Working Paper 94-53
Economics Series 25
November 1994

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9875

# COALITION-PROOF EQUILIBRIUM

Diego Moreno and John Wooders*

Abstract

We characterize the set of agreements that the players of a non-cooperative game may reach when they have the opportunity to communicate prior to play. We show that communication allows the players to correlate their actions. Therefore, we take the set of correlated strategies as the space of agreements. Since we consider situations where agreements are non-binding, they must not be subject to profitable self-enforcing deviations by coalitions of players. A coalition-proof equilibrium is a correlated strategy from which no coalition has an improving and self-enforcing deviation. A coalition-proof equilibrium exists when there is a correlated strategy which (i) has a support contained in the set of actions that survive the iterated elimination of strictly dominated strategies, and (ii) weakly Pareto dominates every other correlated strategy whose support is contained in that set. Consequently, the unique equilibrium of a dominance solvable game is coalition-proof.

*Moreno, Department of Economics, University of Arizona, and Departamento de Economía, Universidad Carlos III de Madrid. This author gratefully acknowledges financial support from the Ministerio Asuntos Sociales administered through the Cátedra Gumersindo Azcárate, and from DGICYT grant PB93-0230.

Wooders, Department of Economics, University of Arizona, and Departamento de Economía, Universidad Carlos III de Madrid. This author gratefully acknowledges support from the Spanish Ministry of Education.

# Introduction

When the players of a noncooperative game have the opportunity to communicate prior to play, they will try to reach an agreement to coordinate their actions in a mutually beneficial way. The aim of this paper is to characterize the set of agreements that the players may reach. Since we consider situations where agreements are non-binding, only those agreements that are not subject to viable (i.e., *self-enforcing*) deviations are of interest. As pre-play communication allows the players to correlate their play, we take the set of all correlated strategies as the space of feasible agreements. We characterize the set of coalition-proof equilibria as the set of agreements from which no coalition has a self-enforcing deviation making all its members better off.

Admitting correlated strategies as feasible agreements alters the set of coalition-proof equilibria of a game in a fundamental way (*viz.*, no inclusion relationship between the notion of coalition-proofness that we propose and others previously introduced is to be found). In fact, there are games where the only plausible agreements are correlated (and not mixed) agreements. We provide examples with this feature and we show that the notion of coalition-proof equilibrium that we propose identifies these agreements. Unfortunately, as with other notions of coalition proofness previously introduced, existence of an equilibrium cannot be guaranteed. We are able to show, however, that if there is a correlated strategy which (i) has a support contained in the set of actions that survive the iterated elimination of strictly dominated strategies, and (ii) weakly Pareto dominates every other correlated strategy whose support is contained in that set, then this strategy is a coalition-proof equilibrium. Consequently, the unique equilibrium of a dominance solvable game is coalition-proof.

Other authors have explored the implications of pre-play communication when agreements are mixed strategy profiles. Aumann [1] introduced the notion of strong Nash equilibrium, which requires that an agreement not be subject to an improving deviation by any coalition of players. This requirement is too strong, since agreements must be resistant to deviations which are not themselves resistant to further deviations. Recognizing this problem, Bernheim, Peleg and Whinston [3] (henceforth referred to as BPW) introduced the notion of coalition-proof Nash equilibrium ($CPNE$), which requires only that an agreement be immune to improving deviations which are self-enforcing. A deviation is self-enforcing if there is no further self-enforcing and improving deviation available to a proper subcoalition of players. This notion of "self-enforcingness" provides a useful means of distinguishing coalitional deviations that are viable from those that are not re-

sistant to further deviations. Only viable deviations can upset potential agreements. A deficiency of $CPNE$, however, is that it does not allow players to agree to correlate their play.

Although the possibility that players correlate their actions when given the opportunity to communicate was recognized as early as in Luce and Raiffa [7], only recently did Einy and Peleg [4] (E&P) introduce a concept of coalition-proof communication equilibrium. The difference between E&P's notion and ours can be better understood if we assume that correlated agreements are carried out with the assistance of a mediator. The mediator selects an action profile according to the agreement and then makes a (private and non-binding) recommendation of an action to each player.

E&P consider situations where the players may plan deviations only after receiving recommendations. In our framework, however, players plan deviations before receiving recommendations, and no further communication is possible after recommendations are issued. This difference manifests itself most clearly in two-person games where an agreement is coalition-proof in our sense only if it is Pareto efficient within the set of correlated equilibria, while an agreement that is coalition-proof in E&P's sense need not be. We provide an example with this feature in Section 3. The second difference is that in our framework deviations may involve the members of a coalition jointly "misreporting" their types, while this possibility is not considered by E&P's notion. In section 3 these differences are discussed in detail. Subsequent to our work, Ray [10] has characterized coalition-proof agreements when the players' possibilities of correlating their actions are exogenously given.

As the following example illustrates, correlated play naturally arises when communication is possible. Therefore one should take the set of correlated strategies as the set of feasible agreements, and one must consider deviations that involve correlated play by members of a deviating coalition.

> *Three Player Matching Pennies Game.* Three players each simultaneously choose heads or tails. If all three faces match, then players 1 and 2 each win a penny while player 3 loses two pennies. Otherwise, player 3 wins two pennies while players 1 and 2 each lose a penny.

This game has two pure strategy and one mixed strategy Nash equilibria: one pure strategy equilibrium consists of players 1 and 2 each choosing heads (tails) and player 3 choosing tails (heads). In the mixed strategy equilibrium each player chooses heads with probability $\frac{1}{2}$.

The game does not have a $CPNE$, as each of the Nash equilibria is upset by a deviation of the coalition of players 1 and 2: in the pure strategy Nash equilibrium where players 1 and 2

both choose heads, they each obtain a payoff of $-1$. By jointly deviating (both choosing tails instead) players 1 and 2 each obtain a payoff of 1. This deviation is self-enforcing as players 1 and 2 each obtain their highest possible payoffs and therefore neither player can improve by a further unilateral deviation. (A symmetric argument shows that the other pure strategy Nash equilibrium is not a $CPNE$ either.) In the mixed strategy Nash equilibrium, players 1 and 2 each obtain an expected payoff of $-\frac{1}{2}$. This equilibrium is not a $CPNE$ as players 1 and 2 can jointly deviate (both choosing heads instead) and obtain a payoff of zero. This deviation is self-enforcing, since given that player 3 chooses heads or tails with equal probability, neither player can obtain more than zero by a further deviation. Since a $CPNE$ must be a Nash equilibrium, this game has no $CPNE$.

Nevertheless, the game does have an agreement that is resistant to improving deviations. This agreement is the correlated strategy where with probability $\frac{1}{2}$ players 1 and 2 both choose heads and with probability $\frac{1}{2}$ both choose tails, and player 3 chooses heads or tails with equal probability. Under this agreement each player has an expected payoff of zero. No single player can deviate and improve upon this agreement: neither player 1 nor player 2 can benefit by unilaterally deviating, as they both loose a penny whenever their faces do not match. Neither does player 3 benefit from deviating: given the probability distribution over the moves of players 1 and 2, he is indifferent between heads and tails. Moreover, since the interests of players 1 and 2 are completely opposed to those of player 3, no coalition involving player 3 can improve upon the given agreement. Finally, given player 3's strategy, players 1 and 2 obtain at most a payoff of zero, and therefore they cannot benefit by deviating. Hence, no coalition can gain by deviating from the agreement.

Notice that the agreement described above is *not* a mixed strategy and so cannot possibly be a $CPNE$. As we shall see, however, when we expand the space of agreements to include all the correlated strategies, this agreement is the unique coalition-proof equilibrium of the game.

The possibility of players correlating their play arises even when communication is limited. Consider, for instance, the following example which is related to a class of games discussed in Farrell [5]: two identical firms must simultaneously decide whether to enter a market which is a natural monopoly. Firm payoffs are given in the following table:

| | Enter | Not Enter |
|---|---|---|
| Enter | -2,-2 | 1,-1 |
| Not Enter | -1,1 | 0,0 |

This game has three Nash equilibria: (Enter, Not Enter), (Not Enter, Enter), and the mixed strategy Nash equilibrium where each firm enters the market with probability $\frac{1}{2}$. Each of these Nash equilibria is also a $CPNE$.

