Robust Strong Nash Implementation

Huiyi Guo* and Nicholas C. Yannelis†

November 4, 2017

Abstract: When a mechanism designer does not know agents’ belief structure about each other’s private information, following the spirit of Bergemann and Morris (2005, 2011, etc.), she designs robust mechanisms that do not rely on agents’ beliefs. We introduce coalitional structures into the belief-free approach. Specifically, we provide necessary and almost sufficient conditions for robust full implementation of a social choice set as an interim strong Nash equilibrium. This enables us to implement new social choice sets compared to robust Nash implementation. As an extension, when the mechanism designer has information on only certain coalitions can be formed, or has no information on which coalitions can be formed, this paper also provides conditions for robust implementation under other coalitional patterns. As different social choice sets may be implementable under different coalitional structures, this paper provides insights into when agents should be allowed to communicate and cooperate with each other.

Keywords: Bayesian implementation, robust implementation, interim strong Nash equilibrium.

1 Introduction

In implementation theory, if a mechanism can be designed such that all its equilibria coincide with an exogenous social choice set, then the set is said to be fully implementable. Under incomplete information, agents’ private information is traditionally modeled by a commonly known type space. However, some details of the type space (e.g., agents’ beliefs) may not

*Department of Economics, University of Iowa. Email address: huiyi-guo@uiowa.edu.
†Department of Economics, University of Iowa. Email address: nicholasyannelis@gmail.com.
be available to the mechanism designer in practice. Therefore, it is suggested to weaken the common knowledge assumption for a more robust conclusion (Wilson, 1987). To relax the common knowledge assumption, Bergemann and Morris (2005, 2011, etc.) adopt a belief-free approach to study robust mechanism design. Consistent with their spirit, we investigate the problem of robust full implementation as an interim strong Nash equilibrium. We establish that robust coalitional incentive compatibility and robust coalitional monotonicity are necessary and almost sufficient for robust strong Nash implementation. With the two conditions, interim strong Nash implementation can be guaranteed under all type spaces. Moreover, the robust implementation framework enables us to go beyond Bayesian implementation, and thus one can include more general preferences, e.g., the maximin expected utility of Gilboa-Schmeidler (1989). Our modeling provides new insights for social choice sets that may not be robustly Nash implementable (e.g., the fine core).

The interim strong Nash equilibrium is immune to all coalitional deviations, and therefore it is more stable than the interim Nash equilibrium. Stability is a concern in economics, e.g., the matching and the network literature. Excluding all collusion possibilities may be impractical. For example, it may be the case that a social choice set is not interim (robust) Nash implementable, but it is interim (robust) strong Nash implementable, and thus the mechanism designer may wish to facilitate agents' communication in order to obtain her objective. Under complete information, the problem of strong Nash implementation has been studied by Maskin (1979) and Dutta and Sen (1991) among others. Under the Bayesian framework, Hahn and Yannelis (2001) consider a related notion in exchange economies.

Our framework to address robust strong Nash implementation provides insights on other implementation concepts, such as interim Nash implementation, pairwise stable Nash implementation, and simultaneous implementation as an interim Nash and strong Nash equilibrium. In fact, by restricting the deviating coalitions to be singletons, our interim strong Nash implementation concept reduces to interim Nash implementation. Similarly, by restricting the deviating coalitions to have cardinality no more than two, our equilibrium has the flavor of pairwise stable Nash equilibrium, which has been used in the network literature. By slightly modifying our mechanism into one that simultaneously implements a social choice set as an interim Nash and strong Nash equilibrium, the mechanism designer can guarantee the desired outcome even when she does not know which coalitions could be formed. Some of these implementation concepts and mechanisms are provided in Section 7. If future researchers were to adopt a different implementation concept with certain stability requirements, the methodology of our paper could be adapted.

From a technical point of view, we construct new arguments in order to prove our
robust strong Nash implementation results. Our basic idea follows the lottery construction of Bergemann and Morris (2011), but their mechanism cannot prevent profitable coalitional deviations. Therefore, the focus on coalitional deviations requires non-trivial modifications.

The interim strong Nash equilibrium is a refinement of the interim Nash equilibrium. This does not mean that our robust strong Nash implementation is stronger than robust Nash implementation because full implementation requires the equilibria to coincide with the social choice set. On the partial implementation side, robust strong Nash implementation implies robust Nash implementation. However, the robust coalitional monotonicity condition is weaker than the robust monotonicity condition of Bergemann and Morris (2011), which is necessary and almost sufficient for robust Nash implementation. This allows us to implement some social choice sets that are not robustly Nash implementable.

Several applications of our results are provided. The first application is the fine core. We show that it is robustly strong Nash implementable. This complements the work of Palfrey and Srivastava (1987), which shows that several core notions fail to be Bayesian Nash implementable. The second application involves a benevolent social planner allocating a unit of private good efficiently. The efficient rule is robustly strong Nash implementable, but not robustly Nash implementable. Then we revisit the public good example of Bergemann and Morris (2009) and show that their efficient social choice function is robustly strong Nash implementable if and only if agents have a common value. This example shows that robust Nash implementation doesn’t imply robust strong Nash implementation and vice versa. In the end, we study two variants of the public good example, one for implementation under another known coalitional pattern, and the other for implementation under unknown coalitional patterns.

The paper proceeds as follows. Section 2 presents the primitives of the environment. The concept of full implementation is given in Section 3. We provide necessary and almost sufficient conditions on robust strong Nash implementation in Section 4 and 5. In Section 6, we extend the results to address a situation where the feasible outcome set is random. Section 7 studies implementation under other coalitional patterns. Section 8 provides applications. In Section 9, we indicate how our results can be extended to accommodate the maximin expected utility.
2 Asymmetric Information Environment

As in Bergemann and Morris (2005), we first consider an asymmetric information environment without any specification on beliefs, namely a payoff environment. Formally, a payoff environment is given by \( \mathcal{E} = \{ I, A, (\Theta_i, u_i)_{i=1}^n \} \), where

- \( I = \{1, \ldots, n\} \) is the set of agents;
- \( A \) is the set of feasible outcomes, i.e., the set of all lotteries over a pure feasible outcome set \( X \); the set \( X \) is assumed to be deterministic, but we relax the assumption in Section \[\]
- \( \Theta = \Theta_1 \times \ldots \times \Theta_n \) is the payoff type set, and \( \theta_i \in \Theta_i \) is agent \( i \)'s payoff type; we focus on the case where each \( \Theta_i \) is finite;
- \( u_i : A \times \Theta \to \mathbb{R} \), agent \( i \)'s random utility function, represents agent \( i \)'s expected utility of consuming a lottery \( a \in A \), when the realized payoff type profile is \( \theta \in \Theta \).

A type space is a collection \( \mathcal{T} = (T_i, \hat{\theta}_i, \pi_i)_{i=1}^I \), where

- \( t_i \in T_i \) is a type of agent \( i \), which represents agent \( i \)'s private information; the set of all type profiles is denoted by \( T = \prod_{i \in I} T_i \); we focus on the case where each \( T_i \) is finite;
- agent \( i \) with a type \( t_i \) has a payoff type \( \hat{\theta}_i(t_i) \), which is defined by an onto mapping \( \hat{\theta}_i : T_i \to \Theta_i \); let \( \hat{\theta} : T \to \Theta \) be the mapping defined by \( \hat{\theta}(t) = (\hat{\theta}_1(t_1), \ldots, \hat{\theta}_n(t_n)) \) for all \( t \in T \);
- agent \( i \) with a type \( t_i \) has a belief type \( \pi_i(t_i) \); each \( \pi_i(t_i) \) is a probability distribution over \( T_{-i} \), assigning probability \( \pi_i(t_i)[t_{-i}] \) to the event that others have types \( t_{-i} \).

In general, we do not require that each \( \pi_i(t_i) \) has full support over \( T_{-i} \). However, we impose such a condition to simplify the statement of some sufficient conditions. We call the restricted type spaces full support type spaces. In a type space, the full support assumption implies the closure condition in the implementation literature, which is needed for implementation of social choice sets.\[\]

A social choice function \( f : \Theta \to A \) is an exogenous allocation rule contingent on agents’ payoff types. A social choice set \( F \) is a set of social choice functions.

---

1Here \( T_{-i} = \prod_{j \neq i} T_j \), and \( \Delta(T_{-i}) \) is the set of all probability distributions on \( T_{-i} \).
2See Jackson (1991) for a more detailed discussion about the closure condition in (potentially not full support) type spaces.
3 Full Implementation

A mechanism is a pair \((M, g) = (\prod_{i \in I} M_i, g)\), where \(M_i\) is the set of all messages that agent \(i\) can submit, i.e., \(M_i\) is the message space of agent \(i\). An outcome function is a mapping \(g : M \rightarrow A\), which assigns to each message profile a feasible allocation.\(^3\) Agent \(i\)'s strategy \(\sigma_i : T_i \rightarrow M_i\) is a private information contingent plan of submitting messages. A strategy profile is given by \(\sigma = (\sigma_1, \sigma_2, ..., \sigma_n)\). For simplicity, denote by \(\sigma_S\) the strategy for all agents in \(S \subseteq I\) and by \(\sigma_{-S}\) the strategy for all agents not in \(S\).

Under a type space \(T\), a mechanism \((M, g)\) fully implements a social choice set \(F\) if the following two conditions are satisfied:

1. for any \(f \in F\), there exists an equilibrium \(\sigma : T \rightarrow M\) of the mechanism \((M, g)\) such that \(g(\sigma(t)) = f(\hat{\theta}(t))\) for all \(t \in T\);
2. if \(\sigma\) is an equilibrium of the mechanism \((M, g)\), then there exists \(f \in F\) such that \(g(\sigma(t)) = f(\hat{\theta}(t))\) for all \(t \in T\).

If the first requirement is satisfied, then the social choice set \(F\) is said to be partially implemented by \((M, g)\).

When the type space is common knowledge among the mechanism designer and the agents, we call the full implementation problem an interim implementation problem. If there exists a mechanism \((M, g)\) that fully implements \(F\) in all type spaces associated with the payoff environment \(\Theta\), then the social choice set is said to be robustly implementable.

In this paper, we adopt the notion of the interim strong Nash equilibrium, which is stronger than the interim (Bayesian) Nash equilibrium in that the interim strong Nash equilibrium is immune to all coalitional deviations. A social choice set \(F\) is said to be robustly strong Nash implementable if there is a mechanism \((M, g)\) that implements \(F\) as an interim strong Nash equilibrium in all type spaces.

Before defining the notion of an interim strong Nash equilibrium, we introduce some notation. Let the symbol \(\setminus\) denote the difference between two sets. For each distribution \(\pi_i(t_i^*)[\cdot]\) and \(S \supseteq i\), the notation \(\pi_i(t_i^*)[t_{S\setminus\{i\}}^*]\) represents the marginal probability that the coalition \(S\setminus\{i\}\) has types \(t_{S\setminus\{i\}}^*\). For \(S \subseteq I\), \(t_S^* \in T_S\), and \(i \in S\), if the marginal probability \(\pi_i(t_i^*)[t_{S\setminus\{i\}}^*] > 0\), Bayes’ rule can be applied. Therefore, we let \(\pi_i(t_i^*)[t_{-i}]\) be the conditional probability that \(t_{-i}\) is the true type profile of agents in \(I \setminus \{i\}\), given that \(S\) has a type profile \(t_S^*\). If the marginal probability \(\pi_i(t_i^*)[t_{S\setminus\{i\}}^*] = 0\), Bayes’ rule cannot be applied. Then we assume that agent \(i\) updates her belief into \(\pi_i(t_i^*)[\cdot]\), which is an arbitrary commonly known

\(^3\)In Sections 6, an outcome function specifies the net trade, rather than the allocation that agents consume.
distribution satisfying $\pi_i(t^*_S)[t^*_S \setminus \{i\}] = 1$.

The definition of an interim strong Nash equilibrium is presented below, which is a variant of the coalitional Bayesian Nash equilibrium of Hahn and Yannelis (2001).

**Definition 3.1:** In a type space $T$, the strategy profile $\sigma^*$ is an **interim strong Nash equilibrium** of the mechanism $(M,g)$ if there does not exist $S \subseteq I$, $t^*_S \in T_S$, and $\sigma'_S : T_S \rightarrow M_S$, such that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i\left(g(\sigma'_S(t^*_S), \sigma^*_{-S}(t_{-S})), \hat{\theta}(t^*_S, t_{-S})\right) \pi_i(t^*_S)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i\left(g(\sigma^*(t^*_S, t_{-S})), \hat{\theta}(t^*_S, t_{-S})\right) \pi_i(t^*_S)[t_{-i}] .$$

The above definition implies that the current paper focuses on implementation under pure strategies. If one considers mixed strategies, then Propositions 4.1, 4.2, 6.1, and Theorem 7.3 still go through.

In the above definition of an interim strong Nash equilibrium, we assume that the coalition members truthfully pool their information within the coalition. When agents pool their information, they obtain higher efficiency. In the Appendix, we present an alternative definition without the information pooling assumption and indicate how our results can be modified to accommodate the new definition.

In Definition 3.1, we require the deviation to be strictly profitable for every member of the coalition. This is consistent with the definition of strong Nash equilibrium in the literature. Alternatively, we can strengthen the notion by requiring the inequalities to be weak for all members, but strict for at least one member. Such a modification does not change our main results, as long as we adjust our necessary and sufficient conditions accordingly.

### 4 Necessary Conditions

In this section, we introduce conditions that are necessary for robust strong Nash implementation. We show that if a social choice set is robustly strong Nash implementable, it satisfies robust coalitional incentive compatibility and robust coalitional monotonicity. The sufficiency of the conditions is discussed in subsequent sections.

---

4 Some other papers impose specific updating rules in the case where Bayes’ rule fails, e.g., Penta (2015) and Müller (2016).

5 See Maskin (1979) for example.
4.1 Incentive Compatibility

In a type space $\mathcal{T}$, a deception of agent $i$ is a mapping $\alpha_i : T_i \rightarrow T_i$, i.e., under $\alpha_i$, the type-$t_i$ agent reports $\alpha_i(t_i)$. We denote by $\alpha$ the deception profile $(\alpha_1, ..., \alpha_n)$. To prevent a coalitional deviation from truth-telling, we need the following condition of coalitional incentive compatibility.

