]

Maskin, Eric and Jean Tirole (1986), "Principals with private information, I: Independent values." Discussion Paper 1234, Department of Economics, Harvard University. [480]

Maskin, Eric and Jean Tirole (1990), "The principal–agent relationship with an informed principal: The case of private values." *Econometrica*, 58, 379–409. [467, 468, 469, 470, 474, 478, 480, 481]

Maskin, Eric and Jean Tirole (1992), "The principal–agent relationship with an informed principal, II: Common values." *Econometrica*, 60, 1–42. [468, 469]

Myerson, Roger B. (1981), "Optimal auction design." *Mathematics of Operations Research*, 6, 58–73. [468]

Myerson, Roger B. (1983), "Mechanism design by an informed principal." *Econometrica*, 51, 1767–1797. [466, 467, 468, 469, 471, 481]

Myerson, Roger B. and Mark A. Satterthwaite (1983), "Efficient mechanisms for bilateral trading." *Journal of Economic Theory*, 29, 265–281. [468, 471]

Mylovanov, Tymofiy and Thomas Tröger (2008), "Optimal auction design and irrelevance of privacy of information." Unpublished paper. [468]

Quesada, Lucia (2010), "A comprehensive note on the informed principal with private values and independent types." Unpublished paper. [467]

Riley, John G. (2001), "Silver signals: Twenty-five years of screening and signaling." *Journal of Economic Literature*, 39, 432–478. [466]

Severinov, Sergei (2008), "An efficient solution to the informed principal problem." *Journal of Economic Theory*, 141, 114–133. [468]

Skreta, Vasiliki (2011), "On the informed seller problem: Optimal information disclosure." *Review of Economic Design*, 15, 1–36. [468]

Tan, Guofu (1996), "Optimal procurement mechanisms for an informed buyer." *Canadian Journal of Economics*, 29, 699–716. [468]

Yilankaya, Okan (1999), "A note on the seller's optimal mechanism in bilateral trade with two-sided incomplete information." *Journal of Economic Theory*, 87, 267–271. [468, 471]

Submitted 2010-5-8. Final version accepted 2011-8-1. Available online 2011-8-2.