Although the mixed Nash equilibrium is a $CPNE$, it is not resistant to improving deviations given the possibility of pre-play communication. The firms can improve by augmenting the game with a round of cheap talk. In the game with cheap talk each firm simultaneously and publicly announces whether it intends to "Enter" or "Not Enter" the market. Following both announcements each firm makes its choice.

Suppose the firms agree to play the following Nash (and subgame perfect) equilibrium of the game with cheap talk. Each firm announces "Enter" with probability $\frac{3}{4}$. If the profile of announcements is either (Enter, Not Enter) or (Not Enter, Enter), then each firm plays its announcement. Otherwise, each firm plays "Enter" with probability $\frac{1}{2}$. This equilibrium yields an expected payoff for each firm of $-\frac{5}{16}$ while in the mixed Nash equilibrium of the original game each firm has an expected payoff of only $-\frac{1}{2}$.

Pre-play communication has enabled the firms to correlate their play. In this Nash equilibrium of the cheap talk game the firms effectively play the correlated strategy of the original game given by

| | Enter | Not Enter |
|---|---|---|
| Enter | $\frac{5}{32}$ | $\frac{11}{32}$ |
| Not Enter | $\frac{11}{32}$ | $\frac{5}{32}$ |

This joint probability distribution is not the product of its marginal distributions and therefore cannot be obtained from a mixed strategy profile of the game without communication. This "correlated deviation" from the mixed strategy equilibrium makes both firms better off. Moreover, it is a self-enforcing deviation since it is a correlated equilibrium of the original game.

Expanding the set of feasible agreements from the mixed strategies (as in $CPNE$) to the set of correlated strategies does not lead simply to an expansion of the set of coalition-proof agreements. In the *Three Player Matching Pennies* game we found a coalition-proof agreement where no $CPNE$ existed. In the entry game we found a $CPNE$ that was not coalition-proof. Thus, there is no inclusion between the set of $CPNE$ and the set of equilibria that are coalition-proof in our sense.

In our framework the primitives are a set of feasible agreements and the concepts of feasible deviation and of self-enforcing deviation by a coalition from a given agreement. The set of feasible deviations by a coalition from a given agreement is the set of all correlated strategies that the

coalition can induce when the complementary coalition behaves according to the given agreement and when the members of the coalition correlate their play. The definition of a self-enforcing deviation is recursive. For a coalition of a single player any feasible deviation is self-enforcing. For coalitions of more than one player, a deviation is self-enforcing if it is feasible and if there is no further self-enforcing and improving deviation by one of its proper subcoalitions. With these concepts, our notion of coalition-proofness is easily formulated: an agreement is coalition-proof if no coalition (not even the grand coalition) has a self-enforcing deviation that makes all its members better off.

Our notion of a self-enforcing deviation coincides with that implicit in the concept of $CPNE$. The difference between our notion of coalition-proofness and $CPNE$ is only that we take the set of correlated strategies as the space of feasible agreements. For games of complete information, if feasible agreements are mixed strategies then our definition of coalition-proofness coincides with $CPNE$. (This is established in Appendix B.) In some situations it may be natural to restrict the space of feasible agreements (e.g., if communication is limited) or to limit the possibilities of players to form deviations. The framework we propose easily accommodates these kinds of changes.

The paper is organized as follows: in Section 1 we discuss our framework and define our notion of equilibrium for games of complete information. In Section 2 we extend the concept of coalition-proofness to games of incomplete information. Of course, the notion of coalition-proof equilibrium for games of incomplete information reduces to that formulated for games of complete information when every player has only a single type. We present separately the notion of coalition-proofness for games of complete information, as the notion's simplicity in this context facilitates the discussion and because we want to stress the fact that our notion of coalition-proofness can be formulated without resorting to games of incomplete information. In Section 3 we compare our notion of coalition-proof equilibrium and E&P's notion of coalition-proof communication equilibrium, and we present some concluding remarks.

## 1. Games of Complete Information

A game in strategic form $\Gamma$ is defined as

$$\Gamma = (N, (A_i)_{i \in N}, (u_i)_{i \in N}),$$

where $N$ is the set of players, and for each $i \in N$, $A_i$ is player $i$'s set of actions (or pure strategies) and $u_i$ is player $i$'s utility (payoff) function, a real valued function on $A = \prod_{i \in N} A_i$. Assume that $N$ and $A$ are nonempty and finite. For any finite set $Z$, denote by $\triangle Z$ the set of probability

distributions over $Z$. In particular, denote by $\triangle A$ the set of probability distributions over $A$, and refer to its members as *correlated strategies.* Given a correlated strategy $\mu$, player $i$'s expected utility when the players' actions are selected according to $\mu$ is

$$U_i(\mu) = \sum_{a \in A} \mu(a) u_i(a).$$

A coalition of players $S$ is a member of $2^N$. When $S$ consists of a single player $i \in N$, we write it as "$i$" rather than the more cumbersome $\{i\}$. For each $S \in 2^N$, $S \neq \emptyset$, denote by $A_S$ the set $\prod_{i \in S} A_i$. Given $a \in A$, we write $a = (a_S, a_{-S})$ where $a_S \in A_S$ and $a_{-S} \in A_{-S}$. If $S = N$, then $(a_S, a_{-S}) = a_S = a$.

### Coalition-Proof Correlated Equilibrium

We conceive of communication and play as proceeding in two stages. In the first stage players communicate, reaching an agreement, and possibly planning deviations from the agreement. Given an agreement $\mu \in \triangle A$, the players implement it with the assistance of a mediator who recommends the action profile $a \in A$ with probability $\mu(a)$. In the second stage, each player privately receives his component of the recommendation and then chooses an action. (No further communication occurs in this stage.)

A deviation by a coalition is a plan for its members to correlate their play in a way different from that prescribed by the agreement. We take a broad view of the ability of coalitions to plan deviations: for every different profile of recommendations received by its members, a deviating coalition may plan a different correlated strategy. Therefore, a deviation for a coalition $S$ is a mapping from the set $A_S$ of profiles of recommendations for its members, to the set $\triangle A_S$ of probability distributions on the set of the coalition's action profiles.

Given an agreement $\bar{\mu}$, if a coalition $S$ plans to deviate according to $\eta_S : A_S \to \triangle A_S$ (while the members of the complement of $S$ play their part of the agreement; i.e. they obey their recommendations), then the induced probability distribution over action profiles for the grand coalition is given for each $a \in A$ by

$$\mu(a) = \sum_{\alpha_S \in A_S} \bar{\mu}(\alpha_S, a_{-S}) \eta_S(a_S | \alpha_S).$$

It will be convenient to define the feasible deviations for coalition $S$ as those correlated strategies $\mu \in A$ which the coalition can induce, rather than as mappings from $A_S$ to $\triangle A_S$. Thus, a correlated strategy is a feasible deviation by coalition $S$ from a given agreement if the members of $S$, using

some plan to correlate their play, can induce the correlated strategy when each member of the complementary coalition obeys his recommendation.

**Definition 1.1.** Let $\bar{\mu} \in \triangle A$ and $S \in 2^N$, $S \neq \emptyset$. We say that $\mu \in \triangle A$ *is a feasible deviation by coalition* $S$ *from* $\bar{\mu}$ if there is a $\eta_S : A_S \to \triangle A_S$, such that for all $a \in A$, we have $\mu(a) = \sum_{\alpha_S \in A_S} \bar{\mu}(\alpha_S, a_{-S})\eta_S(a_S|\alpha_S)$ .

We illustrate our definition of a feasible deviation by describing a procedure that can be thought of as mimicking the process by which players select agreements and plan deviations. Given an agreement $\bar{\mu}$, suppose that the mediator implementing $\bar{\mu}$ mails a sealed envelop to each player containing the player's recommendation. A coalition $S$ deviates from $\bar{\mu}$ by employing a new mediator to which each member of $S$ sends the (unopened) envelop it received from the mediator implementing $\bar{\mu}$. The new mediator opens the envelops, reads the recommendations $\alpha_S$, and then selects a new profile of recommendations according to the correlated strategy $\eta_S(\alpha_S)$. The mediator then mails to each player $i \in S$ a sealed envelop containing his recommended action. When each player opens his envelop and obeys the recommendation it contains, the induced correlated strategy is given by the equation in Definition 1.1.