**Definition 4.1:** In a type space $\mathcal{T}$, a social choice set $F$ is **coalitional incentive compatible** if there is no $f \in F$, $S \subseteq I$, $t^*_S \in T_S$, and $\alpha_S : T_S \rightarrow T_S$ such that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i\left(f\left(\hat{\theta}(\alpha_S(t^*_S), t_{-S}), t^*_S, t_{-S}\right)\right) \pi_i(t^*_i[t_{-i}]) > \sum_{t_{-i} \in T_{-i}} u_i\left(f\left(\hat{\theta}(t^*_S, t_{-S}), \hat{\theta}(t^*_S, t_{-S})\right)\right) \pi_i(t^*_i[t_{-i}]).$$

Following standard arguments in implementation theory and mechanism design, it is straightforward to see that the above condition is necessary for a social choice set to be implementable as an interim strong Nash equilibrium. We thereby omit the proof. We also remark that the coalitional incentive compatibility condition implies (interim) incentive compatibility, which is obvious by setting $S$ to be singletons.

When the type space is not common knowledge, we need the following condition, which is stronger than coalitional incentive compatibility.

**Definition 4.2:** A social choice set $F$ is **robust coalitional incentive compatible** if for all $f \in F$, $S \subseteq I$, and $\theta'_S \neq \theta^*_S$, there exists $i \in S$ such that

$$u_i\left(f\left(\theta^*_S, \theta_{-S}\right), (\theta^*_S, \theta_{-S})\right) \geq u_i\left(f\left(\theta'_S, \theta_{-S}\right), (\theta^*_S, \theta_{-S})\right)$$

for all $\theta_{-S} \in \Theta_{-S}$.

In the Appendix, Proposition A.1 shows that $F$ is robust coalitional incentive compatible if and only if it is coalitional incentive compatible in all type spaces. As coalitional incentive compatibility is necessary for interim strong Nash implementation, we have the following proposition.

**Proposition 4.1:** If a social choice set $F$ is interim strong Nash implementable in all type spaces with payoff environment $\Theta$, then $F$ is robust coalitional incentive compatible.

4.2 Monotonicity

For full implementation, we need a version of the monotonicity condition, which is necessary and almost sufficient for Nash implementation under complete information (e.g., Maskin,
Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1989), and Jackson (1991) adopt related concepts under incomplete information. They prove that the conditions of incentive compatibility and Bayesian monotonicity are necessary and almost sufficient for Bayesian implementation. To study the problem of coalitional Bayesian Nash implementation, Hahn and Yannelis (2001) further propose a coalitional Bayesian monotonicity condition. We modify their condition in order to be consistent with our definition of interim strong Nash equilibrium.

We define first the notion of acceptable deceptions. Given $f \in F$, the deception $\alpha : T \to T$ is acceptable if there exists $f' \in F$ with $f'(t) = f(\alpha(t))$ for all $t \in T$. Otherwise, the deception is unacceptable.

For ease of notation, for each $f \in F$, $S \subseteq I$, and $t_S' \in T_S$, we denote the reward set by $H^f_{t_S'}$, which is the collection of mappings $h : T \to A$ satisfying the following condition. For all $t_S'' \in T_S$, there exists $i \in S$ such that

$$\sum_{t_{-i} \in T_{-i}} u_i \left( f(\hat{\theta}(t_{S'}^n, t_{-S}^n)), \hat{\theta}(t_{S'}^n, t_{-S}^n) \right) \pi_i(t_S^n)[t_{-i}] \geq \sum_{t_{-i} \in T_{-i}} u_i(h(t_{S'}^n, t_{-S}), \hat{\theta}(t_{S'}^n, t_{-S})) \pi_i(t_S^n)[t_{-i}]$$

**Definition 4.3:** Given a type space $T$, a social choice set $F$ satisfies the coalitional monotonicity condition if for all $f \in F$, whenever $\alpha$ is unacceptable, there exists $S \subseteq I$, $t_S^* \in T_S$, and $h \in H^f_{\alpha S}(t_S^*)$ such that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i(h(\alpha(t_S^*, t_{-S}^*)), \hat{\theta}(t_S^*, t_{-S}^*)) \pi_i(t_S^*)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i \left( f(\hat{\theta}(\alpha(t_S^*, t_{-S}^*))), \hat{\theta}(t_S^*, t_{-S}^*) \right) \pi_i(t_S^*)[t_{-i}]$$.\n
We remark that the coalitional monotonicity condition is weaker than the Bayesian (interim) monotonicity condition, which is obvious by setting $S$ to be singletons.

**Lemma 4.1:** In a type space $T$, if a social choice set $F$ is implementable as an interim strong Nash equilibrium, then $F$ satisfies the coalitional monotonicity condition.

**Proof.** For $f \in F$, as $F$ is implementable, there exists a mechanism $(M, g)$ and its interim strong Nash equilibrium $\sigma^*$ such that $g(\sigma^*(t)) = f(\hat{\theta}(t))$ for all $t \in T$. If $\alpha : T \to T$ is

\[\text{The requirement that each } h \text{ is a mapping from } T \text{ to } A \text{ is consistent with the classical Bayesian implementation literature. However, it should be noted that for a fixed } S, H^f_{t_S'} \text{ is independent of the realization of } t_S'. \text{ This implies that we can denote the reward set by } H^f_{S'} \text{ instead of } H^f_{t_S'}, \text{ and let each } h \in H^f_{S} \text{ be a mapping from } T_{-S} \text{ to } A. \text{ We did not choose this notation because it requires some explanations if } S = I \text{ and } S^c = \emptyset. \]
unacceptable, then \(g(\sigma^*(\alpha(t))) = f(\hat{\theta}(\alpha(t)))\) implies that \(\sigma^* \circ \alpha\) is not an interim strong Nash equilibrium. Hence, there exists \(S \subseteq I\), \(t'_S \in T_S\), and \(\sigma'_S : T_S \to M_S\) such that for all \(i \in S\),

\[
\sum_{t_{-i} \in T_{-i}} u_i \left( g\left( \sigma'_S(t'_S), \sigma'_{-S}(\alpha_{-S}(t_{-S})) \right), \hat{\theta}(t'_S, t_{-S}) \right) \pi_i(t'_S)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i \left( g\left( \sigma^*(\alpha(t'_S), t_{-S}) \right), \hat{\theta}(t'_S, t_{-S}) \right) \pi_i(t'_S)[t_{-i}].
\]

Define \(h : T \to A\) by \(h(t) = g(\sigma'_S(t'_S), \sigma'_{-S}(t'_S))\) for all \(t \in T\). Then we have

\[
\sum_{t_{-i} \in T_{-i}} u_i \left( h(\alpha(t'_S, t_{-S})), \hat{\theta}(t'_S, t_{-S}) \right) \pi_i(t'_S)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i \left( f(\hat{\theta}(\alpha(t'_S), t_{-S})), \hat{\theta}(t'_S, t_{-S}) \right) \pi_i(t'_S)[t_{-i}].
\]

As \(\sigma^*\) is an interim strong Nash equilibrium, at all \(t''_S \in T_S\), \(\sigma'_S(t''_S)\) cannot be a profitable deviating message to send. This observation as well as the definition of \(h\) implies that \(h \in H^I_{\alpha_S(t'_S)}\). Therefore, we have established the coalitional monotonicity condition. \(\square\)

Bergemann and Morris (2011) propose a robust monotonicity condition, which is necessary and almost sufficient for robust Nash implementation. For the purpose of robust strong Nash implementation, we focus on a weaker condition, the robust coalitional monotonicity condition.

A deception (of payoff types) of agent \(i\) is a set-valued mapping \(\beta_i : \Theta_i \to 2^{\Theta_i \setminus \emptyset}\). Let \(\beta\) denote the deception profile \((\beta_1, ..., \beta_n)\). Given the social choice set \(F\) and \(f \in F\), the deception is acceptable if for any selection \(\beta \in \beta\), there exists \(f' \in F\) such that \(f(\beta(\theta)) = f'(\theta)\) for all \(\theta \in \Theta\). Otherwise the deception is unacceptable.

For each \(i \in I\) and \(\theta'_i \in \Theta_i\), define \(\beta^{-1}_i(\theta'_i) = \{\theta_i : \theta'_i \in \beta_i(\theta_i)\}\), which is the set of all possible true payoff types of agent \(i\), given she reports \(\theta'_i\). Let a reasonable conjecture of agent \(i\) be a mapping \(\psi_i : \Theta_{-i} \to \Delta(\Theta_{-i})\), where each \(\psi_i(\theta'_{-i})\) is a distribution with support in \(\beta^{-1}_i(\theta'_{-i})\). The interpretation is that agent \(i\) has a conjecture of others’ true payoff types, given they report \(\theta'_{-i}\). Knowing that the coalition \(S\) with \(i \in S \subseteq I\) has payoff types \(\theta_S\) and reports \(\theta'_S\), for all \(\theta'_{-S} \in \Theta_{-S}\), whenever \(\psi_i(\theta'_{S \setminus \{i\}}, \theta'_{-S})[\cdot]\) has nonzero marginal distribution on \(\theta_{S \setminus \{i\}}\), agent \(i\) updates the conjecture into \(\psi_i(\theta'_{S \setminus \{i\}}, \theta'_{-S}; \theta_S)[\cdot]\) using Bayes’ rule; when \(\psi_i(\theta'_{S \setminus \{i\}}, \theta'_{-S})[\cdot]\) has zero marginal distribution on \(\theta_{S \setminus \{i\}}\), we assign an arbitrary updating rule. We call the above described rule a reasonable updating rule.

For each \(f \in F\), \(S \subseteq I\), and \(\theta'_S \in \Theta_S\), we denote the robust reward set by \(Y^f_{\theta'_S}\), which is the collection of mappings \(y : \Theta \to A\) satisfying the following property. For all
$\theta''_S \in \Theta_S$, there exists $i \in S$ such that $u_i(f(\theta''_S, \theta_{-S}), (\theta''_S, \theta_{-S})) \geq u_i(y(\theta'_S, \theta_{-S}), (\theta''_S, \theta_{-S}))$ for all $\theta_{-S} \in \Theta_{-S}$.

Now we are ready to present the definition of robust coalitional monotonicity.

**Definition 4.4:** A social choice set $F$ satisfies the robust coalitional monotonicity condition if for each $f \in F$ and unacceptable deception $\beta$, there exists $S \subseteq I$, $\theta^*_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta^*_S)$, such that for any reasonable conjecture $(\psi_i)_{i \in S}$ and reasonable updating rule, there exists $y \in Y^f_{\theta^*_S}$ such that for all $i \in S$ and $\theta'_{-S} \in \beta_{-S}(\Theta_{-S})$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta'_S, \theta'_{-S}), (\theta^*_S, \theta_{-S})) \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{-i}] > \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta'_S, \theta'_{-S}), (\theta^*_S, \theta_{-S})) \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{-i}]$$.

If for each $f$ and unacceptable deception $\beta$, there exists a singleton coalition $S \subseteq I$, $\theta^*_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta^*_S)$ such that the requirements in the above condition are satisfied, then the condition becomes the robust monotonicity condition of Bergemann and Morris (2011). It is easy to see that the robust coalitional monotonicity condition is weaker than the robust monotonicity condition.

In the Appendix, Proposition A.2 shows that robust coalitional monotonicity is equivalent to coalitional monotonicity in all type spaces. Proposition A.2 and Lemma 4.1 imply the following result.

**Proposition 4.2:** If a social choice set $F$ is interim strong Nash implementable in all type spaces with payoff environment $\Theta$, then $F$ satisfies the robust coalitional monotonicity condition.

## 5 Sufficient Conditions

The sufficient conditions to robustly strong Nash implement a social choice set $F$ are usually slightly stronger than the necessary conditions. We will impose some additional conditions, and construct a mechanism $(M, g)$ that implements $F$ robustly.

One condition imposed for robust strong Nash implementation is a weak form of ex-post Pareto efficiency. It is weak in the sense that there doesn’t exist an alternative feasible allocation to strictly improve the payoff of every agent. Although the ex-post weak Pareto efficiency condition is natural, it is not necessary for robust strong Nash implementation.
Definition 5.1: A social choice set $F$ is **ex-post weak Pareto efficient** if there does not exist $f \in F$, $\theta \in \Theta$, and $y : \Theta \rightarrow A$ such that $u_i(y(\theta), \theta) > u_i(f(\theta), \theta)$ for all $i \in I$.

5.1 Exchange Economies

As in Section 2, a payoff environment is $\mathcal{E} = \{I, A, (\Theta_i, u_i)_{i=1}^n\}$, where $A$ is the set of all lotteries over a pure feasible outcome set $X$. The set $X$ is abstract in general. However, in an exchange economy with deterministic initial endowments, the set $X$ can be formally defined. Given the initial endowments of all agents $(e_1, ..., e_n) \in \mathbb{R}_+ \times ... \times \mathbb{R}_+^i$, the pure feasible outcome set is $X = \{(x_1, ..., x_n) \in \mathbb{R}_+ \times ... \times \mathbb{R}_+^i | \sum_{i \in I} x_i \leq \sum_{i \in I} e_i\}$. In such an environment, it is without loss of generality to assume that the mechanism designer collects all initial endowments from agents, and then distributes the resources according to the designed mechanism $(M, g)$.

In this environment, we assume that for all $i \in I$ and $\theta \in \Theta$, the random utility function $u_i(\cdot, \theta)$ is continuous and strictly increasing in each component of $x_i$. It is also assumed to be independent of $x_j$ for $j \neq i$, i.e., all goods are private.

We separate this environment from the ones described in Sections 5.2 and 6 because the sufficient conditions in this environment are particularly simple. Furthermore, the mechanism presented here does not rely on lotteries. It should be noted that in the following mechanism, one can replace the integer game with a modular game, as in Jackson (1991).

**Theorem 5.1:** In an exchange economy with deterministic initial endowments and $n \geq 3$, if a social choice set $F$ is robust coalitional incentive compatible, robust coalitional monotonic and ex-post weak Pareto efficient, then there exists a mechanism $(M, g)$ that robustly implements $F$ as an interim strong Nash equilibrium in all full support type spaces.

**Proof.** Consider a mechanism $(M, g)$ where each agent $i$ reports a message in her message space $M_i = \Theta_i \times F \times \mathbb{N}_+ \times (\emptyset \cup \{y : \Theta \rightarrow A\})$. We decompose the message by $m_i = (m_i^1, m_i^2, m_i^3, m_i^4)$. Let $m = (m_1, m_2, ... m_n)$ be a message profile.

The message space $M = \prod_{i \in I} M_i$ is partitioned into $M^1$ to $M^5$ below:

$M^1 = \{m|m_i = (\cdot, f, 0, \emptyset), \forall i \in I\}$,

$M^2(S) = \{m|m_i = (\cdot, f, K, y), \forall i \in S; K \geq 0; y \in Y^{f}_{m_i}, m_j = (\cdot, f, 0, \emptyset), \forall j \notin S\}$,

$M^2 = \bigcup_{\emptyset \subseteq S \subseteq I} M^2(S)$,

$M^3(S) = \{m|m_i = (\cdot, f', 0, \emptyset), \forall i \in S; m_j = (\cdot, f, 0, \emptyset), \forall j \notin S; f \neq f'\}$,

$M^3 = \bigcup_{\emptyset \subseteq S \subseteq I} M^3(S)$.
\[ M^4(S) = \{ m \notin \bigcup_{k=1,2,3} M^k | m_j = (\cdot, f, 0, \emptyset), \forall j \notin S; m_i = (\cdot, f', K, \emptyset), \forall i \in S, \text{ or } m_i = (\cdot, f', K, y), \forall i \in S; K \geq 0; f \text{ and } f' \text{ may be equal} \}, \]

\[ M^4 = \bigcup_{\emptyset \subseteq S \subseteq I} M^4(S), \]

\[ M^5 = M \setminus \bigcup_{k=1,\ldots,4} M^k. \]

The mechanism designer assigns to the agents an allocation according to the message received and the function \( g : M \to A \) below. Agent \( i \) consumes the \( i \)-th component of \( g \).

If \( m \in M^1 \cup M^4(S) \) for nonempty \( S \subseteq I \), the outcome allocation is \( g(m) = f(m^1) \).

If \( m \in M^2(S) \) for nonempty \( S \subseteq I \), the outcome allocation is \( g(m) = y(m^1) \).

For \( m \in M^3 \), if either \( S \) or its complement \( S^c \) is a singleton, let \( i^* \) be the agent with the smallest index in the non-singleton coalition; and otherwise, let \( i^* \) be 1. The outcome function gives all the resources to agent \( i^* \).

If \( m \in M^5 \), among the agents who report the highest integer, let the one with the lowest index win all the resources.

The interpretation of the mechanism is as follows. The message space is partitioned into five basic sets. There exists a truthful equilibrium in \( M^1 \) because of robust coalitional incentive compatibility, the definition of \( M^2(S) \), and the uneven distribution in \( M^3 \cup M^5 \). The robust coalitional monotonicity condition, as well as the integer game instrument, dissolves all potential equilibria outside of \( M^1 \). Therefore, \((M, g)\) implements \( F \) in all full support type spaces. The interpretation is formalized by the following claims.

**Claim 5.1:** Fix an arbitrary type space \( T \) and \( f \in F \), \( \sigma_\cdot^*(t_i) = (\hat{\theta}_i(t_i), f, 0, \emptyset) \) for all \( i \in I \) and \( t_i \in T_i \) constitutes an interim strong Nash equilibrium of \((M, g)\).

**Proof:** We wish to show that for any \( S \subseteq I \), \( t_S \in T_S \), and strategy profile \( \sigma'_S, \sigma^*_S \) is not a profitable deviation from \( \sigma^*_S \).

By ex-post weak Pareto efficiency of the social choice set, the grand coalitional deviation cannot be not profitable. Suppose there exists \( t_{-S} \in T_{-S} \) such that \((\sigma'_S(t_S), \sigma^*_S(t_{-S})) \in M^1 \cup M^4(S) \), then by robust coalitional incentive compatibility, \( \sigma'_S \) is not a profitable deviation. If \((\sigma'_S(t_S), \sigma^*_S(t_{-S})) \in M^2(S) \) for some \( t_{-S} \), the deviation is not profitable by the definition of \( M^2(S) \). Suppose \((\sigma'_S(t_S), \sigma^*_S(t_{-S})) \in M^3 \) for some \( t_{-S} \). With \( |S| > 1 \), the deviation is not profitable for \( S \), as only one agent receives all the resources; with \( |S| = 1 \), the deviator consumes zero and is weakly worse off.\(^7\) Singleton deviations cannot lead to

\(^7\) The notation \( |S| \) means the cardinality of \( S \).
\((\sigma_S^*(t_S), \sigma^*_S(t_{-S})) \in M^3\), and for a non-singleton \(S\), at most one agent is strictly better off. This completes the proof of the claim.

**Claim 5.2:** In an arbitrary full support type space \(T\), if \(\sigma\) is an interim strong Nash equilibrium of the mechanism \((M, g)\), then \(\sigma(t) \in M^1\) for all \(t \in T\).

**Proof:** Decompose agent \(i\)'s strategy \(\sigma_i : T_i \rightarrow M_i\) by \(\sigma_i = (\sigma^1_i, \sigma^2_i, \sigma^3_i, \sigma^4_i)\). Suppose by way of contradiction that there exists \(t \in T\) such that \(\sigma(t) \notin M^1\). Below we show that there exists \(j \in I\) who is strictly better off with the strategy \(\sigma'_j\) defined as \(\sigma'_j(t_j) = (\sigma^1_j(t_j), \sigma^2_j(t_j), 1 + \max_{i \in I} \max_{t_i \in T_i} \sigma^3_i(t_i), \sigma^4_j(t_j))\) and \(\sigma'_j(t'_j) = \sigma_i(t'_j)\) for \(t'_j \neq t_j\). This contradicts the fact that \(\sigma\) is an interim strong Nash equilibrium.

Suppose that there exists \(S \subseteq I\) and \(t \in T\) such that \(\sigma(t) \in M^2(S) \cup M^3(S) \cup M^4(S)\). If \(|S| = 1\), let \(j\) be an agent in \(S^c\) who does not receive all the resources under \(g(\sigma(t))\); if \(1 < |S| < n\), let \(j\) be an agent in \(S\) who does not receive all the resources. Suppose that there does not exist \(S \subseteq I\) such that \(\sigma(t) \in M^2(S) \cup M^3(S) \cup M^4(S)\), let \(j\) be an arbitrary agent who does not receive all the resources.

By the full support condition, agent \(j\) with a type \(t_j\) believes that \(t_{-j}\) occurs with positive probability. In that case, \(j\) is strictly better off by deviating to \(\sigma'_j\). For all \(t'_{-j} \in T_{-j}\), from the choice of \(j\), the pair of outcomes \(g(\sigma(t_j, t'_{-j}))\) and \(g(\sigma'_j(t_j), \sigma_{-j}(t'_{-j}))\) can only be one of the following three cases: (1) they are the same; (2) \(g(\sigma(t_j, t'_{-j}))\) gives agent \(j\) no resources; or (3) \(g(\sigma'_j(t_j), \sigma_{-j}(t'_{-j}))\) gives agent \(j\) all the resources. Thus, deviating with \(\sigma'_j\) at \(t_j\) is weakly better for all \(t'_{-j} \in T_{-j}\). Therefore, we have established that \(\sigma'_j\) is a profitable deviation, and this completes the proof of the claim.

**Claim 5.3:** Let \(\sigma\) be an interim strong Nash equilibrium of \((M, g)\) in an arbitrary full support type space. Define the correspondence \(c\) by \(c(\theta) = \cup_{\{t | \hat{\theta}(t) = \theta\}} \sigma(t)\) for all \(\theta \in \Theta\). Then for any selection \(c \in c\), there exists \(f' \in F\) such that \(g(c(\theta)) = f'(\theta)\) for all \(\theta \in \Theta\).

**Proof:** From the previous claim, there exists \(f \in F\) such that for any \(c \in c\), \(g(c(\theta)) = f(c^1(\theta))\). Suppose by way of contradiction that there exists \(c \in c\), such that no \(f' \in F\) satisfies \(g(c(\theta)) = f'(\theta)\) for all \(\theta \in \Theta\), i.e., \(c^1\) is unacceptable. By the robust coalitional monotonicity condition, there exists \(\theta^*_S, \theta^*_S \in c^1(\theta^*_S)\) and \(y \in Y_{\theta_S}^f\) such that the inequality in Definition 7.3 is satisfied for all \(i \in S\). For all \(i \in S\), Define \(\sigma^i_S \) by \(\sigma^i_S(t_i) = (\sigma^1_i(t_i), \sigma^2_i(t_i), K^*, y)\) for \(t_i\) satisfying \(\hat{\theta}_i(t_i) = \theta^*_S\) and \(\sigma^i(t_i) = \theta^*_i\), and define \(\sigma^i_S(t_i) = \sigma_i(t_i)\) elsewhere; \(K^* \geq 0\). Then for \(t_S \in T_S\) satisfying \(\hat{\theta}_S(t_S) = \theta^*_S\) and \(\sigma^*_S(t_S) = \theta^*_S\), the new message is in \(M^2(S)\). This is strictly profitable for \(S\) in all full support type spaces, a contradiction.
In view of the three claims, we have established that \((M, g)\) implements \(F\) robustly.

One can slightly modify the above proof by removing the second part of the message space and get the following result on implementing a social choice function.

**Corollary 5.1:** In an exchange economy with deterministic initial endowments and \(n \geq 3\), if a social choice function \(f\) is robust coalitional incentive compatible and robust coalitional monotonic, then \(f\) is implementable as an interim strong Nash equilibrium in all full support type spaces.

### 5.2 A General Environment

In a general environment, the goods are not restricted to be private. Therefore, the mechanism designed in Theorem 5.1 no longer works. Below we strengthen the sufficient conditions of Section 5.1 in order to accommodate a general environment.

In Section 5.1 to dissolve unwanted equilibria, the winner of the integer game wins all the resources. Without assuming all goods are private, one may consider Jackson’s (1991) design, where the winner of the modular game becomes the dictator. However, this design cannot be applied to robust strong Nash implementation, because the possibility of common interest may lead to a profitable coalitional deviation.

A condition called the "bad outcome property" is added to prove the sufficiency of the conditions introduced in Section 4. The bad outcome property guarantees that there is an outcome that is bad for all agents, and hence no coalition has an incentive to achieve this outcome.

**Definition 5.2:** A social choice set \(F\) satisfies the **bad outcome property** if there exists \(a \in A\) such that for all \(f \in F, \theta, \theta' \in \Theta,\) and \(i \in I, u_i(f(\theta'), \theta) > u_i(a, \theta).\)

As in Bergemann and Morris (2011), we allow for lotteries. In Theorem 7.3 and Proposition 6.1, the use of lotteries relaxes the restriction that \(n \geq 3\) in Theorem 5.1. It also allows us to relax the full support condition if we focus on implementing a social choice function. With lotteries, the mechanisms can also implement social choice sets under mixed strategies. The mechanism below is an adaptation of Bergemann and Morris (2011).

**Theorem 5.2:** If a social choice set \(F\) satisfies robust coalitional incentive compatibility, robust coalitional monotonicity, ex-post weak Pareto efficiency, and the bad outcome property,
then it is implementable by a mechanism \((M, g)\) as an interim strong Nash equilibrium in all full support type spaces.

**Proof.** In the mechanism, each agent \(i\) reports a message \(m_i = (m_{i1}, m_{i2}, m_{i3}, m_{i4}, m_{i5})\), where \(m_{i1} \in \Theta_i, m_{i2} \in F, m_{i3} \in \mathbb{N}_+, m_{i4} \in \mathbb{N}_+, m_{i5} \in \{y : \Theta \to A\}\). We partition the message space into \(M_1, M_2,\) and \(M_3\) as follows:

\[
M_1 = \{m|m_i = (\cdot, f, 0, 0, \cdot), \forall i \in I\},
\]
\[
M_2(S) = \{m|m_i = (\cdot, f, K_1, K_2, y), \forall i \in S; K_1, K_2 > 0; y \in Y_{m_{i3}}^f; m_j = (\cdot, f, 0, 0, \cdot), \forall j \not\in S\},
\]
\[
M_3 = M\setminus\{M_1 \cup M_2\}.
\]

Let \(\sigma\) be the "bad outcome" in the bad outcome property. Pick any \(f' \in F,\) and let \(\sigma'(m_1)\) be \(f'(m_1)\) with probability \(\epsilon > 0\) and \(\sigma\) with probability \(1 - \epsilon,\) where \(\epsilon\) is sufficiently small such that \(\epsilon u_i(f'(\theta'), \theta) + (1 - \epsilon) u_i(\sigma, \theta) < u_i(f''(\theta'), \theta)\) for all \(\theta, \theta' \in \Theta, i \in I,\) and \(f'' \in F.\)

If \(m \in M_1,\) let the outcome allocation be \(g(m) = f(m_1).\)

If \(m \in M_2(S)\) for \(S \subseteq I,\) let \(g(m)\) be a lottery \(\tilde{g}(m_1),\) which has a realization of \(y(m_1)\) with probability \(K_1/(K_1 + 1),\) \(\sigma'(m_1)\) with probability \(K_2/((K_1 + 1)(K_2 + 1)),\) and \(\sigma\) with probability \(1/((K_1 + 1)(K_2 + 1)).\)

If \(m \in M_3,\) let \(g(m)\) be \(\sigma\) with probability \((\max_{i \in I} m_{i4} + 1)\) and \(\sigma'(m_1)\) with probability \((\max_{i \in I} m_{i4} + 1).\)

It is easy to see that for \(f \in F, \sigma_\ast^i(t_i) = (\hat{\theta}_i(t_i), f, 0, 0, \cdot)\) for all \(i \in I\) and \(t_i \in T_i\) constitutes an interim strong Nash equilibrium in all type spaces.