Given a coalition $S \in 2^N$, $S \neq \emptyset$, and an agreement $\mu \in \triangle A$, let $D(\mu, S)$ denote the set of feasible deviations by coalition $S$ from $\mu$; note that $\mu \in D(\mu, S)$, since a coalition always has the trivial "deviation" consisting of each member of the coalition obeying his own recommendation. Also note that for every $\mu \in \triangle A$, we have $D(\mu, N) = \triangle A$. A correlated equilibrium is a correlated strategy from which no individual has a feasible improving deviation.

**Correlated Equilibrium.** A correlated strategy $\mu$ is a *correlated equilibrium* if no individual $i \in N$, has a feasible deviation $\tilde{\mu} \in D(\mu, i)$, such that $U_i(\tilde{\mu}) > U_i(\mu)$.

The definition of strong Nash equilibrium suggests the following definition of strong correlated equilibrium[1]: a strong correlated equilibrium is a correlated strategy from which no coalition has a deviation which makes every member of the coalition better off.

[1]A notion of strong correlated equilibrium was informally proposed in Moulin [8].

**Definition 1.2.** A correlated strategy $\mu \in \triangle A$ is a *strong correlated equilibrium* if no coalition $S \in 2^N$, $S \neq \emptyset$, has a feasible deviation $\tilde{\mu} \in D(\mu, S)$, such that for each $i \in S$, we have $U_i(\tilde{\mu}) > U_i(\mu)$.

The agreement described in the introduction for the *Three Player Matching Pennies* game is, for example, the unique strong correlated equilibrium of that game. Like strong Nash equilibrium, the notion of strong correlated equilibrium is too strong. A strong correlated equilibrium must be resistant to *any* feasible deviation by *any* coalition. In particular, it must be resistant to deviations which are not themselves resistant to further deviations. Consider, for example, the *Prisoner's Dilemma* game.

| | C | D |
|---|---|---|
| C | 1,1 | -1,2 |
| D | 2,-1 | 0,0 |

This game has a unique correlated equilibrium where $(D, D)$ is played with probability one. This correlated equilibrium is not a strong correlated equilibrium since the correlated strategy $\tilde{\mu}$ consisting of playing $(C, C)$ with probability one is a feasible deviation which makes both players better off. Since a strong correlated equilibrium must be a correlated equilibrium, this game has no strong correlated equilibrium. Notice, however, that either player can unilaterally deviate from $\tilde{\mu}$ and increase his payoff. Hence $\tilde{\mu}$ should not undermine an agreement to play $(D, D)$ with probability one.

In order to be able to distinguish those deviations that are viable from those that are not (and which therefore should not upset an agreement as coalition-proof), we introduce the notion of self-enforcing deviation: a correlated strategy $\mu$ is a self-enforcing deviation by coalition $S$ from correlated strategy $\bar{\mu}$ if $\mu$ is a feasible deviation and if no proper subcoalition of $S$ has a further self-enforcing and improving deviation. This notion of self-enforcingness is identical to the one implicit in the concept of $CPNE$.

**Definition 1.3.** Let $\bar{\mu} \in \triangle A$ and $S \in 2^N$, $S \neq \emptyset$. *The set of self-enforcing deviations by coalition $S$ from $\bar{\mu}$*, $SED(\bar{\mu}, S)$, is defined, recursively, as follows.

(i) If $\mid S \mid = 1$, then $SED(\bar{\mu}, S) = D(\bar{\mu}, S)$;

(ii) If $\mid S \mid > 1$, then $SED(\bar{\mu}, S) = \{\mu \in D(\bar{\mu}, S) \mid \nexists\, [R \in 2^S \backslash S,\ R \neq \emptyset,\ \tilde{\mu} \in SED(\mu, R)]$ such that $\forall i \in R : U_i(\tilde{\mu}) > U_i(\mu)\}$.

Since a coalition consisting of a single player has no proper (nonempty) subcoalitions, any feasible deviation by a one-player coalition is self-enforcing. With this notion of a self-enforcing deviation, a coalition-proof correlated equilibrium is defined to be a correlated strategy from which no coalition has a self-enforcing and improving deviation.

**Definition 1.4.** A correlated strategy $\mu$ is a *coalition-proof correlated equilibrium* if no coalition $S \in 2^N$, $S \neq \emptyset$, has a deviation $\tilde{\mu} \in SED(\mu, S)$, such that for each $i \in S$, we have $U_i(\tilde{\mu}) > U_i(\mu)$.

It is clear that a strong correlated equilibrium is a coalition-proof correlated equilibrium, which in turn is a correlated equilibrium. For two-player games the set of coalition-proof correlated equilibria is the set of correlated equilibria which are not Pareto dominated by other correlated equilibria. Thus, for two-player games, the set of coalition-proof correlated equilibria is nonempty. Although existence of a $CPCE$ cannot be guaranteed in general games, we have identified conditions under which a $CPCE$ exists.

### On the Existence of Coalition-Proof Correlated Equilibrium

We show that a $CPCE$ exists whenever there is a correlated strategy whose support is the set of action profiles that survive iterated elimination of strictly dominated strategies and which Pareto dominates every other correlated strategy with support in this set. First, we define formally the notion of strict dominance.

**Definition 1.5.** Let $B = \prod_{i \in N} B_i \subset A$ arbitrary. An action $\bar{a}_j \in B_j$ is said to be *strictly dominated* in $B$ if there is $\sigma_j \in \triangle B_j$ such that for each $a_{-j} \in B_{-j}$

$$\sum_{a_j \in B_j} \sigma_j(a_j) u_j(a_j, a_{-j}) > u_i(\bar{a}_j, a_{-j}).$$

Note that if $\bar{a}_j$ is strictly dominated in $B$, then it is also strictly dominated in $A_j \times B_{-j}$. The set of action profiles that survive iterated elimination of strictly dominated strategies, which we write as $A^\infty$, is now easily defined.

**Definition 1.6.** *The set $A^\infty$ of action profiles that survive the iterated elimination of strictly dominated strategies,* is defined by $A^\infty = \prod_{i\in N} A_i^\infty$, where each $A_i^\infty = \bigcap_{n=0}^\infty A_i^n$, $A_i^n$ is the set of actions that are not strictly dominated in $A^{n-1} = \prod_{i\in N} A_i^{n-1}$, and $A_i^0 = A_i$.

The following proposition establishes that only correlated strategies whose support is $A^\infty$ can be self-enforcing deviations from a correlated strategy whose support is $A^\infty$. For each $\mu \in \Delta A$ and $S \in 2^N$, $S \neq \emptyset$, write $A_S^+(\mu)$ for the set $\{a_S \in A_S \mid \mu(a_S, a_{-S}) > 0 \text{ some } a_{-S} \in A_{-S}\}$, and write $A^+(\mu)$ for the set $A_N^+(\mu)$.

**Proposition.** *Let $S \in 2^N$, $S \neq \emptyset$, and let $\mu \in \Delta A$ be such that $A^+(\mu) \subset A^\infty$. If $\tilde{\mu} \in SED(\mu, S)$, then $A^+(\tilde{\mu}) \subset A^\infty$.*

**Proof:** Let $S$ and $\mu \in \Delta A$ be as in the Proposition, and let $\tilde{\mu} \in SED(\mu, S)$. By the definition of feasible deviation (Definition 1.1) $A^+(\tilde{\mu}) \subset A_S \times A_{-S}^\infty$. We show that in fact $A^+(\tilde{\mu}) \subset A^\infty$. Suppose by way of contradiction that $A_S^+(\tilde{\mu}) \not\subset A_S^\infty$. Let $n^*$ be the largest $n$ such that $A_S^+(\tilde{\mu}) \subset A_S^n$. Hence there is $j \in S$ and $\bar{a}_j \in A_j^+(\tilde{\mu})$ such that $\bar{a}_j$ is strictly dominated in $A^{n^*}$. Thus $\bar{a}_j$ is also strictly dominated in $A_j \times A_{-j}^{n^*}$; i.e., there is $\sigma_j \in \Delta A_j$ such that for each $a_{-j} \in A_{-j}^{n^*}$

$$\sum_{a_j \in A_j} \sigma_j(a_j) u_j(a_j, a_{-j}) > u_i(\bar{a}_j, a_{-j}). \qquad (*)$$