Now we show that if the strategy profile \(\sigma\) is an interim strong Nash equilibrium of the mechanism \((M, g)\) in an arbitrary type space \(T,\) then for all \(t \in T, \sigma(t) \in M_1.\) Suppose by way of contradiction that there exists \(t \in T\) and \(S \subseteq I\) such that \(\sigma(t) \in M_2(S) \cup M_3.\)

Suppose \(\sigma(t) \in M_2(S)\) for some \(t \in T\) and \(S \subseteq I.\) Assume without loss of generality that there does not exist \(t'' \) and \(S'' \subseteq S\) such that \(\sigma(t'') \in M_2(S'')\), otherwise we can apply the same argument to one largest set \(S''\) (in partial order) and the corresponding \(t''.\) Let agents in \(S\) deviate with \(\sigma_\ast'_S(t_S),\) where for all \(i \in S, \sigma'_i(t) = (\sigma^2_1(t_i), \sigma^2_2(t_i), \sigma^3_2(t_i), 1 + \max_{j \in I} \max_{t_j \in T_J} \sigma^4_3(t_j), \sigma^5_3(t_i))\) and \(\sigma'_i(t'_i) = \sigma_i(t'_i)\) for \(t'_i \neq t_i.\) Notice that \(\sigma(t_S, t'_{-S}) \in M_2(S) \cup M_3\) for all \(t'_{-S} \in T_{-S}.\) The deviation \(\sigma'_S\) leads to a strictly better lottery in \(M_2(S) \cup M_3.\) Hence, \(\sigma\) cannot be an interim strong Nash equilibrium, a contradiction. Suppose \(\sigma(t) \in M_3,\) then the grand coalition can deviate with the strategy profile \(\sigma'\) and obtain a strictly better result in \(M_3,\) a contradiction.
As in Theorem 5.1, we can prove that if $\sigma$ is an interim strong Nash equilibrium of $(M, g)$ in an arbitrary full support type space, then there exists $f' \in F$ such that $g(c(\theta)) = f'(\theta)$ for all $c \in c$ and $\theta \in \Theta$. Suppose not, by the robust coalitional monotonicity condition, there exists $S \subseteq I$, $\theta_S^* \in \Theta_S$, $\theta_S' \in \Theta_S^*\{\theta_S^*\}$, and $y \in Y_{\theta_S'}^f$ such that the inequality in Definition 7.3 holds for all $i \in S$. Let $\sigma''_S$ be a strategy for agents in the coalition $S$, where $\sigma''_i(t_i) = (\sigma_1^i(t_i), \sigma_2^i(t_i), K_1^*, K_2^*, y)$ for all $t_i$ satisfying $\hat{\theta}_i(t_i) = \theta_i^*$ and $\sigma_1^i(t_i) = \theta_i'$, and $\sigma''_i(t_i) = \sigma_i(t_i)$ elsewhere. We require the above $K^*_1$ to be sufficiently large and $K^*_2 > 0$. Then for any $t_S \in T_S$ satisfying $\hat{\theta}_S(t_S) = \theta_S^*$ and $\sigma^1_S(t_S) = \theta_S'$, the new message is in $M^2(S)$, bringing coalition $S$ a strictly better outcome. Therefore, $\sigma''_S$ is a profitable coalitional deviation, a contradiction.

6 Random Initial Endowments

In this section, we relax the assumption that agents’ initial endowments in an exchange economy are deterministic. This results in a random feasible outcome set and adds difficulties compared to Section 5. Therefore, in this section, one needs to strengthen the sufficient conditions of Theorem 5.1.

We assume that agents’ initial endowments are private payoff type relevant and nonzero, i.e., agent $i$ has an initial endowment $e_i(\theta_i) \in \mathbb{R}^l_+ \setminus \{0\}$ when her payoff type is $\theta_i$. As an agent’s initial endowment is part of her private information, the mechanism designer can no longer collect all agents’ initial endowments. Instead, we let the mechanism designer control the net trade between agents.

Under the payoff type $\theta$, the set of pure feasible net trades $X(\theta)$ is defined by $X(\theta) = \{z(\theta) \in \mathbb{R}^l \times \ldots \times \mathbb{R}^l | \forall i \in I, z_i(\theta) + e_i(\theta_i) \in \mathbb{R}^l_+, \text{ and } \sum_{i \in I} z_i(\theta) \leq 0\}$. All lotteries over $X(\theta)$ are denoted by $A(\theta)$. To guarantee that the mechanism $(M, g)$ is feasible, we follow Hurwicz et al. (1994) and require each agent $i \in I$ to place her initial endowment $e_i(\theta_i)$ at the table when she reports $\theta_i \in \Theta_i$. Hence, agents can never have deceptions involving an over-report in initial endowments. Nevertheless, an under-report may occur. This restriction needs to be taken into consideration in the definitions of robust coalitional incentive compatibility and robust coalitional monotonicity with random initial endowments. However, for consistency of notation we do not explicitly state it.

Under such an environment, we interpret a feasible net trade plan $z : \Theta \to A$ as a mapping from $\Theta$ to $\bigcup_{\theta \in \Theta} A(\theta)$ satisfying $z(\theta) \in A(\theta)$ for all $\theta \in \Theta$. Social choice functions
are socially desired net trade plans. When the payoff type profile is \( \theta \), the initial endowments of all agents are \( e \), and the realized net trade is \( \bar{z} \in A(\theta) \), agent \( i \) consumes \( e_i(\theta_i) \) and the \( i \)-th component of \( \bar{z} \). Under a random utility function \( \bar{u}_i(\cdot, \cdot) \), agent \( i \)'s ex-post utility level is \( \bar{u}_i(e(\theta) + \bar{z}, \theta) \). We denote \( \bar{u}_i(e(\theta) + \bar{z}, \theta) \) by \( u_i(\bar{z}, \theta) \), and therefore the conditions of incentive compatibility, monotonicity, and ex-post weak Pareto efficiency are consistent with those in Sections 4 and 5.

With random initial endowments, we interpret the robust reward set \( Y^f_{\Theta} \) as the collection of mappings \( y : \Theta \rightarrow A \) satisfying the following property. Whenever \( \theta''_S \in \Theta_S \) satisfies \( e_i(\theta''_S) \geq e_i(\theta'_S) \) for all \( i \in S \), there exists \( i \in S \) such that \( u_i(f(\theta''_S, \theta_{-S}), (\theta''_S, \theta_{-S})) \geq u_i(f(\theta'_S, \theta_{-S}), (\theta'_S, \theta_{-S})) \) for all \( \theta_{-S} \in \Theta_{-S} \).

The following definition is the familiar condition of ex-post individual rationality, which requires that the social choice set is better than no trade for every agent. It is weaker than the bad outcome property but plays a similar role in the proof. In the definition, \( \Theta_{I \times n} \) represents no trade.

**Definition 6.1:** A social choice set \( F \) satisfies the **ex-post individual rationality** condition if for all \( f \in F, i \in I \), and \( \theta \in \Theta \), \( u_i(f(\theta), \theta) \geq u_i(\Theta_{I \times n}, \theta) \).

Below we present a sufficiency theorem on robust strong Nash implementation in exchange economies with random initial endowments.

**Proposition 6.1:** In an exchange economy with random initial endowments, if a social choice set \( F \) satisfies robust coalitional incentive compatibility, robust coalitional monotonicity, ex-post weak Pareto efficiency, and ex-post individual rationality, then there exists a mechanism \( (M, g) \) that robustly implements \( F \) as an interim strong Nash equilibrium in all full support type spaces.

**Proof.** Each agent \( i \) reports a message \( m_i = (m^1_i, m^2_i, m^3_i, m^4_i) \), where \( m^1_i \in \Theta_i, m^2_i \in F, m^3_i \in N_+, m^4_i \in \{ y : \Theta \rightarrow A \} \). We partition the message space into \( M^1, M^2, \) and \( M^3 \) below:

\[
M^1 = \{ m | m_i = (\cdot, f, 0, \cdot), \forall i \in I \},
\]

\[
M^2(S) = \{ m | \forall i \in S, m_i = (\cdot, f, K, y); K > 0; y \in Y^f_{m^2_S}; e_i(m^1_i) + y_i(m^1_S, \theta_{-S}) \in \mathbb{R}_+ \setminus \{ 0 \}, \forall i \in S \text{ and } \theta_{-S} \in \Theta_{-S}; m_j = (\cdot, f, 0, \cdot), \forall j \notin S \},
\]

\[
M^2 = \bigcup_{\emptyset \neq S \subseteq I} M^2(S),
\]

\[
M^3 = M \setminus \{ M^1 \cup M^2 \}.
\]

If \( m \in M^1 \), let the net trade be \( g(m) = f(m^1) \).
Remark 6.1: If \( m \in M^2(S) \) for some \( S \subseteq I \), let \( g(m) \) be \( y(m^1) \) with probability \( K/(K + 1) \) and \(-e(m^1) \) with probability \( 1/(K + 1) \).

If \( m \in M^3 \), let \( g_i(m) = -e_i(m_i^1)/(1 + m_i^3) \) for all \( i \in I \).

The robust coalitional incentive compatibility condition, the definition of \( M^2(S) \), and the ex-post individual rationality condition imply that for each \( f \in F \), \( \sigma_i^r(t_i) = (\hat{\theta}_i(t_i), f, 0, \cdot) \) for all \( i \in I \) and \( t_i \in T_i \) constitutes an interim strong Nash equilibrium in all type spaces.

Let \( \sigma \) be an interim strong Nash equilibrium in an arbitrary full support type space. As in Theorem 7.3, we know \( c(\theta) \in M^1 \) for all \( \theta \in \Theta \) and \( c \in c \), i.e., agents agree on a function \( f \in F \). Furthermore, \( c^1 \) is an acceptable deception. Otherwise, by the robust coalitional monotonicity condition, there exists \( S \subseteq I, \theta^*_S \in \Theta_S, \theta'_S \in c^1_S(\theta^*_S) \), and \( y \in Y^f_\theta \) such that the strict inequality in Definition 7.3 holds for all \( i \in S \).

We can assume without loss of generality that \( \forall i \in S \) and \( \theta_S \in \Theta_S \), \( e_i(c^1_S(\theta^*_S), \theta_S) \in \mathbb{R}^l_+ \setminus \{0\} \). To see this, if there exists \( i \in S \) and \( \theta_S \in \Theta_S \) such that \( e_i(c^1_S(\theta^*_S)) + y_i(c^1_S(\theta^*_S), \theta_S) = 0 \), we fix the \( \theta_S \) for the argument. Let \( \hat{y}_i(c^1_S(\theta^*_S), \theta_S) = y_i(c^1_S(\theta^*_S), \theta_S) + \eta_i \) for all \( i \in S \) with \( e_i(c^1_S(\theta^*_S)) + y_i(c^1_S(\theta^*_S), \theta_S) = 0 \), where \( \eta_i \in \mathbb{R}^l_+ \setminus \{0\} \).

Let \( \hat{y}_i(c^1_S(\theta^*_S), \theta_S) = y_i(c^1_S(\theta^*_S), \theta_S) + \eta_i \in \mathbb{R}^l_+ \) for all other \( i \in I \), where \( \eta_i \in \mathbb{R}^l_- \). Elsewhere let \( \hat{y}(\cdot) = y(\cdot) \). When all arguments of each of the vectors \( (\eta_i)_{i \in I} \) are sufficiently close to zero and \( \sum_{i \in I} \{y_i(c^1_S(\theta^*_S), \theta_S) + \eta_i \} \leq 0 \), \( \hat{y} \) is feasible. Then we have \( e_i(c^1_S(\theta^*_S)) + \hat{y}_i(c^1_S(\theta^*_S), \theta_S) \in \mathbb{R}^l_+ \setminus \{0\} \) for all \( i \in S \). Next we must prove that \( \hat{y} \in Y^f_{m^S} \).

From ex-post individual rationality of \( F \) and nonzero initial endowments, for all \( i \in S \) with \( e_i(c^1_S(\theta^*_S)) + y_i(c^1_S(\theta^*_S), \theta_S) = 0 \), \( u_i(f(\theta^*_S, \theta_S), (\theta^*_S, \theta_S)) \geq u_i(0_{1 \times n}, (\theta^*_S, \theta_S)) > u_i(-e(c^1_S(\theta^*_S), \theta_S), (\theta^*_S, \theta_S)) = u_i(y_i(c^1_S(\theta^*_S), \theta_S), (\theta^*_S, \theta_S)) \).

Hence, \( u_i(f(\theta^*_S, \theta_S), (\theta^*_S, \theta_S)) \geq u_i(\hat{y}(c^1_S(\theta^*_S), \theta_S), (\theta^*_S, \theta_S)) \) for all such \( i \in S \) and \( \eta_i \) sufficiently close to \( 0 \). For other \( i \in I \), we notice that \( \hat{y}(c^1_S(\theta^*_S), \theta_S) \leq y(c^1_S(\theta^*_S), \theta_S) \). Therefore, we have established that \( \hat{y} \in Y^f_{m^S} \).

Furthermore, as \( y \) satisfies the strict inequality in Definition 7.3 for all \( i \in S \), the strict inequality also holds for \( \hat{y} \) when all arguments of each of the vectors \( (\eta_i)_{i \in I} \) are sufficiently close to zero. If necessary, one can repeatedly apply this argument, and renew \( \hat{y} \) such that (1) all requirements in Definition 7.3 are satisfied, and (2) for all \( i \in S \) and \( \theta_S \in \Theta_S \), \( e_i(c^1_S(\theta^*_S)) + \hat{y}_i(c^1_S(\theta^*_S), \theta_S) \in \mathbb{R}^l_+ \setminus \{0\} \). We thereby can replace \( y \) with \( \hat{y} \).

Define for all \( i \in S \) a deviating strategy \( \sigma_i^r \) as in Theorem 5.1. Let \( K^* \) be sufficiently large. Then for \( t_S \in T_S \) satisfying \( \hat{\theta}_S(t_S) = \theta^*_S \) and \( \sigma^r_{S}(t_S) = \theta'_S, \sigma^r_{S} \) is strictly profitable in all full support type spaces, a contradiction. 

\[ \square \]

Remark 6.1: There are two ways to relax the full support assumption in Theorem 7.3 and
Proposition 6.1. First, one can focus on implementing a social choice function. Second, one can introduce a closure condition as in the Bayesian implementation literature, see Postlewaite and Schmeidler (1986) for example.

7 Extension: Other Coalitional Patterns

The notion of strong Nash equilibrium assumes that all agents can communicate freely with each other, and therefore all coalitions are allowed to be formed. In reality, it might be the case that some coalitions can be formed while some cannot. This could be a result of many factors, for instance, culture differences, language barriers, or geographic isolation.

We demonstrate in this section that our results in Sections 4 and 5 can be extended to environments with other coalition patterns. In the first extension, we assume that the coalition pattern is arbitrary but known to the mechanism designer. In the second extension, the mechanism designer does not know which coalitions could be formed.

These problems are studied under complete information by Suh (1996, 1997). We extend these problems to incomplete information environments and belief-free implementation.

7.1 Implementation under Known Coalition Pattern

Let $S$ be a collection of coalitions $S \subseteq I$, representing the set of all coalitions that can be formed. We assume that all singletons of $I$ are included in $S$, i.e., agents can always choose not to communicate with others, which implies that the following equilibrium notion is a section of Interim Nash equilibrium.

Definition 7.1: In a type space $\mathcal{T}$, the strategy profile $\sigma^*$ is an interim $S$-Nash equilibrium of the mechanism $(M, g)$ if there does not exist $S \in S$, $t^*_S \in T_S$, and $\sigma^*_S : T_S \to M_S$, such that for all $i \in S$,

\[
\sum_{t_{-i} \in T_{-i}} u_i\left(g\left(\sigma^*_S(t^*_S), \sigma^*_{-S}(t_{-S})\right), \hat{\theta}(t^*_S, t_{-S})\right)\pi_i(t^*_S)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i\left(g\left(\sigma^*(t^*_S, t_{-S})\right), \hat{\theta}(t^*_S, t_{-S})\right)\pi_i(t^*_S)[t_{-i}].
\]

If there exists a mechanism $(M, g)$ that fully implements a social choice set $F$ as an interim $S$ Nash equilibrium in all type spaces with payoff environment $\Theta$, then $F$ is said to be robustly $S$-Nash implementable.
We present the following two conditions for robust $S$ Nash implementation.

**Definition 7.2:** A social choice set $F$ is **robust $S$ incentive compatible** if for all $f \in F$, $S \in S$, and $\theta'_S \neq \theta^*_S$, there exists $i \in S$ such that

$$u_i\left(f(\theta^*_S, \theta_i), (\theta^*_S, \theta_i)\right) \geq u_i\left(f(\theta'_S, \theta_i), (\theta^*_S, \theta_i)\right) \text{ for all } \theta_i \in \Theta_i.$$

**Definition 7.3:** A social choice set $F$ satisfies the **robust $S$ monotonicity** condition if for each $f \in F$ and unacceptable deception $\beta$, there exists $S \in S$, $\theta^*_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta^*_S)$, such that for any reasonable conjecture $(\psi_i)_{i \in S}$ and reasonable updating rule, there exists $y \in Y^f_{\theta^*_S}$ such that for all $i \in S$ and $\theta'_i \in \beta_S(\Theta_i)$,

$$\sum_{\theta_i \in \Theta_i} u_i\left(y(\theta'_S, \theta'_i), (\theta^*_S, \theta_i)\right) \psi_i(\theta'_i; \theta^*_S) \{\theta_i\} \theta'_i \geq \sum_{\theta_i \in \Theta_i} u_i\left(f(\theta'_S, \theta'_i), (\theta^*_S, \theta_i)\right) \psi_i(\theta'_i; \theta^*_S) \{\theta_i\} \theta'_i.$$

We claim that the two conditions are necessary for implementing a social choice set $F$. The proofs are natural extensions of the propositions in Section 4. Subsequently, we provide a sufficiency result on implementing a social choice function $f$.

**Corollary 7.1:** If a social choice function $f$ satisfies robust $S$ incentive compatibility, robust $S$ monotonicity, and the bad outcome property, then it is robustly $S$ implementable by a mechanism $(M, g)$.

**Proof.** The mechanism is an adaptation of the one in Theorem 7.3. As we focus on implementation of a social choice function, we remove the second component of the mechanism, i.e., each agent $i$ reports a message $m_i = (m^1_i, m^3_i, m^4_i, m^5_i)$, where $m^1_i \in \Theta_i$, $m^3_i \in N_+$, $m^4_i \in N_+$, $m^5_i \in \{y : \Theta \rightarrow A\}$. The message space partition is slightly differs from the original one. The set $M^1$ is the same with the original partition; the set $M^2$ is different: $M^2 = \bigcup_{S \in S} M^2(S)$, where for all $S \subseteq I$, $M^2(S)$ is defined in the same way as in Theorem 7.3, the set $M^3$ contains all messages not in $M^1 \cup M^2$.

The allocation rule and the remaining of the proof mostly follow those in Theorem 7.3 except for modifying Claim 5.2 into the following version.

**Claim 7.1:** In an arbitrary type space $\mathcal{T}$, if $\sigma$ is an $S$ Nash equilibrium of the mechanism $(M, g)$, then $\sigma(t) \in M^1$ for all $t \in T$.

**Proof:** Define a deviating message $\sigma'_j$ in the same way as in Claim 5.2 except for ignoring the second component of the message in Claim 5.2.
Suppose that there exists $S \subseteq I$ and $t \in T$ such that $\sigma(t) \in M^2 \cup M^3$. This implies that there is an agent $j \in I$ with type $t_j$ such that $\sigma_j^3(t_j) > 0$. If $j$ deviates with strategy $\sigma'_j$, for all $t_{-j} \in T_{-j}$, the message either leads to a strictly better lottery in $M^2$ or a strictly better lottery in $M^3$. This contradicts with the assumption that $\sigma$ is an $S$ Nash equilibrium. 

\[\square\]

### 7.2 Double Implementation

The mechanism designer may not have any information on which coalitions could be formed. In another word, the mechanism designer does not know what the real $S$ is, except that $S$ contains all singletons of $I$.

This framework addresses another layer of uncertainty, i.e., the uncertainty of coalition patterns, in addition to uncertainty of the type spaces. If the mechanism designer wishes to guarantee a desirable outcome regardless of the coalition patterns, the following implementation concept can be adopted.

If there exists a mechanism $(M, g)$ that robustly $S$ implements a social choice set $F$ for all $S \subseteq 2^I$ satisfying $\{i\} \in S$ for all $i \in I$, then $F$ is said to be **robustly double implementable**.

The above notion is called double implementation because it is equivalent to robust Nash and simultaneously robust strong Nash implementation of $F$. This comes from the fact that there are two extreme cases of $S$. The minimal $S$ only contains all singletons of $I$, and this $S$ implementation corresponds to robust implementation of Bergemann and Morris (2011). The maximal $S$ equals $2^I$, and corresponds to robust strong Nash implementation of this paper. If there exists a mechanism $(M, g)$ that robustly implements $F$ as an interim Nash and strong Nash equilibrium, then the set of interim Nash equilibrium outcomes equals the set of interim strong Nash equilibrium outcomes, which further equals outcomes consistent with social choice set $F$. For an arbitrary $S$, the set of all interim $S$ Nash equilibria is a subset of all interim Nash equilibria, and is a superset of all interim strong Nash equilibria. Hence, when $F$ is robustly Nash and simultaneously strong Nash implemented by $(M, g)$, $F$ is robustly $S$ Nash implemented for all coalition pattern $S$. The other direction of the equivalence is obvious.

We claim that robust coalitional incentive compatibility and robust monotonicity condition (a social choice set version of this condition is defined by restricting $S$ to contain only singletons of $I$) are necessary for robust doublem implementation of $F$. 

21
A sufficiency result is stated below.

**Corollary 7.2:** If a social choice function $f$ satisfies robust coalitional incentive compatibility, robust monotonicity, and the bad outcome property, then it is robustly double implemented by a mechanism $(M, g)$.

*Proof.* The mechanism is an adaptation of the one in Theorem 7.3 by removing the second component of the original mechanism.

The allocation rule and the remaining of the proof mostly follow those in Theorem 7.3 except that the proof of Claim 5.2 needs to be modified in the same way as in Claim 7.1.

**7.3 Social Choice Sets**

Although the necessity results in Sections 7.1 and 7.2 take into account social choice sets, the sufficiency results only look at implementation of a social choice function. This is because in general, sufficiency results of fully implementation of social choice sets need an extra condition, the full support condition.

In Bayesian implementation literature, the condition is assumed by works including Postlewaite and Schmeidler (1986), Palfrey and Srivastava (1986, 1989), and Jackson (1991). Bergemann and Morris (2005) also has a footnote, "The full support assumption is crucial when we look at full implementation..."

Bergemann and Morris (2011) does not assume the full support condition. However, that paper look at full implementation of a social choice function, rather than a set. Theorem 7.3 of the current paper avoids this condition, because we consider robust strong Nash equilibrium and allows for grand coalition.

As an illustration, we provide a sufficiency result on robust double implementation of social choice set $F$. Sufficient conditions robust $S$ can be obtained in a similar way.

**Corollary 7.3:** If a social choice set $F$ satisfies robust coalitional incentive compatibility, robust monotonicity, local ex-post weak Pareto efficiency, robust closure, and the bad outcome property, then it is double implemented by a mechanism $(M, g)$ in all full support type spaces.

---

8Jackson (1991) does not assume full support explicitly. However, in his definition of full implementation, equilibria outcomes and social choice outcomes only need to coincide at type profiles occurring with strictly positive probability.
Proof. The mechanism is identical to the one in Theorem 7.3. The proof of this corollary follows that theorem, except that the proof of Claim 5.2 needs the following modification.

**Claim 7.2:** In an arbitrary type space $T$ with full support, if $\sigma$ is simultaneously an interim Nash and strong Nash equilibrium, then $\sigma(t) \in M^1$ for all $t \in T$.

**Proof:** Define a deviating message $\sigma'_j$ in the same way as in Claim 5.2.

Suppose that there exists $S \subseteq I$ and $t \in T$ such that $\sigma(t) \in M^2$. The argument follows the original proof.

Suppose that there exist $t \in T$ such that $\sigma(t) \in M^3$. Any agent $i \in I$ has an incentive to deviate with $\sigma'_i$ under the full support assumption. To see this, let $\tilde{t}_{-i}$ be an arbitrary element in $T_{-i}$. If $\sigma(t_i, \tilde{t}_{-i}) \in M^1$, $(\sigma'_i(t_i), \sigma_{-i}(\tilde{t}_{-i})) \in M^1$, and thus deviating does not change the outcome. If $\sigma(t_i, \tilde{t}_{-i}) \in M^2$, $(\sigma'_i(t_i), \sigma_{-i}(\tilde{t}_{-i})) \in M^2$ leads to a strictly better lottery in $M^2$. If $\sigma(t_i, \tilde{t}_{-i}) \in M^3$, $(\sigma'_i(t_i), \sigma_{-i}(\tilde{t}_{-i})) \in M^3$ leads to a strictly better lottery in $M^3$. Recall that the last case occurs with positive probability because of the full support assumption and the supposition that $\sigma(t) \in M^3$. Therefore, $\sigma'_i$ is a profitable deviation in all full support type spaces.

Both cases contradicts the fact that $\sigma$ is interim strong Nash equilibrium. This completes the proof of the claim.

8 Applications

8.1 Fine Core

In this section, we provide applications of our implementation results. We start with a social choice set that is robustly strong Nash implementable in an exchange economy described in Section 6. We present the definition of the fine core and show that the net trade associated with the fine core is implementable under some conditions.

For each agent $i \in I$, define $\mathcal{F}_i = \{ \bigcup_{t_{-i} \in T_{-i}} (\theta_{t_{-i}}) | \theta_i \in \Theta_i \}$, which represents the private payoff information partition of agent $i$. We denote by $\sigma(\mathcal{F}_i)$ the $\sigma$-algebra generated by $\mathcal{F}_i$. Define $\sigma(\mathcal{F}_S) = \bigvee_{i \in S} \sigma(\mathcal{F}_i)$, which is the pooled information of the coalition $S$. Recall in Section 6 we defined that $u_i(\bar{z}, \theta) := \bar{u}_i(e, \theta) + \bar{z}, \theta)$ and that a social choice function $f : \Theta \rightarrow A$ is a feasible net trade plan.
Below we define the fine core (Yannelis, 1991).

**Definition 8.1:** In an exchange economy with a type space \( T \) and random initial endowments \( e \), the fine core is the set of all allocations \( e + f : \Theta \to \mathbb{R}_+^I \times \cdots \times \mathbb{R}_+^I \) satisfying:

1. for all \( i \in I \) and \( \theta \in \Theta \), \( f_i \) is \( \sigma(\mathcal{F}_i) \)-measurable and \( \sum_{i \in I} f_i(\theta) \leq 0 \),
2. there does not exist \( S \subseteq I \), \( t^*_S \in T_S \), and a net trade plan \( y : \Theta \to \mathbb{R}_+^I \times \cdots \times \mathbb{R}_+^I \) such that:
   2.1 \( y_i : \Theta \to \mathbb{R}_+^I \) is \( \sigma(\mathcal{F}_S) \)-measurable for all \( i \in S \), and \( y_j(\theta) = 0 \) for all \( j \not\in S \) and \( \theta \in \Theta \),
   2.2 \( \sum_{i \in S} y_i(\theta) \leq 0 \) and \( e_i(\theta) + y_i(\theta) \in \mathbb{R}_+^I \) for all \( \theta \in \Theta \) and \( i \in S \),
   2.3 \( \sum_{t_{-i} \in T_{-i}} u_i(y(\hat{\theta}(t^*_S, t_{-S})), \hat{\theta}(t^*_S, t_{-S})) \pi_i(t^*_S)[t_{-i}] \)
   \[ > \sum_{t_{-i} \in T_{-i}} u_i(f(\beta(\hat{\theta}(t^*_S, t_{-S}))), \hat{\theta}(t^*_S, t_{-S})) \pi_i(t^*_S)[t_{-i}] \]
   for all \( i \in S \).

Given the initial endowments \( e \), the set of all net trade plans \( f : \Theta \to \mathbb{R}_+^I \times \cdots \times \mathbb{R}_+^I \) satisfying the above conditions is called the fine core net trade, and it is denoted by \( F \).

In order to obtain the following result, we assume that each \( \hat{u}_i(\cdot, \theta) \) is \( \sigma(\mathcal{F}_i) \)-measurable in addition to the assumptions in Section 6. This implies that \( u_i(\cdot, \theta) \) is \( \sigma(\mathcal{F}_i) \)-measurable.

**Proposition 8.1:** Under the assumptions imposed in this section, the fine core net trade \( F \) is implementable as an interim strong Nash equilibrium in all full support type spaces.

**Proof.** In view of Proposition 6.1, it suffices to prove that \( F \) is robust coalitional incentive compatible and robust coalitional monotonic, as the other two conditions can be easily verified.

To check the robust coalitional monotonicity condition, suppose \( f \in F \), but there exists an unacceptable deception \( \beta \) such that \( f \circ \beta \not\in F \) for some \( \beta \in \beta \). Then there exists \( S \subseteq I \), \( t^*_S \in T_S \), and \( y : \Theta \to \mathbb{R}_+^I \times \cdots \times \mathbb{R}_+^I \) such that:

1. \( y_i : \Theta \to \mathbb{R}_+^I \) is \( \sigma(\mathcal{F}_S) \)-measurable for all \( i \in S \), and \( y_j(\theta) = 0 \) for all \( j \not\in S \) and \( \theta \in \Theta \),
2. \( \sum_{i \in S} y_i(\theta) \leq 0 \) and \( e_i(\theta) + y_i(\theta) \in \mathbb{R}_+^I \) for all \( \theta \in \Theta \) and \( i \in S \),
3. \( \sum_{t_{-i} \in T_{-i}} u_i(y(\hat{\theta}(t^*_S, t_{-S})), \hat{\theta}(t^*_S, t_{-S})) \pi_i(t^*_S)[t_{-i}] \)
   \[ > \sum_{t_{-i} \in T_{-i}} u_i(f(\beta(\hat{\theta}(t^*_S, t_{-S}))), \hat{\theta}(t^*_S, t_{-S})) \pi_i(t^*_S)[t_{-i}] \]
   for all \( i \in S \).