Consider the deviation $\mu'$ by player $j$ from $\tilde{\mu}$ where player $j$ chooses an action according to $\sigma_j$ when recommended $\bar{a}_j$, and takes the recommended action otherwise. Formally, the deviation $\eta_j$ is defined as follows: for each $a_j \in A_j$ such that $a_j \neq \bar{a}_j$, let $\eta_j(a_j \mid \alpha_j) = 1$ if $a_j = \alpha_j$, and $\eta_j(a_j \mid \alpha_j) = 0$ if $a_j \neq \alpha_j$; for $\alpha_j = \bar{a}_j$, let $\eta_j(a_j \mid \alpha_j) = \sigma_j(a_j)$. Again by the definition of feasible deviation $A^+(\mu') \subset A_j \times A_{-j}^{n^*}$. Then

$$U_j(\mu') = \sum_{a\in A} \mu'(a) u_j(a) = \sum_{a \in A_j \times A_{-j}^{n^*}} \left( \sum_{\alpha_j \in A_j} \tilde{\mu}(\alpha_j, a_{-j}) \eta_j(a_j \mid \alpha_j) \right) u_j(a).$$

substituting $\eta_j$ as defined above we have

$$U_j(\mu') = \sum_{a \in (A_j \setminus \{\bar{a}_j\}) \times A_{-j}^{n^*}} \tilde{\mu}(a) u_j(a) + \sum_{a \in \{\bar{a}_j\} \times A_{-j}^{n^*}} \tilde{\mu}(\bar{a}_j, a_{-j}) \left( \sum_{\alpha_j \in A_j} \sigma_j(\alpha_j) u_j(\alpha_j, a_{-j}) \right).$$

Since $\tilde{\mu}(\bar{a}_j, a_{-j}) > 0$ for some $a_{-j} \in A_{-j}^{n^*}$, equation (*) implies

$$U_j(\mu') > \sum_{a \in (A_j \setminus \{\bar{a}_j\}) \times A_{-j}^{n^*}} \tilde{\mu}(a) u_j(a) + \sum_{a \in \{\bar{a}_j\} \times A_{-j}^{n^*}} \tilde{\mu}(\bar{a}_j, a_{-j}) u_j(\bar{a}_j, a_{-j});$$

i.e.,

$$U_j(\mu') > \sum_{a \in A} \tilde{\mu}(a) u_j(a) = U_j(\tilde{\mu}).$$

Hence $\tilde{\mu}$ is not a self-enforcing deviation by $S$ from $\mu$; i.e., $\tilde{\mu} \notin SED(\mu, S)$. This contradiction establishes that $A^+(\tilde{\mu}) \subset A^\infty$. □

Suppose that $\mu$ is a correlated strategy with support in $A^\infty$ and that $\mu$ weakly Pareto dominates any other correlated strategy with support in $A^\infty$ (i.e., for each $\tilde{\mu}$ such that $A^+(\tilde{\mu}) \subset A^\infty$ we have $U_i(\mu) \geq U_i(\tilde{\mu})$, for each $i \in N$). Then the support of any feasible and improving deviation by a coalition $S$ from $\mu$ cannot be contained in $A^\infty$, and by the above proposition such a deviation is not self-enforcing. Therefore, we obtain the following corollary.

**Corollary.** *Let $\mu \in \triangle A$ be such that $A^+(\mu) \subset A^\infty$ and such that it weakly Pareto dominates every other $\tilde{\mu} \in \triangle A$ for which $A^+(\tilde{\mu}) \subset A^\infty$. Then $\mu$ is a $CPCE$.*

*Dominance solvable* games are those for which the set $A^\infty$ is a singleton. Our corollary implies that a dominance solvable game has an attractive property: its unique equilibrium is coalition-proof (i.e. it is a $CPCE$).

In Appendix B we show that the set of coalition-proof Nash equilibria of a game can be characterized as the set of mixed strategies from which no coalition has a self-enforcing deviation which makes all its members better off. The proposition above is easily modified to show that if a mixed strategy profile $\sigma$ has a support in $A^\infty$ then any self-enforcing mixed deviation from $\sigma$ also has a support in $A^\infty$. Thus, a $CPNE$ exists in dominance solvable games. In fact, a $CPNE$ exists whenever there is a mixed strategy profile in $A^\infty$ which weakly Pareto dominates every other mixed strategy profile in $A^\infty$.[2]

### A Game Where a $CPCE$ does not exist

Unfortunately, as the following example shows, there are games with more than two players with no coalition-proof correlated equilibria. Consider the following three-player game, taken from Einy and Peleg, where player 1 chooses the row, player 2 chooses the column, and player 3 chooses the matrix.

[2]Paul Milgrom has reported that for games with strategic complementarities if either (1) the equilibrium is unique or (2) the Pareto ranking theorem applies, then the Pareto-best Nash equilibrium is also coalition-proof.

| 3,2,0 | 0,0,0 |
|---|---|
| 2,0,3 | 2,0,3 |

| 3,2,0 | 0,3,2 |
|---|---|
| 0,0,0 | 0,3,2 |

We show that there does not exist a coalition-proof correlated equilibrium of this game. Let $\bar{\mu}$ be an arbitrary correlated equilibrium and suppose that player 1 has the lowest payoff of the three players. Then $U_3(\bar{\mu}) \leq \frac{13}{5}$. (This can be proven by maximizing player 3's utility over the set of correlated equilibria $\mu$ satisfying $U_1(\mu) \leq \max\{U_2(\mu), U_3(\mu)\}$.) Moreover, $U_1(\bar{\mu}) \leq \frac{13}{5}$ since player 1 has the lowest payoff. Now consider the following deviation from $\bar{\mu}$ by players 1 and 3: player 1 chooses the bottom row and player 3 chooses the left matrix. This deviation is improving as players 1 and 3 now receive payoffs of 2 and 3, respectively. To demonstrate that $\bar{\mu}$ is not a coalition-proof correlated equilibrium we need only show that this deviation is self-enforcing. Clearly player 3 does not deviate further as he now obtains 3, his highest possible payoff. It can be shown that player 1 obtains at most $\frac{5}{3}$ by deviating further and choosing the top row.[3] (The details of this calculation are in the appendix.) Thus, $\bar{\mu}$ is not a coalition-proof correlated equilibrium as players 1 and 3 have a self-enforcing and improving deviation.

There was no loss of generality in assuming that player 1 has the lowest payoff. If player 2 has the lowest payoff, then there is a self-enforcing and improving deviation by players 2 and 1. If player 3 has the lowest payoff, then there is a self-enforcing and improving deviation by players 3 and 2. Since any correlated equilibrium has a self-enforcing deviation by two players which makes both players better off, this game has no coalition-proof correlated equilibrium. (This game also does not have a $CPNE$.)

## 2. Games of Incomplete Information

In this section we extend our notion of coalition proofness to games of incomplete information. A (finite) game of incomplete information (or Bayesian game) $G$ is defined by

$$G = (N, (T_i)_{i \in N}, (A_i)_{i \in N}, (p_i)_{i \in N}, (u_i)_{i \in N}),$$

where $N$ is the set of players, and for each $i \in N$, $T_i$ is the set of possible types for player $i$, $A_i$ is player $i$'s action set, $p_i : T_i \to \Delta T_{-i}$ is player $i$'s prior probability distribution over the set of

[3]Following the deviation by players 1 and 3, player 1 is choosing the bottom row with probability one. Hence, when considering a further deviation by player 1 there is no loss of generality in restricting attention to the deviation where he chooses the top row with probability one. If this deviation does not make him better off, then no deviation does.

type profiles for the other players in the game ($T_{-i} = \prod_{j \in N \setminus \{i\}} T_j$), and $u_i : T \times A \to R$ is player $i$'s utility (payoff) function ($A = \prod_{i \in N} A_i$, $T = \prod_{i \in N} T_i$). We assume that the sets $N$, $A$, and $T$ are nonempty and finite. For every coalition of players $S \in 2^N$, $S \neq \emptyset$, we denote by $T_S$ the set $\prod_{i \in S} T_i$.

A *correlated strategy* is a function $\mu : T \to \triangle A$. We let $C$ denote the set of all correlated strategies. Given $\mu \in C$, if each player reports his type truthfully and obeys his recommendation, then player $i$'s expected payoff when he is of type $t_i \in T_i$ is

$$U_i(\mu|t_i) = \sum_{t_{-i} \in T_{-i}} \sum_{a \in A} p_i(t_{-i}|t_i)\mu(a|t)u_i(a,t).$$

Notice that in order for the players to play according to a correlated strategy, information about the players' types must be *revealed* so that an action profile can be selected according to the probability distribution specified by the given correlated strategy. We therefore must allow deviations by a coalition in which the players reveal a type profile different from their true one, as well as deviations where the players take actions different from those recommended. In the conceptual framework of mediation, the members of a coalition can deviate from a correlated strategy $\bar{\mu}$ by *misreporting* their type profile to the mediator or by *disobeying* the mediator's recommendations.

Intuitively, a deviation can be conceived of as follows: A coalition $S$ carries out a deviation by employing a new mediator who *represents* the coalition with the mediator implementing $\mu$ and with whom the members of $S$ communicate. Each member of $S$ reports his type to this mediator who then (1) selects according to some $f_S : T_S \to \triangle T_S$ a type profile for the coalition (which he reports to the mediator implementing $\mu$) and, upon receiving from the mediator implementing $\mu$ the recommendations for the members of $S$, (2) selects according to some $\eta_S : T_S \times T_S \times A_S \to \triangle A_S$ an action profile (which he recommends to the coalition members). The action profile recommended by the new mediator depends upon the type profile reported to it, the type profile it reported to the mediator implementing $\mu$, and the actions recommended by the mediator implementing $\mu$. This deviation generates a new correlated strategy which can be calculated from $f_S$ and $\eta_S$ according to the formula given in Definition 2.1.

**Definition 2.1.** Let $\bar{\mu} \in C$ and $S \in 2^N$, $S \neq \emptyset$. A correlated strategy $\mu$ is a *feasible deviation by coalition $S$ from $\bar{\mu}$* if there are $f_S : T_S \to \triangle T_S$, and $\eta_S : T_S \times T_S \times A_S \to \triangle A_S$, such that for each

$t \in T$ and each $a \in A$

$$\mu(a|t) = \sum_{\tau_S \in T_S} \sum_{\alpha_S \in A_S} f_S(\tau_S|t_S)\bar{\mu}(\alpha_S, a_{-S}|\tau_S, t_{-S})\eta_S(a_S|\tau_S, t_S, \alpha_S).$$

The set of feasible deviations by coalition $S$ from a correlated strategy is the set of correlated strategies that the coalition can induce by means of some $f_S$ and $\eta_S$. Given $\mu \in C$ and $S \in 2^N$, $S \neq \emptyset$, denote by $D(\mu, S)$ the set of all feasible deviations by coalition $S$ from correlated strategy $\mu$. As in Section 1, for every $\mu \in C$ and $S \in 2^N$, we have $\mu \in D(\mu, S)$ and $D(\mu, N) = C$.

For expositional ease, we explicitly introduce a concept of Pareto dominance: a correlated strategy $\tilde{\mu}$ *Pareto dominates* another correlated strategy $\mu$ *for coalition* $S$ if no member of $S$ is worse off under $\tilde{\mu}$ than under $\mu$ for any type profile, and if for at least one type profile every member of $S$ is better off under $\tilde{\mu}$ than under $\mu$.

**Definition 2.2.** Let $S \in 2^N$, $S \neq \emptyset$, and let $\mu, \tilde{\mu} \in C$. We say that $\tilde{\mu}$ *Pareto dominates* $\mu$ *for coalition* $S$ (or that $\tilde{\mu}$ *Pareto S-dominates* $\mu$) if

(2.1.1) For each $t_S \in T_S$, and each $i \in S$ : $U_i(\tilde{\mu}|t_i) \geq U_i(\mu|t_i)$, and

(2.1.2) There is $\tilde{t}_S \in T_S$ such that for each $i \in S$ : $U_i(\tilde{\mu}|\tilde{t}_i) > U_i(\mu|\tilde{t}_i)$.

In our framework, the notion of Pareto dominance used determines whether a deviation is an improvement for a coalition. Consequently, alternative notions of Pareto dominance will lead to different notions of coalition-proof communication equilibrium. There are two alternative notions worth considering.

We say that $\tilde{\mu}$ *weakly Pareto S-dominates* $\mu$ if no member of $S$ is worse off under $\tilde{\mu}$ than under $\mu$ for any type profile (i.e., if (2.1.1) is satisfied), and if at least one member of $S$ is better off under $\tilde{\mu}$ than under $\mu$ for some type profile (i.e., if (2.1.2) is satisfied for some $i \in S$ rather than for all $i \in S$). The notion of weak Pareto dominance does not seem appropriate: an agreement will be ruled out if a coalition has a self-enforcing deviation which makes only a proper subset of its members better off, even though there are not clear incentives for such a coalition to form.

We say that $\tilde{\mu}$ *strongly Pareto S-dominates* $\mu$ if the members of $S$ are better off under $\tilde{\mu}$ than under $\mu$ for all possible type profiles (i.e., if the inequalities (2.1.1) are satisfied with strict inequality). Strong Pareto dominance is sometimes too strong. For example, if the utility function of some player is constant for one of his types, then there is no deviation which is improving for

this player. Using strong Pareto dominance rules out the possibility of this player participating in any deviation.

It is easy to see that a correlated strategy $\mu$ is a *communication equilibrium* if no single player $i \in N$ has a feasible deviation which Pareto $i$-dominates $\mu$. In the spirit of the notion of strong Nash equilibrium, a strong communication equilibrium can be defined as follows: correlated strategy $\mu$ is a *strong communication equilibrium* if no coalition $S$ has a feasible deviation which Pareto $S$-dominates $\mu$. We only want to require, however, that an agreement not be Pareto dominated by self-enforcing deviations. The notion of self-enforcingness we define is identical to that introduced in Section 1.

**Definition 2.3.** Let $\bar{\mu} \in C$ and $S \in 2^N$, $S \neq \emptyset$. *The set of self-enforcing deviations by coalition* $S$ *from* $\bar{\mu}$, $SED(\bar{\mu}, S)$, is defined, recursively, as follows.

(i) If $\mid S \mid = 1$, then $SED(\bar{\mu}, S) = D(\bar{\mu}, S)$;

(ii) If $\mid S \mid > 1$, then $SED(\bar{\mu}, S) = \{\mu \in D(\bar{\mu}, S) \mid \nexists [R \in 2^S \backslash S, R \neq \emptyset, \tilde{\mu} \in SED(\mu, R)]$ such that $\tilde{\mu}$ Pareto $R$-dominates $\mu\}$.

With this notion of self-enforcingness, a coalition-proof communication equilibrium is defined to be any correlated strategy $\mu$ from which no coalition $S$ has a self-enforcing deviation which Pareto $S$-dominates $\mu$.

**Definition 2.4.** A correlated strategy $\mu$ is a *coalition-proof communication equilibrium* ($CPCE$) if no coalition $S \in 2^N, S \neq \emptyset$, has a self-enforcing deviation $\tilde{\mu} \in SED(\mu, S)$ such that $\tilde{\mu}$ Pareto $S$-dominates $\mu$.

When the set of type profiles $T$ is a singleton, the concepts of strong and coalition-proof *communication* equilibrium reduce to, respectively, strong and coalition-proof *correlated* equilibrium. Note that a strong communication equilibrium is a coalition-proof communication equilibrium, which in turn is a communication equilibrium.

In two-player Bayesian games, the set of coalition-proof communication equilibria consists of the communication equilibria that are not Pareto $N$-dominated by any other communication equilibrium (i.e., the set of *interim efficient* communication equilibria).[4] Hence, for two-player

[4]See Holmstrom and Myerson.

Bayesian games a $CPCE$ always exists. As established by example in Section 1, games with more than two players need not have a $CPCE$.

## 3. Discussion

In this section we discuss the relation of $CPCE$ to Einy and Peleg's notion of coalition-proof communication equilibrium (which we denote by $CPCE_{EP}$), and we present some concluding remarks.

In $CPCE$ deviations are evaluated *prior* to the players receiving recommendations: a deviation is improving if it makes each member of the deviating coalition better off, conditional on his type, for at least one of his types and no worse off for any of his types. In contrast, in $CPCE_{EP}$ deviations are considered *after* players receive recommendations: a deviation is improving if it makes each member of the deviating coalition better off, conditional on *both* his type and his recommendation, for each combination of types and recommendations that occur with positive probability.