For all \( i \in I \) and \( \theta \in \Theta \), define a net trade plan \( \tilde{y} \) as follows, for all \( i \in I \),

\[
\tilde{y}_i(\theta_S, t_{-S}) = \begin{cases} 
  y_i(\hat{\theta}_S(t^*_S), \theta_{-S}) & \text{if } i \in S \text{ and } \theta_S = \beta_S(\hat{\theta}_S(t^*_S)), \\
  0 & \text{otherwise.}
\end{cases}
\]

By the \( \sigma(\mathcal{F}_i) \)-measurability of \( u_i \), we write \( u_i(\cdot, \theta_S) \) instead of \( u_i(\cdot, \theta) \). The measurability
conditions of \( f_i \) and \( y_i \) imply that for all \( i \in S \) and \( \theta'_S \in \Theta_S \),
\[
    u_i\left(\tilde{y}(\beta_S(\hat{\theta}_S(t^*_S)), \theta'_S), \hat{\theta}_i(t^*_i)\right) = u_i\left(y(\hat{\theta}_S(t^*_S), \theta'_S), \hat{\theta}_i(t^*_i)\right) > u_i\left(f(\beta_S(\hat{\theta}_S(t^*_S)), \theta'_S), \hat{\theta}_i(t^*_i)\right).
\]
Hence, we have established the strict inequality in the robust coalition monotonicity condition.

For \( \theta'^S \in \Theta_S \) with \( \left( e_i(\theta'_i) \geq e_i(\beta_i(\hat{\theta}_i(t^*_i))) \right)_{i \in S} \), we have \( e_i(\theta'_i) + \tilde{y}_i(\beta_S(\hat{\theta}_S(t^*_S)), \theta_S) \geq e_i(\beta_i(\hat{\theta}_i(t^*_i))) + \tilde{y}_i(\beta_S(\hat{\theta}_S(t^*_S)), \theta_S) \in \mathbb{R}_+ \) for all \( i \in S \) and \( \theta_S \in \Theta_S \). Also notice that \( \sum_{i \in S} \tilde{y}_i(\beta_S(\hat{\theta}_S(t^*_S)), \theta_S) \leq 0 \) for all \( \theta_S \in \Theta_S \). From the fact that \( f \) is a net trade plan associated with the fine core as well as the measurability conditions of \( f_i \) and \( \tilde{y}_S \), we have
\[
    u_i\left(f(\theta'_S, \theta_S), \theta'_i\right) \geq u_i\left(\tilde{y}(\beta_S(\hat{\theta}_S(t^*_S)), \theta_S), \theta'_i\right)
\]
for some \( i \in S \) and all \( \theta_S \in \Theta_S \). Hence, we have established the robust coalition monotonicity condition.

To check the robust coallitional incentive compatibility condition, first notice that there is no profitable grand coalitional deviation because each \( f \) is ex-post weak Pareto efficient. Suppose by way of contradiction that there exists \( S \not\subseteq I \), \( f \in F \), and \( \theta'^S, \theta'^S \in \Theta_S \) such that for all \( i \in S \), \( u_i(f(\theta'^S, \theta_S), \theta^*_i) > u_i(f(\theta'^S, \theta_S), \theta^*_i) \) for all \( \theta_S \in \Theta_S \). Note that the private information measurability condition of \( f \) is used in the above hypothesis. For all \( j \not\in S \), \( f_j \) is \( \sigma(F_j) \)-measurable, and therefore \( u_j(f(\theta'^S, \theta_S), \theta^*_S) = u_j(f(\theta'^S, \theta_S), \theta^*_S) \) for all \( \theta_S \in \Theta_S \). Fix one \( \theta_S \in \Theta_S \). For all \( i \in I \), let \( \mathbf{e}_i \in \mathbb{R}_+ \setminus \{0\} \) satisfy \( \sum_{i \in S} \mathbf{e}_i \leq \sum_{i \in S} \mathbf{e}_i \). Let \( g_i(\theta'^S, \theta_S) = f_i(\theta'^S, \theta_S) - \mathbf{e}_i \) for all \( i \in S \) and \( g_j(\theta'^S, \theta_S) = f_j(\theta'^S, \theta_S) + \mathbf{e}_j \) for all \( j \not\in S \). Let \( g(\cdot) \) be \( 0 \times n \) at other payoff type profiles. If all elements of each of the vectors \( (\mathbf{e}_i)_{i \in I} \) are small enough, then \( g \) is a feasible net trade plan. By continuity and monotonicity of preferences, \( g(\theta'^S, \theta_S) \) makes all agents strictly better off than \( f(\theta'^S, \theta_S) \) at \( (\theta'^S, \theta_S) \), which contradicts the supposition that \( f \) is ex-post weak Pareto efficient.

\( \square \)

### 8.2 Allocating a Private Good

Suppose a benevolent social planner aims to efficiently allocate an indivisible private good to two agents, whose evaluation of the good is private information. We allow for transfers in this problem. Assume that the type space is not common knowledge and that random utility functions are of the form \( u_i((q(\theta), t(\theta)), \theta_i) = q_i(\theta)\theta_i + t_i(\theta) \).

Suppose \( \Theta_1 = \{1, 2\} \) and \( \Theta_2 = \{2, 3\} \). The efficient allocation rule, namely the provision probability, \( q^a(\cdot) \) is given by \( q^a(1, 2) = q^a(1, 3) = q^a(2, 3) = (0, 1) \) and \( q^a(2, 2) = (a, 1-a) \) with \( a \in [0, 1] \). A transfer rule \( t^a(\cdot) \) is given by \( t^a(1, 2) = t^a(1, 3) = t^a(2, 3) = (1, -1) \) and \( t^a(2, 2) = (1 - 2a, 2a - 1) \). A social choice function \( f^a \) is defined as \( f^a(\theta) = (q^a(\theta), t^a(\theta)) \) for all \( \theta \in \Theta \). The social choice set \( F \) is given by \( F = \{f^a : a \in [0, 1]\} \).

25
We apply Theorem 7.3 to show that $F$ is interim strong Nash implementable in all full support type spaces. It suffices to show that $F$ satisfies robust coalitional incentive compatibility and robust coalitional monotonicity. The remaining conditions are easy to verify, as we allow for transfers.

To check the robust coalitional incentive compatibility condition, by ex-post weak Pareto efficiency of each $f^a$, we just need to consider singleton deviations. Notice that for each $a$, $f^a(1, 2) = f^a(1, 3) = f^a(2, 3)$, and hence it suffices to check the following inequalities:

$$u_1(f^a(1, 2), 1) = 1 \geq u_1(f^a(2, 2), 1) = 1 - a,$$
$$u_1(f^a(2, 2), 2) = 1 \geq u_1(f^a(1, 2), 2) = 1,$$
$$u_2(f^a(2, 2), 2) = 1 \geq u_2(f^a(2, 3), 2) = 1,$$
$$u_2(f^a(2, 3), 3) = 2 \geq u_2(f^a(2, 2), 3) = 2 - a.$$

We now check the robust coalitional monotonicity condition. Suppose $\beta : \Theta \rightarrow 2^\Theta \setminus \emptyset$ is unacceptable for some $f^a$. It must be the case that $a > 0$ and $(2, 2) \in \beta(1, 2) \cup \beta(1, 3) \cup \beta(2, 3)$. Suppose $(2, 2) \in \beta(1, 2)$. Let $S = \{1, 2\}$, $\theta^* = (1, 2)$, $\theta' = (2, 2)$, and $y : \Theta \rightarrow A$ be any feasible plan satisfying $y_1(\theta') = (0, 1 - 0.5a)$ and $y_2(\theta') = (1, 0.5a - 1)$. Then

$$u_1(y(\theta'), 1) = 1 - 0.5a > u_1(f^a(\theta'), 1) = 1 - a,$$
$$u_2(y(\theta'), 2) = 1 + 0.5a > u_2(f^a(\theta'), 2) = 1,$$
$$u_1(f^a(\theta), \theta_1) = 1 \geq u_1(y(\theta'), \theta_1) = 1 - 0.5a \text{ for all } \theta \in \Theta.$$

Suppose $(2, 2) \in \beta(1, 3)$. Let $S = \{1, 2\}$, $\theta^* = (1, 3)$, $\theta' = (2, 2)$, and $y : \Theta \rightarrow A$ be any feasible plan satisfying $y_1(\theta') = (0, 1)$ and $y_2(\theta') = (1, -1)$. Then

$$u_1(y(\theta'), 1) = 1 > u_1(f^a(\theta'), 1) = 1 - a,$$
$$u_2(y(\theta'), 3) = 2 > u_2(f^a(\theta'), 3) = 2 - a,$$
$$u_1(f^a(\theta), \theta_1) = 1 \geq u_1(y(\theta'), \theta_1) = 1 \text{ for all } \theta \in \Theta.$$

Suppose $(2, 2) \in \beta(2, 3)$. Let $S = \{1, 2\}$, $\theta^* = (2, 3)$, $\theta' = (2, 2)$, and $y : \Theta \rightarrow A$ be any feasible plan satisfying $y_1(\theta') = (0.1a, 1)$ and $y_2(\theta') = (1 - 0.1a, -1)$. Then

$$u_1(y(\theta'), 2) = 1 + 0.2a > u_1(f^a(\theta'), 2) = 1,$$
$$u_2(y(\theta'), 3) = 2 - 0.3a > u_2(f^a(\theta'), 3) = 2 - a,$$
$$u_2(f^a(\theta), \theta_2) = 1 \geq u_2(y(\theta'), \theta_2) = 1 - 0.2a \text{ for } \theta = (1, 2) \text{ and } (2, 2),$$
$$u_2(f^a(\theta), \theta_2) = 2 \geq u_2(y(\theta'), \theta_2) = 2 - 0.3a \text{ for } \theta = (1, 3) \text{ and } (2, 3).$$

Therefore, we have established the robust coalitional monotonicity condition.
We remark that the social choice set $F$ is not robustly Nash implementable in the sense of Bergemann and Morris (2011), as $F$ fails to satisfy the robust monotonicity condition (for a proof, see Guo and Yannelis, 2016).

8.3 Revisiting the Public Good Example of Bergemann and Morris

We modify the transfer rules of the public good example of Bergemann and Morris (2009). The modified social choice function is still robustly Nash implementable when the interdependence of preferences is small. However, it may not be robustly strong Nash implementable with the same parameter values. We will show that the social choice function is robustly strong Nash implementable if and only if agents have a common value.

Consider an environment with $n$ agents, and each $\Theta_i$ is a fine grid on $[0,1]$. The social planner chooses to provide $x_0 \geq 0$ unit of public good with a cost function $c(x_0) = x_0^2/2$. Agent $i$’s utility function is $u_i(x,\theta) = (\theta_i + \gamma \sum_{j \neq i} \theta_j)x_0 + x_i$. A total-surplus-maximizing social planner has an efficient public good provision level $x_0(\theta) = (1 + \gamma(n-1))\sum_{i=1}^n \theta_i$. Let the transfer rule be $x_i(\theta) = -(1 + \gamma(n-1))(\gamma \theta_i \sum_{j \neq i} \theta_j + \theta_i^2/2 + (\sum_{j \neq i} \theta_j)^2/2)$ for all $i \in I$. The social choice function is $f = (x_0, (x_i)_{i \in I})$.

In this modified version of the public good example of Bergemann and Morris (2009), we add a term that is independent of $i$’s private type $\theta_i$ to $i$’s VCG transfer. As a result, the social choice function is still ex-post incentive compatible. According to Bergemann and Morris (2009), the social choice function is robustly Nash implementable if and only if the interdependence of preferences is small ($|\gamma| < 1/(n-1)$). This is because when $\gamma$ is large, one can exploit the large strategic externality and construct a type space with unwanted equilibria.

Different from their robust Nash implementation results, we show that the social choice function is robustly implementable as an interim strong Nash equilibrium if and only if agents have a common value, i.e. $\gamma = 1$. The interpretation is as follows. If $\gamma = 1$, each agent has the same utility function with the social planner’s surplus function. In this case, an unacceptable deception lowers social welfare and thus every agent’s welfare. If $\gamma \neq 1$, one can find a non-singleton coalition with a profitable deviation from truth-telling.

To prove the result, we apply Theorem 7.3 and Remark 6.1. As transfers are allowed in this example, it is easy to verify the bad outcome property. It is also easy to verify the ex-post weak Pareto efficiency condition when $\gamma = 1$. Therefore, it suffices to focus on robust coalitional incentive compatibility and robust coalitional monotonicity.
Consider the following deviation of coalition $S$ from truth-telling. For all $i \in S$, the reported payoff type is $\theta_i' = \theta_i + k_i\Delta$. Suppose other agents tell the truth, i.e., for all $j \not\in S$, the reported payoff type is $\theta_j' = \theta_j + k_j\Delta$ with $k_j = 0$. Then under the payoff type profile $\theta$ and the claimed payoff type profile $\theta'$, the utility of agent $i \in S$ is given by $(1 + \gamma(n - 1)) \times \frac{(\theta_i + \gamma \sum_{j \not\in i} \theta_j)(\sum_{i \in I} (\theta_i + k_i\Delta) - \gamma(\theta_i + k_i\Delta))\sum_{j \not\in i} (\theta_j + k_j\Delta) - (\theta_i + k_i\Delta)^2 + (\sum_{j \not\in i} (\theta_j + k_j\Delta))^2}{2}$. 

Take the first order condition with respect to $\Delta$, then the optimal $\Delta_i^*$ for agent $i$ is

$$\Delta_i^* = \frac{(\gamma - 1)(\sum_{j \not\in i} \theta_j - \theta_i)\sum_{j \not\in i} k_j}{k_i^2 + (\sum_{j \not\in i} k_j)^2 + 2\gamma k_i \sum_{j \not\in i} k_j}.$$ 

Therefore, if $\gamma = 1$, $\Delta_i^* = 0$ for all $i \in I$, i.e., it is a weakly dominant strategy for all agents to report the true payoff type. If $\gamma < 1$, let $S = \{1, 2\}$, $\theta_1 = 0$, $\theta_2 = 1$ and $\theta_j = 0$ for all $j \not\in S$. Reporting $\theta_1' = \theta_2' = 0.5$ makes both agents in $S$ strictly better off. If $\gamma > 1$, let $S = \{1, 2\}$, $\theta_1 = \theta_2 = 0.5$ and $\theta_j = 0$ for all $j \not\in S$. Reporting $\theta_1' = 0$ and $\theta_2' = 1$ makes both agents in $S$ strictly better off. Hence, $f$ is robust coalitional incentive compatible if and only if $\gamma = 1$.

When $\gamma = 1$, we must prove that $f$ satisfies the robust coalitional monotonicity condition. Assume that there exists a payoff type profile $\theta^*$ and an unacceptable report $\theta'$, i.e., $\sum_{i \in I} \theta_i^* \neq \sum_{i \in I} \theta_i'$. Let $S = I$ and $y$ satisfy $y_0(\theta') = n \sum_{i \in I} \theta_i^*$ and $y_i(\theta') = -n(\sum_{i \in I} \theta_i^*)^2/2$ for all $i \in I$. Then one can verify the robust coalitional monotonicity condition.