Consequently, for two person games, while a $CPCE$ must be interim efficient a $CPCE_{EP}$ need not be. This is illustrated by the game *Chicken* given below by the left matrix. The right matrix describes the correlated equilibrium $\bar{\mu}$ which yields an expected payoff of 5 for each player.

| | L | R |
|---|---|---|
| T | 6,6 | 2,7 |
| B | 7,2 | 0,0 |

$\bar{\mu}$ :

| | L | R |
|---|---|---|
| T | 1/3 | 1/3 |
| B | 1/3 | 0 |

This correlated strategy is not a $CPCE$ as the grand coalition has the self-enforcing deviation $\tilde{\mu}$ given by

| | L | R |
|---|---|---|
| T | 1/2 | 1/4 |
| B | 1/4 | 0 |

which yields an expected payoff of 5.25 for each player. (The deviation $\tilde{\mu}$ is self-enforcing since it is a correlated equilibrium and therefore is immune to further deviations by a single player.)

Nonetheless, $\bar{\mu}$ is a $CPCE_{EP}$. In this game each player has only a single type; therefore, for a deviation to be improving in Einy and Peleg's sense, it must make each player better off, conditional on his recommendation, for each of his possible recommendations. Consider player 1

given the recommendation $B$. His expected payoff conditional on his recommendation is 7. Since 7 is player 1's highest possible payoff, no coalition involving player 1 can improve upon $\bar{\mu}$.[5]

One interpretation of E&P's framework is that players have the opportunity to communicate only after each player has received his recommendation. Thus, when determining whether or not an agreement is a $CPCE_{EP}$, the agreement is elevated to the position of a *status quo* agreement. It is required to be resistant to deviations following recommendations, but it is not confronted with alternative agreements which are improving at the stage prior to each player receiving his recommendation. If players have the opportunity to discuss their play prior to receiving recommendations, however, they will exhaust the opportunities for improvements at this stage. For the game *Chicken*, if the players must decide whether to play $\bar{\mu}$ or $\tilde{\mu}$, they should choose $\tilde{\mu}$ as, given that both are resistant to further deviations, $\tilde{\mu}$ gives a higher expected payoff to each player.

The second fundamental way in which the notions of coalition proofness differ is that Einy and Peleg do not admit the possibility that members of a coalition jointly "misreport" their types. A $CPCE_{EP}$ must be a communication equilibrium, and so a $CPCE_{EP}$ is immune to deviations where a *single* player misreports his type and disobeys his recommendation. However, in Einy and Peleg's framework, at the stage where deviations are considered, the players are assumed to have already truthfully reported their types. Thus, deviations may not involve the members of a coalition jointly misreporting their types, or involve one member of a coalition misreporting his type and another member of the coalition disobeying his recommendation. An example of a $CPCE_{EP}$ which fails to be immune to this latter kind of deviation is illustrated in the game of incomplete information below. The game is the same as the *Three Player Matching Pennies* game described in the Introduction, except that player 1's moves have now become his types.

$t_1 = H_1$ :

| | $H_3$ | $T_3$ |
|---|---|---|
| $H_2$ | 1,1,-2 | -1,-1,2 |
| $T_2$ | -1,-1,2 | -1,-1,2 |

$t_1 = T_1$ :

| | $H_3$ | $T_3$ |
|---|---|---|
| $H_2$ | -1,-1,2 | -1,-1,2 |
| $T_2$ | -1,-1,2 | 1,1,-2 |

Player 1 now has two possible types $\{H_1, T_1\}$ and no actions, while players 2 and 3 both have a singleton type set and their action sets remain, respectively, $\{H_2, T_2\}$ and $\{H_3, T_3\}$. Assume that the priors of players 2 and 3 over player 1's types are, respectively, $p_2(H_1) = p_3(H_1) = \frac{1}{2}$.

[5]It can be shown that there is no improving deviation upon $\bar{\mu}$ in E&P's sense even with the weaker requirement that a deviation makes each member of the deviating coalition better off for at least one recommendation and at least as well off for all recommendations.

The correlated strategy $\mu$ given by $\mu(H_2, T_3|H_1) = 1$ and $\mu(T_2, H_3|T_1) = 1$, is a communication equilibrium of the game which yields expected payoffs of $U_1(\mu|H_1) = U_1(\mu|T_1) = -1$, $U_2(\mu) = -1$, and $U_3(\mu) = 2$. It is also a $CPCE_{EP}$: in E&P's framework, a deviation by a coalition is a mapping from the set of type and action (recommendation) profiles for the coalition to probability distributions over the coalition's set of action profiles. The coalition $\{1,2\}$ has no improving deviation since, if player 1 is of type $H_1$, then player 3 moves $T_3$ with probability one and players 1 and 2 have a payoff of $-1$ regardless of the action taken by player 2. By the same argument, the coalition cannot improve if player 1 is of type $T_1$. No coalition involving player 3 has an improving deviation as the interests of players 1 and 2 are completely opposed to the interests of player 3. That no single player has an improving deviation follows from the fact that $\mu$ is a communication equilibrium.

In contrast, $\mu$ is not a $CPCE$ of the game. Consider the deviation by the coalition $\{1,2\}$ where player 1 reports $T_1$ when his type is $H_1$ and he reports $H_1$ when his type is $T_1$, and where player 2 moves $H_2$ when recommended $T_2$ and moves $T_2$ when recommended $H_2$. This deviation results in the correlated strategy $\tilde{\mu}$ given by $\tilde{\mu}(H_2, H_3|H_1) = \tilde{\mu}(T_2, T_3|T_1) = 1$, which yields expected payoffs of $U_1(\mu|H_1) = U_1(\mu|T_1) = 1$ and $U_2(\mu) = 1$. The deviation makes both players better off and is also self-enforcing (as both players attain their maximum possible payoff). Hence $\mu$ is not a $CPCE$.

Note that even if players can communicate only following the receipt of recommendations, $CPCE_{EP}$ assumes a certain myopia on the part of player 1. Consider again the $CPCE_{EP}$ of the *Three Player Matching Pennies* game where $\mu(H_2, T_3|H_1) = 1$ and $\mu(T_2, H_3|T_1) = 1$. If player 1 is of type $H_1$ and if he anticipates the opportunity to communicate following player 2's receipt of his recommendation, then player 1 should report type $T_1$ and, at the communication stage, suggest to player 2 that he should move $H_2$. Player 2 should follow player 1's suggestion given that his interests are coincident with player 1's.

This game has a unique $CPCE$ (which is also a $CPCE_{EP}$) where player 3 moves $H_3$ with probability $\frac{1}{2}$ regardless of player 1's type, and player 2 moves $H_2$ when player 1's type is $H_1$ and moves $T_2$ when player 1's type is $T_1$. This is essentially the same agreement predicted for the complete information version of the game. In fact, given that the interests of players 1 and 2 are coincident and opposed to those of player 3, this seems the only reasonable outcome.

We conclude by emphasizing our findings. First, we show that when players can communicate they will reach correlated agreements. For example, in the *Three Player Matching Pennies Game*

the only intuitive agreement is a correlated (and not mixed) agreement. Second, we offer a natural definition of coalition-proof equilibrium when correlated agreements are possible, and we show that no inclusion relationship between this new notion and $CPNE$ is to be found. (Consequently, the notion of coalition proofness is sensitive to the possibility of correlated agreements.) And third, we obtain conditions under which a coalition-proof equilibrium exists.

# Appendix A

In this appendix we present two examples. The first example is a game that has no coalition-proof correlated equilibrium. The second example is the *Three Player Matching Pennies* game; we show that the correlated strategy described in the introduction is the unique coalition-proof correlated equilibrium (and the unique strong correlated equilibrium) of the game.

### A game with no coalition proof correlated equilibrium

We show that the game below has no coalition-proof correlated equilibrium.

$c_1$

| | $b_1$ | $b_2$ |
|---|---|---|
| $a_1$ | 3,2,0 | 0,0,0 |
| $a_2$ | 2,0,3 | 2,0,3 |

$c_2$

| | $b_1$ | $b_2$ |
|---|---|---|
| $a_1$ | 3,2,0 | 0,3,2 |
| $a_2$ | 0,0,0 | 0,3,2 |

We represent a correlated strategy for this game as $\mu = (\mu_{ijk})_{i,j,k\in\{1,2\}}$, where $\mu_{ijk} \geq 0$ denotes the probability that players $1, 2$ and $3$ are recommended, respectively, actions $a_i$, $b_j$, and $c_k$.