### 8.4 A Public Good Example - Implementation under Known Coalition Pattern

Consider a variant of the previous example. Suppose there are two islands in a country, the east one and the west one. Let $I_E = \{1, \ldots, n_E\}$ denote all citizens on the east island and $I_W = \{1 + n_W, \ldots, n_E + n_W\}$ denote all citizens on the south island. Citizens on each island do not communicate with those on the other island, but they communicate frequently with those on the same island. In another word, the coalition pattern is given by $S = \{S \neq \emptyset : S \subseteq I_E$ or $S \subseteq I_W\}$.

The government wants to build a bridge with quality level $x_0$ between the two islands to maximizes social welfare. The production function is $x_0^2/2$. For each agent $i \in I_E$, $i$'s utility from the bridge and a transfer $x_i$ to her is $(\sum_{j \in I_E} \theta_j)x_0 + x_i$. Similarly, for agent $i \in I_W$, $i$'s utility is $(\sum_{j \in I_W} \theta_j)x_0 + x_i$. 

28
A welfare maximizing quality level is $q(\theta) = (n_E \sum_{j \in I_E} \theta_j + n_W \sum_{j \in I_W} \theta_j)$. For each agent $i \in I_E$, $i$'s utility from the bridge and a transfer $x_i$ to her is $(\sum_{j \in I_E} \theta_j)x_0 + x_i$. Similarly, for agent $i \in I_W$, $i$'s utility is $(\sum_{j \in I_W} \theta_j)x_0 + x_i$. Let the transfer of agent $i \in I_E$ be $t_i(\theta) = -n_E(\sum_{j \in I_E} \theta_j)^2/2$ and agent $i \in I_W$ be $t_i(\theta) = -n_W(\sum_{j \in I_W} \theta_j)^2/2$. Denote the social choice function $f$ by $(q, t_1, \ldots, t_n, t_{n+1}, \ldots, t_{n+n})$.

To check robust $S$ incentive compatibility, we let $S$ be a nonempty subset of $I_E$. For all $\theta_{-S}$, the payoff for agent $i \in S$ to report $\hat{\theta}_S$ is

$$(n_E \sum_{j \in S} \hat{\theta}_j + n_E \sum_{j \in I_E \setminus S} \theta_j + n_W \sum_{j \in I_W} \theta_j)(\sum_{j \in S} \theta_j) - n_E(\sum_{j \in S} \hat{\theta}_j + \sum_{j \in I_E \setminus S} \theta_j)^2/2$$

$$\leq (n_E \sum_{j \in S} \theta_j + n_E \sum_{j \in I_E \setminus S} \theta_j + n_W \sum_{j \in I_W} \theta_j)(\sum_{j \in S} \theta_j) - n_E(\sum_{j \in S} \theta_j + \sum_{j \in I_E \setminus S} \theta_j)^2/2.$$  

A similar argument holds for a subset of $I_W$. Therefore, robust coalitional incentive compatibility holds.

To check the robust $S$ monotonicity condition, suppose at payoff type profile $\theta^*$, agents report $\hat{\theta}$, where $\sum_{j \in I_E} \hat{\theta}_j \neq \sum_{j \in I_E} \theta^*_j$. Let $y = (q^y, t^y_1, \ldots, t^y_{n+n})$ be any feasible rule where for all $\theta \in \Theta$ and $i \in I_E$, $q^y(\theta) = (n_E \sum_{j \in I_E} \theta^*_j + n_W \sum_{j \in I_W} \theta_j)$ and $t_i(\theta) = -n_E(\sum_{j \in I_E} \theta^*_j)^2/2$. Then for all $\theta \in \Theta$ and $i \in I_E$,

$$(n_E \sum_{j \in I_E} \theta_j + n_W \sum_{j \in I_W} \theta_j)(\sum_{j \in I_E} \theta_j) - n_E(\sum_{j \in I_E} \theta_j)^2/2$$

$$\geq (n_E \sum_{j \in I_E} \theta^*_j + n_W \sum_{j \in I_W} \theta_j)(\sum_{j \in I_E} \theta^*_j) - n_E(\sum_{j \in I_E} \theta^*_j)^2/2.$$  

In addition, as $\sum_{j \in I_E} \hat{\theta}_j \neq \sum_{j \in I_E} \theta^*_j$, for all reported $\hat{\theta}_{I_W} \in \Theta_{I_W}$,

$$(n_E \sum_{j \in I_E} \theta^*_j + n_W \sum_{j \in I_W} \hat{\theta}_j)(\sum_{j \in I_E} \theta^*_j) - n_E(\sum_{j \in I_E} \theta^*_j)^2/2$$

$$>(n_E \sum_{j \in I_E} \hat{\theta}_j + n_W \sum_{j \in I_W} \hat{\theta}_j)(\sum_{j \in I_E} \theta^*_j) - n_E(\sum_{j \in I_E} \hat{\theta}_j)^2/2.$$  

For $\sum_{j \in I_W} \hat{\theta}_j \neq \sum_{j \in I_W} \theta^*_j$, a similar argument holds. Hence, we have established the robust $S$ monotonicity condition.

It is easy to verify the local ex-post weak Pareto efficiency condition and the bad outcome property. Therefore, Corollary 4.1 implies that the given social choice function is robustly implementable under the coalition pattern $S$.  

29
8.5 A Public Good Example - Double Implementation

Consider another variant of the example in Section 8.3. There are $n$ agents in the environment and each one has a private value in $\Theta_i = \{0, 1\}$. On this smaller payoff type space, we will show that robust coalitional incentive compatibility holds. The mechanism designer is not sure which coalitions can be formed. The cost of producing $x_0 \geq 0$ unit of public good is $c(x_0) = x_0^2/2$. Agent $i$’s utility function is $u_i(x, \theta) = \theta_i \cdot x_0 + x_i$ with $x_0$ unit of public good and $x_i$ unit of transfer. The efficient public good provision level is $x_0(\theta) = \sum_{i=1}^{n} \theta_i$. Let the transfer rule be $x_i(\theta) = -\theta_i^2/2$ for all $i \in I$.

By Bergemann and Morris (2011), the social choice function defined by $f = (x_0, (x_i)_{i \in I})$ satisfies the robust monotonicity condition. It is easy to verify the bad outcome property. Therefore, it remains to show the robust coalitional incentive compatibility.

When the true payoff type profile is $\theta$, suppose that each agent $i$ in a coalition $S$ misreports $\theta_i + \epsilon_i$. For $i \in S$, the net gain from mireporting is

$$((\sum_{j \in S} \epsilon_j + \sum_{j \in I} \theta_j) \cdot \theta_i - (\theta_i + \epsilon_i)^2/2) - ((\sum_{j \in I} \theta_j) \cdot \theta_i - \theta_i^2/2)$$

$$= (\sum_{j \in S, j \neq i} \epsilon_j) \cdot \theta_i - \epsilon_i^2/2.$$

If there exists $i \in S$ such that $\theta_i = 0$, then the coalition $S$ cannot be strictly better-off by deviating. When $\theta_j = 1$ for all $j \in S$, as $\Theta_j = \{0, 1\}$, it must be the case that $\sum_{j \in S, j \neq i} \epsilon_j \leq 0$, which means that $S$ does not benefit from deviating. Therefore, the robust coalitional incentive compatibility condition also holds. By Corollary 7.2, the social choice function $f$ is robustly double implementable.

9 Concluding Remarks

This paper introduces coalitional structure into the belief-free approach of Bergemann and Morris (2005, 2011, etc.). Specifically, we prove that the conditions of robust coalitional incentive compatibility and robust coalitional monotonicity are necessary and almost sufficient for robustly fully implementing a social choice set as an interim strong Nash equilibrium. Our modeling provides insights on implementing social choice sets that may not be robustly Nash implementable, as is shown in Section 8.

The methodology of the paper can be extended to accommodate the maximin expected
utility (MEU) of Gilboa and Schmeidler (1989). The introduction of MEU solves the conflict between efficiency and ambiguity, and therefore it makes incentive compatibility and partial implementation easier to obtain (e.g., De Castro and Yannelis, 2008; De Castro et al., 2015). It is natural to study the effect of MEU on full implementation. We provide an outline below.

First, we must generalize the type spaces to accommodate MEU. An **ambiguous type space** is a collection \( T = (T_i, \hat{\theta}_i, \Pi_i)_{i=1}^I \), where agent \( i \)'s **belief type** \( \Pi_i(t_i) \) is defined by the function \( \Pi_i : T_i \rightarrow 2^{\Delta(T_{-i})} \) that maps each type of agent \( i \) into a nonempty, compact, convex set of probability distributions over others' types; each element \( \pi_i(t_i) \in \Pi_i(t_i) \) is a probability distribution over \( T_{-i} \). Such an ambiguous type space and common knowledge of the ambiguous type space can be formally defined. For each allocation plan \( x : T \rightarrow A \) and each type \( t_i \in T_i \), agent \( i \)'s maximin expected utility is defined as

\[
V_i(x, t_i) = \min_{\pi_i(t_i) \in \Pi_i(t_i)} \sum_{t_{-i} \in T_{-i}} u_i(x(t), \hat{\theta}(t)) \pi_i(t_i)[t_{-i}].
\]

In an ambiguous type space, if one considers interim implementation in pure strategies, the interim conditions of this paper can be adopted. If one considers mixed strategies, a bigger change in the interim conditions is needed because mixed strategies can hedge against ambiguity. For example, incentive compatibility would require that there does not exist a profitable lottery of misreports.

However, if one wants to obtain robust strong Nash implementation in all ambiguous type spaces, the conditions of robust coalitional incentive compatibility and robust coalitional monotonicity are still necessary and almost sufficient. Hence, with the above modifications, the main results of the paper still hold for robust strong Nash implementation under MEU framework.

### A Appendix

#### A.1 Ommitted Proofs

**Proposition A.1:** Given a payoff environment \( \Theta \), a social choice set \( F \) is robust coalitional incentive compatible if and only if it is coalitional incentive compatible in all type spaces with payoff environment \( \Theta \).

\(^9\)For example, see Epstein and Wang (1996).
Proof. We prove the "only if" part of the proposition first. Let $F$ be a robust coalitional incentive compatible social choice set. Suppose by way of contradiction that there exists a type space $T$, $f \in F$, $S \subseteq I$, and $t_S^*, t'_S \in T_S$ such that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i\left(f(\hat{\theta}(t_S^*, t_{-S})), \hat{\theta}(t_S^*, t_{-S})\right) \pi_i(t_S^*)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i\left(f(\hat{\theta}(t'_S, t_{-S})), \hat{\theta}(t'_S, t_{-S})\right) \pi_i(t'_S)[t_{-i}].$$

For all $i \in S$, let $\theta^*_i = \hat{\theta}(t_i^*)$ and $\theta'_i = \hat{\theta}(t'_i)$. The above inequality shows that for all $i \in S$, there exists $\theta^*_{i-S}$ such that $u_i(f(\theta^*_S, \theta^*_{i-S}), (\theta^*_S, \theta^*_{i-S})) > u_i(f(\theta'_S, \theta'_i), (\theta'_S, \theta'_i))$, contradicting the robust coalitional incentive compatibility condition.

To prove the "if" part, suppose that $F$ does not satisfy the robust coalitional incentive compatibility condition, i.e., there exists $f \in F$, $S \subseteq I$, and $\theta^*_S, \theta'_S \in \Theta_S$ such that for all $i \in S$, there exists $\theta^*_{i-S} \in \Theta_{-S}$ such that $u_i(f(\theta^*_S, \theta^*_{i-S}), (\theta^*_S, \theta^*_{i-S})) > u_i(f(\theta'_S, \theta'_i), (\theta'_S, \theta'_i))$. Then we let $T$ be a payoff type space, i.e., a type space where each $\hat{\theta}_i$ is a bijection between $T_i$ and $\Theta_i$, satisfying the following condition. For all $i \in S$ and $t_i \in T_{i}$, $\pi_i(t_i)[t_{-i}] = 1$, if $t_{-i}$ has payoff types $\theta^*_{i-S}$, and $\pi_i(t_i)[t_{-i}] = 0$ elsewhere. For each $i \in S$, let $t'_i$ be the type with a payoff type $\theta'_i$. The above inequality implies that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i\left(f(\hat{\theta}(t'_S, t_{-S})), \hat{\theta}(t'_S, t_{-S})\right) \pi_i(t'_S)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i\left(f(\hat{\theta}(t'_S, t_{-S})), \hat{\theta}(t'_S, t_{-S})\right) \pi_i(t'_S)[t_{-i}].$$

Therefore, $F$ is not coalitional incentive compatible in $T$, a contradiction. 

Proposition A.2: Given a payoff environment $\Theta$, a social choice set $F$ is robust coalitional monotonic if and only if it is coalitional monotonic in all type spaces with payoff environment $\Theta$.

Proof. We begin with proving the "if" part of the equivalence relation. Let $F$ satisfy the coalitional monotonicity condition in all type spaces, but suppose by way of contradiction that the robust coalitional monotonicity condition fails. Then there exists $f \in F$ and an unacceptable deception $\beta$, such that for all $S \subseteq I$, $\theta_S \in \Theta_S$, and $\theta'_S \in \beta_S(\theta_S)$, there is a reasonable conjecture $(\psi_i)_{i \in S}$ and a reasonable updating rule, such that whenever $y \in Y_{\theta'_S}^f$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i\left(y(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S})\right) \psi_i(\theta'_{-i}; \theta_S)[\theta_{-i}] \leq \sum_{\theta_{-i} \in \Theta_{-i}} u_i\left(y(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S})\right) \psi_i(\theta'_{-i}; \theta_S)[\theta_{-i}]$$

for some $i \in S$ and $\theta'_{-S} \in \beta_{-S}(\Theta_{-S})$.

Therefore, for all $\theta_i \in \Theta_i$ and $\theta'_i \in \beta_i(\theta_i)$, there exists $\theta'_{-i} \in \Theta_{-i}$ and $\theta_{-i} \in \beta_{-i}(\theta'_{-i})$ such that for all $\theta_S \in \Theta_S$, $\theta'_S \in \beta_S(\theta_S)$, and $y \in Y_{\theta'_S}^f$, there exists $i \in S$ such that
\begin{align}
\sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta'_S, \theta'^i_S), (\theta_S, \theta_{-S})) & = u_i(f(\theta'_S, \theta'^i_S), (\theta_S, \theta_{-S})).
\end{align}

Therefore, for all \( \theta_i \in \Theta_i \) and \( \theta'_i \in \beta_i(\theta_i) \), there exists a reasonable conjecture \( \psi_i \), a reasonable updating rule, and \( \theta'^i_{-i} \in \Theta_{-i} \) such that for all \( \theta_S \in \Theta_S, \theta'_S \in \beta_S(\theta_S) \), and \( y \in Y^f_{\theta'_S} \), there exists \( i \in S \) such that

\begin{align}
\sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta'_S, \theta'^i_S), (\theta_S, \theta_{-S})) \psi_i(\theta'_{S\setminus\{i\}}, \theta'^i_{-S}; \theta_S)[\theta_{-i}] & \leq \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta'_S, \theta'^i_S), (\theta_S, \theta_{-S})) \psi_i(\theta'_{S\setminus\{i\}}, \theta'^i_{-S}; \theta_S)[\theta_{-i}].
\end{align}

(A.1)

The proof proceeds as follows. First, we construct a type space \( T \) with each \( \Xi_i = \{(\theta_i, \theta'_i) : \theta_i \in \Theta_i \) and \( \theta'_i \in \beta_i(\theta_i)\} \), as well as the \( \theta'^i_{-i} \) and \( \psi_i \) above. Second, we define an unacceptable deception \( \alpha : T \rightarrow T \). Third, we specify an updating rule for \( T \) when Bayes’ rule cannot be applied, as we focus on interim strong Nash implementation. Fourth, a contradiction is derived.

Let \( \hat{\theta} \in \Theta \) be an arbitrary payoff type profile, and fix it throughout our argument. The following type space is a modification of Bergemann and Morris (2008).

For each \( i \in I \), let \( T^1_i \) be a type set isomorphic to \( \Xi_i \) with a bijection \( \xi^1_i : T^1_i \rightarrow \Xi_i \). For each \( t_i \in T^1_i \) with \( \xi^1_i(t_i) = (\theta_i, \theta'_i) \), let \( \hat{\theta}_i(t_i) = \theta_i \). Define \( \pi_i(t_i)[t_{-i}] = \psi_i(\theta'^i_{-i})[\theta_{-i}] \) if \( t_{-i} = (\{\xi^1_j\}^{-1}(\theta_j, \theta'^j_j))_j \neq i \in T^1_{-i}; \pi_i(t_i)[t_{-i}] = 0 \) elsewhere.

Define another type set \( T^2_i \) isomorphic to \( \Theta \) with a bijection \( \xi^2_i : T^2_i \rightarrow \Theta \). For each \( t_i \in T^2_i \) with \( \xi^2_i(t_i) = (\theta_i, \theta_{-i}) \), let \( \hat{\theta}_i(t_i) = \theta_i \). Define \( \pi_i(t_i)[t_{-i}] = 1 \) if for all \( j \neq i, t_j = [\xi^2_j]^{-1}(\theta_j, \theta_{-j}) \in T^2_j; \pi_i(t_i)[t_{-i}] = 0 \) elsewhere.

Let \( T_i = T^1_i \cup T^2_i \) be the set of all types for agent \( i \), and we obtain a type space \( T \).

For \( i \in I \) and \( S \) satisfying \( i \in S \subset I \), when \( S \) has types \( t_S \) and \( \pi_i(t_i)[t_{S\setminus\{i\}}] \neq 0 \), apply Bayes’ rule to update the belief. When \( \pi_i(t_i)[t_{S\setminus\{i\}}] = 0 \), Bayes’ rule fails and hence we specify an updating rule as follows. If there exists \( j \in S \) such that \( t_j \in T^1_j \), let \( \pi_i(t_S)[t_{S\setminus\{i\}}, t_{-S}] = \psi_i(\hat{\theta}_{S\setminus\{i\}}(\alpha_{S\setminus\{i\}}(t_{S\setminus\{i\}})) \theta'^S_{-S}; \hat{\theta}_S(t_S)) [\theta'_{S\setminus\{i\}}(t_{S\setminus\{i\}}), \theta_{-S}] \) for \( t_{-S} = ([\xi^1_j]^{-1}(\theta_j, \theta'^j_j))_{j \in S} \), where \( \psi_i \) and \( \theta'^i_{-i} \) are associated with the pair \( \hat{\theta}_i(t_i) \) and \( \alpha_i(\hat{\theta}_i(t_i)) \); elsewhere, let \( \pi_i(t_S)[\cdot] = 0 \). If \( t_S \in T^2_S \), let \( \pi_i(t_S)[\cdot] \) satisfy that the marginal probabilities \( \pi_i(t_S)[t_{S\setminus\{i\}}] = 1 \) and \( \pi_i(t_S)[t_{-S}] = \pi_i(t_i)[t_{-S}] \) for each \( t_{-S} \in T_{-S} \).

Let an unacceptable deception \( \alpha : T \rightarrow T \) be:

\[
\alpha_i(t_i) = \begin{cases} 
\text{[\xi^2_j]^{-1}(\theta'_j, \theta_{-j})} & \text{if } t_i = [\xi^1_i]^{-1}(\theta_i, \theta'_i) \in T^1_i, \\
 t_i & \text{elsewhere}.
\end{cases}
\]

As \( \alpha \) is unacceptable in \( T \), there exists \( f \in F, t^*_S \in T_S, \) and \( h \in H^f_{\alpha_S(t^*_S)} \) such that
the strict inequality in Definition 4.3 holds for all $i \in S$. For all $\hat{\theta} \in \Theta$, define $y(\hat{\theta}) = h\left(\alpha_S(t^*_S), \left(\xi^{-1}_S(\hat{\theta}_j; \hat{\theta}_-j)\right)_{j \notin S}\right)$. In the definition of $h \in H^f_{\alpha_S(t^*_S)}$, by letting $t''_S$ go over all type profiles in $T^2_S$, one can verify that $y \in Y^f_{\tilde{\theta}_S(\alpha_S(t^*_S))}$. Also, we know $t^*_S \notin T^2_S$, as otherwise letting $t''_S = t^*_S$ would result in a contradiction between the strict inequalities in Definition 4.3 and that $h \in H^f_{\alpha_S(t^*_S)}$. As there exists $i \in S$ such that $t^*_i \in T^1_i$, the strict inequalities in Definition 4.3 as well as the described updating rule imply that for all $i \in S$,

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S})) \psi_i(\theta'_{S\setminus\{i\}}, \theta'_{-S}; \theta_S)[\theta_{-i}] > \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta'_S, \theta'_{-S}), (\theta_S, \theta_{-S})) \psi_i(\theta'_{S\setminus\{i\}}, \theta'_{-S}; \theta_S)[\theta_{-i}],$$

a contradiction to expression (A.1).

Now we prove the "only if" half of the equivalence relation. Let $T$ be an arbitrary type space with payoff environment $\Theta$. Let $\alpha : T \to T$ be an unacceptable deception for $f \in F$. Define a correspondence $\beta : \Theta \to 2^{\Theta \setminus \emptyset}$ by $\beta_i(\theta_i) = \{\theta'_i : \text{there exists } t_i \in T_i \text{ with } \hat{\theta}_i(t_i) = \theta_i \text{ and } \hat{\theta}_i(\alpha_i(t_i)) = \theta'_i\}$ for all $\theta_i \in \Theta_i$ and $i \in I$, which is unacceptable for $f$. Suppose the social choice set $F$ satisfies the robust coalitional monotonicity condition, then there exists $S \subseteq I$, $\theta^*_S \in \Theta_S$, and $\theta'_{-S} \in \beta_S(\theta^*_S)$, such that for any reasonable conjecture $(\psi_i)_{i \in S}$ and reasonable updating rule, there exists $y \in Y^f_{\tilde{\theta}_S}$ such that for all $i \in S$ and $\theta'_{-S} \in \beta_S(\Theta_{-S})$:

$$\sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta'_S, \theta'_{-S}), (\theta^*_S, \theta_{-S})) \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{-i}] > \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta'_S, \theta'_{-S}), (\theta^*_S, \theta_{-S})) \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{-i}].$$

Let $t^*_S$ be a type profile satisfying $\hat{\theta}_S(t^*_S) = \theta^*_S$ and $\hat{\theta}_S(\alpha_S(t^*_S)) = \theta'_{-S}$. For all $i \in S$ and $\theta'_{-S} \in \Theta_{-S}$, let $\psi_i(\theta'_{-i})[\cdot]$ be an arbitrary distribution with a reasonable updating rule such that

$$\psi_i(\theta'_{-i}; \theta^*_S)[\theta_{S\setminus\{i\}}, \theta_{-S}] = \sum_{\{t_{-S}; \theta_{-S}(t_{-S}) = \theta_{-S}, \theta_S = \theta^*_S\}} \pi_i(t^*_S)|t^*_S, t_{-S}| \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{S\setminus\{i\}}, \theta_{-S}]$$

for all $\theta_{-S} \in \Theta_{-S}$, if the denominator is nonzero. For all $t \in T$, let $h(t) = y(\theta'_S, \hat{\theta}_S(t_{-S}))$. Then,

$$\sum_{t_{-S} \in T_{-S}} u_i\left(h(\alpha(t^*_S, t_{-S}), \hat{\theta}(t^*_S, t_{-S})) \pi_i(t^*_S)[t_{-i}]\right)$$

$$= \sum_{\theta'_{-S} \in \beta_{-S}(\Theta_{-S})} \sum_{\rho_{-S}(\theta'_{-S}) \rho'_{-S}(\theta'_{-S})} u_i(y(\theta'_S, \theta'_{-S}), (\theta^*_S, \theta_{-S})) \left(\sum_{\{t_{-S}; \hat{\theta}_{-S}(t_{-S}) = \theta'_{-S}\}} \pi_i(t^*_S)|t^*_S, t_{-S}| \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{S\setminus\{i\}}, \theta_{-S}]\right) \left(\sum_{\{t_{-S}; \theta_{-S}(t_{-S}) = \theta_{-S}\}} \pi_i(t^*_S)|t^*_S, t_{-S}| \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{S\setminus\{i\}}, \theta_{-S}]\right) \sum_{\{t_{-S}; \theta_{-S}(t_{-S}) = \theta_{-S}\}} \pi_i(t^*_S)|t^*_S, t_{-S}| \psi_i(\theta'_{-i}; \theta^*_S)[\theta_{S\setminus\{i\}}, \theta_{-S}]$$
A.2 An Alternative Definition of Interim Strong Nash Equilibrium

Definition 3.1 assumes that agents within a coalition pool their private information, which helps the coalition obtain higher efficiency. For the methodology of the paper to work, the information pooling assumption is not required. For instance, if there is no information sharing within the coalition, we have an alternative definition of an interim strong Nash equilibrium. Under this new definition, the fine core and the rational expectations equilibrium are still robustly strong Nash implementable. However, the efficiency level is reduced compared to Definition 3.1 in general.

Definition A.1: In a type space $\mathcal{T}$, the strategy profile $\sigma^*$ is an interim strong Nash equilibrium of the mechanism $(M, g)$ if there does not exist $S \subseteq I$, $t^*_S \in \mathcal{T}_S$, and $m_S \in M_S$, such that for all $i \in S$,

$$\sum_{t_{-i} \in T_{-i}} u_i\left(g\left(m_S, \sigma^*_{-S}(t_{-S})\right), \hat{g}(t^*_i, t_{-i})\right) \pi_i(t^*_i)[t_{-i}] > \sum_{t_{-i} \in T_{-i}} u_i\left(g\left(\sigma^*(t^*_i, t_{-i})\right), \hat{g}(t^*_i, t_{-i})\right) \pi_i(t^*_i)[t_{-i}].$$

Namely, under any type profile, there does not exist a coalition who can strictly benefit from committing within the coalition to sending a certain message. By setting $S$ to be singletons, we see that the above definition is a subset of the interim Nash equilibrium.

Based on Definition A.1, we modify the conditions of robust coalitional incentive compatibility and robust coalitional monotonicity as follows.

Definition A.2: A social choice set $F$ is robust coalitional incentive compatible if for all $f \in F$, $S \subseteq I$, and $\theta'_S \neq \theta^*_S$, there exists $i \in S$ such that

$$u_i\left(f(\theta'_i, \theta_{-i}), (\theta^*_i, \theta_{-i})\right) \geq u_i\left(f(\theta'_S, \theta_{-S}), (\theta^*_i, \theta_{-i})\right) \text{ for all } \theta_{-i} \in \Theta_{-i}. $$
Definition A.3: A social choice set $F$ satisfies robust coalitional monotonicity if for each $f \in F$ and unacceptable deception $\beta$, there exists $S \subseteq I$, $\theta^*_S \in \Theta_S$, and $\theta^i_S \in \beta_S(\theta^*_S)$, such that for any reasonable conjecture $(\psi_i)_{i \in S}$, there exists $y : \Theta \rightarrow A$ such that

1. for all $i \in S$ and $\theta^i_{-i} \in \beta_{-i}(\Theta_{-i})$
   \[
   \sum_{\theta_{-i} \in \Theta_{-i}} u_i(y(\theta^i_S, \theta^i_{-i}), (\theta^*_i, \theta_{-i})) \psi_i(\theta^i_{-i})[\theta_{-i}] > \sum_{\theta_{-i} \in \Theta_{-i}} u_i(f(\theta^i, \theta^i_{-i}), (\theta^*_i, \theta_{-i})) \psi_i(\theta^i_{-i})[\theta_{-i}],
   \]

2. for all $\theta''_S \in \Theta_S$, there exists $i \in S$ such that
   \[
   u_i(f(\theta''_i, \theta_{-i}), (\theta''_i, \theta_{-i})) \geq u_i(y(\theta^i_S, \theta_{-S}), (\theta''_i, \theta_{-i})), \forall \theta_{-i} \in \Theta_{-i}.
   \]

Following the arguments in this paper, one can prove that the two conditions are necessary and almost sufficient for robust strong Nash implementation under the alternative definition of interim strong Nash equilibrium. Note that in the sufficiency theorems, the mechanisms, as well as the ex-post weak Pareto efficiency condition, need some modification.

The methodology in this paper allows for more intricate communication within the coalition, as long as the coalition commits to submitting some message $m_S$. For example, consider the case where there is no direct communication within the coalition about private information, but the fact that all members benefit from submitting $m_S$ reveals additional information of the coalition’s true types. In this case, we can denote the updated belief by $\pi_i(t^*_i, m_S)[\cdot]$ and modify the above three definitions accordingly. It should be noted that the commitment within the coalition to submitting $m_S$ plays a role in the proof. As the messages submitted to the mechanism designer are verifiable, the commitment can be enforced within the coalition.

References