If $\mu$ is a correlated equilibrium, then it satisfies the system of inequalities $(I)$ given by

$$
\begin{aligned}
(I.a_1) \quad & \mu_{111} - 2\mu_{121} + 3\mu_{112} \geq 0 \\
(I.a_2) \quad & -\mu_{211} + 2\mu_{221} - 3\mu_{212} \geq 0 \\
\\
(I.b_1) \quad & 2\mu_{111} - \mu_{112} - 3\mu_{212} \geq 0 \\
(I.b_2) \quad & -2\mu_{121} + \mu_{122} + 3\mu_{222} \geq 0 \\
\\
(I.c_1) \quad & -2\mu_{121} + 3\mu_{211} + \mu_{221} \geq 0 \\
(I.c_2) \quad & 2\mu_{122} - 3\mu_{212} - \mu_{222} \geq 0
\end{aligned}
$$

We show that for each correlated equilibrium there is a coalition of two players which has an improving and self-enforcing deviation. Therefore, since a coalition-proof correlated equilibrium must be a correlated equilibrium, the set of $CPCE$ of this game is empty.

Let $\bar{\mu}$ be an arbitrary correlated equilibrium and suppose that player 1 has the lowest payoff of the three players. We show that the coalition of players 1 and 3 has a self-enforcing and improving deviation. If player 1 has the lowest payoff in a correlated equilibrium, then player 3's payoff is no larger than $\frac{13}{5}$, which is the value of the solution to the linear programing problem

$$\max_{\mu \in \Delta A} U_3(\mu) \quad \text{subject to} \quad (I),\ U_1(\mu) \leq U_2(\mu), \text{ and } U_1(\mu) \leq U_3(\mu).$$

We also have $U_1(\bar{\mu}) \leq \frac{5}{3}$ since player 1 has the lowest payoff.

Consider the deviation $\tilde{\mu}$ induced by players 1 and 3 playing $(a_2, c_1)$ with probability one for each profile of recommendations. (Then $\tilde{\mu}_{211} = \mu_{111} + \mu_{211} + \mu_{112} + \mu_{212}$, $\tilde{\mu}_{221} = \mu_{121} + \mu_{221} + \mu_{122} + \mu_{222}$, and $\tilde{\mu}_{ijk} = 0$ otherwise.) Given this deviation, then regardless of player 2's action, players 1 and 3 obtain payoffs of, respectively, 2 and 3. Hence $U_1(\tilde{\mu}) = 2 > U_1(\bar{\mu})$ and $U_3(\tilde{\mu}) = 3 > U_3(\bar{\mu})$ and so $\tilde{\mu}$ is an improving deviation for $\{1, 3\}$.

We now show that $\tilde{\mu}$ is self-enforcing. Clearly player 3 does not have a further improving deviation as he obtains his highest possible payoff. Player 1 has an improving deviation if the expected payoff of deviating to $a_1$, which is $3\tilde{\mu}_{211}$, is greater than $U_1(\tilde{\mu}) = 2$ (his expected payoff when he follows a recommendation to play $a_2$). However, this payoff is not larger than $\frac{5}{3}$, which is the value of the solution to the linear programing problem

$$\max_{\mu \in \Delta A} 3(\mu_{111} + \mu_{211} + \mu_{112} + \mu_{212}) \quad \text{subject to} \quad (I),\ U_1(\mu) \leq U_2(\mu), \text{ and } U_1(\mu) \leq U_3(\mu).$$

The value of this maximization problem is the maximum payoff that player 1 can obtain by a further deviation to $a_1$ from the correlated strategy $\tilde{\mu}$ given then the original agreement $\mu$ was a correlated equilibrium in which player 1 had the lowest payoff. Hence player 1 has no further improving deviation.

There was no loss of generality in assuming that player 1 has the lowest payoff. Given the symmetry of this game, we can construct the following self-enforcing and improving deviations in each case: If player 2 has the lowest payoff, then players 1 and 2 deviate to $\{a_1, b_1\}$. If player 3 has the lowest payoff, then players 2 and 3 deviate to $\{b_2, c_2\}$. Therefore, this game has no coalition-proof correlated equilibrium.

### Three Player Matching Pennies

The payoff matrix for the *Three Player Matching Pennies* game is given by

$H_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | 1,1,-2 | -1,-1,2 |
| $T_1$ | -1,-1,2 | -1,-1,2 |

$T_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | -1,-1,2 | -1,-1,2 |
| $T_1$ | -1,-1,2 | 1,1,-2 |

In the Introduction we demonstrated that the correlated strategy $\mu^*$ given in the table below is a strong correlated equilibrium.

$H_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | $\frac{1}{4}$ | 0 |
| $T_1$ | 0 | $\frac{1}{4}$ |

$T_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | $\frac{1}{4}$ | 0 |
| $T_1$ | 0 | $\frac{1}{4}$ |

We now establish that $\mu^*$ is the unique coalition-proof correlated equilibrium of this game. (A strong correlated equilibrium is also a coalition-proof correlated equilibrium, therefore $\mu^*$ is also the unique strong correlated equilibrium). Let $\mu$ be any correlated strategy. We reduce notation by writing for $\mu_{xyz}$ for the probability $\mu(x_1, y_2, z_3)$, where $(x_1, y_2, z_3) \in \{H_1, T_1\} \times \{H_2, T_2\} \times \{H_3, T_3\}$; e.g., we write $\mu_{TTH}$ for $\mu(T_1, T_2, H_3)$. If $\mu$ is a correlated equilibrium, then it must satisfy the system of inequalities (1) given by

$$\begin{aligned}
(1.H_1) \quad & 2\mu_{HHH} - 2\mu_{HTT} \geq 0 \\
(1.T_1) \quad - & 2\mu_{THH} + 2\mu_{TTT} \geq 0 \\
\\
(1.H_2) \quad & 2\mu_{HHH} - 2\mu_{THT} \geq 0 \\
(1.T_2) \quad - & 2\mu_{HTH} + 2\mu_{TTT} \geq 0 \\
\\
(1.H_3) \quad - & 4\mu_{HHH} + 4\mu_{TTH} \geq 0 \\
(1.T_3) \quad & 4\mu_{HHT} - 4\mu_{TTT} \geq 0
\end{aligned}$$

Note that $(1.H_3)$ implies $\mu_{TTH} \geq \mu_{HHH}$, and $(1.T_3)$ implies $\mu_{HHT} \geq \mu_{TTT}$. Hence player 3's payoff,

$$U_3(\mu) = 2(-\mu_{HHH} + \mu_{HTH} + \mu_{THH} + \mu_{TTH} + \mu_{HHT} + \mu_{HTT} + \mu_{THT} - \mu_{TTT}),$$

satisfies $U_3(\mu) \geq 0$. Since for each $(x_1, y_2, z_3) \in \{H_1, T_1\} \times \{H_2, T_2\} \times \{H_3, T_3\}$, $u_1(x_1, y_2, z_3) + u_2(x_1, y_2, z_3) + u_3(x_1, y_2, z_3) = 0$, we have $U_1(\mu) = U_2(\mu) \leq 0$.

We now establish the following result.

CLAIM: If $\mu$ is a $CPCE$, then $U_1(\mu) = U_2(\mu) = 0$, and $\mu_{HHH} + \mu_{HTH} + \mu_{THH} + \mu_{TTH} = \frac{1}{2}$.

PROOF: Let $\mu$ be a coalition-proof correlated equilibrium. We have shown that in any correlated equilibrium $U_1(\mu) = U_2(\mu) \leq 0$. Suppose by way of contradiction that $U_1(\mu) = U_2(\mu) < 0$. Consider the deviation where players 1 and 2 play $(H_1, H_2)$ with probability $\frac{1}{2}$ and $(T_1, T_2)$ with probability $\frac{1}{2}$, regardless of their recommendations. This deviation induces the correlated strategy $\bar{\mu}$ given by

$H_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | $\frac{\lambda}{2}$ | 0 |
| $T_1$ | 0 | $\frac{\lambda}{2}$ |

$T_3$

| | $H_2$ | $T_2$ |
|---|---|---|
| $H_1$ | $\frac{1-\lambda}{2}$ | 0 |
| $T_1$ | 0 | $\frac{1-\lambda}{2}$ |

where $\lambda = \mu_{HHH} + \mu_{HTH} + \mu_{THH} + \mu_{TTH}$. This deviation is improving since $U_1(\bar{\mu}) = U_2(\bar{\mu}) = 0$. It it also self-enforcing since a further deviation by either player 1 or 2 makes the player strictly worse off. The existence of such a deviation contradicts that $\mu$ is a $CPCE$. Thus, we must have $U_1(\mu) = U_2(\mu) = 0$; i.e.,

$$(2) \qquad \mu_{HHH} - \mu_{HTH} - \mu_{THH} - \mu_{TTH} - \mu_{HHT} - \mu_{HTT} - \mu_{THT} + \mu_{TTT} = 0.$$

We now show that $\lambda = \frac{1}{2}$. Suppose that $\lambda > \frac{1}{2}$. Then the deviation by players 1 and 2 where, regardless of their recommendation, they move $(H_1, H_2)$ with probability one is improving (players 1 and 2 each have an expected payoff of $\lambda + (1 - \lambda) > 0$) and self-enforcing, contradicting that $\mu$ is a $CPCE$. The case $\lambda < \frac{1}{2}$ is symmetric. Therefore $\lambda = \frac{1}{2}$; i.e.,

$$(3) \qquad \mu_{HHH} + \mu_{HTH} + \mu_{THH} + \mu_{TTH} = \frac{1}{2}. \square$$

Finally, we show that if $\mu$ is a $CPCE$, then $\mu_{HHH} = \mu_{HHT} = \mu_{TTT} = \mu_{TTH} = \frac{1}{4}$. As $\mu$ is a correlated strategy, we have

$$(4) \qquad \mu_{HHH} + \mu_{HTH} + \mu_{THH} + \mu_{TTH} + \mu_{HHT} + \mu_{HTT} + \mu_{THT} + \mu_{TTT} = 1.$$

Adding (2) and (4) we get

$$(*) \qquad \mu_{HHH} + \mu_{TTT} = \frac{1}{2}.$$

Also $(1.H_3)$ and $(1.T_3)$ yield $\mu_{TTH} \geq \mu_{HHH}$, and $\mu_{HHT} \geq \mu_{TTT}$. Adding these two inequalities and noticing (4) we get

$$(**) \qquad \mu_{TTH} + \mu_{HHT} = \frac{1}{2}.$$

Thus, (4) implies $\mu_{HTH} = \mu_{THH} = \mu_{HTT} = \mu_{THT} = 0$. Substituting in (3) we get

$$(***) \qquad \mu_{HHH} + \mu_{TTH} = \frac{1}{2}.$$

Subtracting $(***)$ from $(*)$ we get $\mu_{TTT} - \mu_{TTH} = 0$; i.e., $\mu_{TTT} = \mu_{TTH}$. Substituting in $(**)$ and subtracting $(*)$, we have $\mu_{HHT} - \mu_{HHH} = 0$; i.e., $\mu_{HHT} = \mu_{HHH}$. Using $(1.H_3)$ and $(1.T_3)$ again implies

$$\mu_{HHH} = \mu_{HHT} \geq \mu_{TTT} = \mu_{TTH} \geq \mu_{HHH}.$$

Hence $\mu_{HHH} = \mu_{HHT} = \mu_{TTT} = \mu_{TTH} = \frac{1}{4}$.□

## Appendix B

In this appendix we prove Proposition B.1 which characterizes the set of coalition-proof Nash equilibria as the set of mixed strategies from which no coalition has a self-enforcing deviation which makes all its members better off.

For $S \in 2^N$, $S \neq \emptyset$, let $\Sigma_S$ denote the set of probability distributions $\sigma_S$ over $A_S = \prod_{i \in S} A_i$ satisfying $\sigma_S(a_S) = \prod_{i \in S} \sigma_i(a_i)$ for all $a_S \in A_S$, where $\sigma_i(a_i) = \sum_{a_{S\setminus\{i\}} \in A_{S\setminus\{i\}}} \sigma_S(a_{S\setminus\{i\}}, a_i)$ is the marginal distribution of $\sigma_S$ over $A_i$. Write $\Sigma$ for the set $\Sigma_N$, and refer to its members as *mixed strategies*. If $\mid N \mid > 1$, then $\Sigma$ is a proper subset of $\triangle A$. Given $\sigma \in \Sigma$ and $S \in 2^N$, we denote by $\sigma_S$ the marginal distribution of $\sigma$ over $A_S$ (i.e., $\forall a_S \in A_S : \sigma_S(a_S) = \sum_{\alpha_{-S} \in A_{-S}} \sigma(a_S, \alpha_{-S})$ ). Here a mixed strategy is a probability distribution over $A$. A mixed strategy $\sigma \in \Sigma$ has an equivalent and more conventional representation as a strategy profile, $(\sigma_1, ..., \sigma_n)$.

Given an agreement $\bar{\sigma} \in \Sigma$, define the set of feasible *mixed* deviations by coalition $S$ from $\bar{\sigma}$ as those mixed strategies that are obtained when each player $i$, $i \in S$, randomizes independently according to some $\tilde{\sigma}_i$, while each player $j$, $j \in N\setminus S$, follows the agreement and randomizes according to $\bar{\sigma}_j$. In other words, $\sigma$ is a feasible deviation from $\bar{\sigma}$ by coalition $S$ if $\sigma$ can be written as a mixed strategy profile $((\tilde{\sigma}_i)_{i \in S}, (\bar{\sigma}_j)_{j \in N\setminus S})$ where $(\tilde{\sigma}_i)_{i \in S}$ is some mixed strategy profile for members of $S$. This is established formally in Definition B.1.

**Definition B.1.** Let $\bar{\sigma} \in \Sigma$ and $S \in 2^N$, $S \neq \emptyset$. We say that $\sigma \in \Sigma$ is a *feasible mixed deviation by coalition $S$ from $\bar{\sigma}$* if there is a $\tilde{\sigma}_S \in \Sigma_S$, such that for all $a \in A$, we have $\sigma(a) = \tilde{\sigma}_S(a_S)\bar{\sigma}_{-S}(a_{-S})$.

Let $D_M(\bar{\sigma}, S)$ denote the set of feasible mixed deviations by coalition $S$ from $\bar{\sigma}$. It is clear that a mixed strategy is a Nash equilibrium if no single player has a feasible mixed deviation which

makes him better off. A mixed strategy is a strong Nash equilibrium if no coalition has a feasible and improving mixed deviation.

The definition of a self-enforcing mixed deviation is obtained by replacing in Definition 1.3 the set of deviations with the set of mixed deviations. Hence, a mixed strategy $\sigma$ is a self-enforcing mixed deviation by coalition $S$ from $\bar{\sigma}$ if $\sigma$ is a feasible mixed deviation and if no proper subcoalition of $S$ has a further self-enforcing and improving mixed deviation from $\sigma$.

**Definition B.2.** Let $\bar{\sigma} \in \Sigma$ and $S \in 2^N$, $S \neq \emptyset$. *The set of self-enforcing mixed deviations by coalition $S$ from $\bar{\sigma}$*, $SED_M(\bar{\sigma}, S)$, is defined, recursively, as follows

(i) If $\mid S \mid = 1$, then $SED_M(\bar{\sigma}, S) = D_M(\bar{\sigma}, S)$;

(ii) If $\mid S \mid > 1$, then $SED_M(\bar{\sigma}, S) = \{\sigma \in D_M(\bar{\sigma}, S) \mid \nexists\, [R \in 2^S \backslash S,\ R \neq \emptyset,\ \tilde{\sigma} \in SED_M(\sigma, R)]$ such that $\forall i \in R : U_i(\tilde{\sigma}) > U_i(\sigma)\}$.

Using the notions of feasible and of self-enforcing deviation by a coalition from a mixed strategy, we define the notion of coalition-proof Nash equilibrium as follows.

**Definition B.3.** Let $\Gamma = (N, (A_i)_{i \in N}, (u_i)_{i \in N})$ be a game in strategic form. A strategy profile $\sigma \in \Sigma$ is a $CPNE'$ if no coalition $S \in 2^N$, $S \neq \emptyset$, has a self-enforcing mixed deviation $\tilde{\sigma} \in SED_M(\sigma, S)$ such that for each $i \in S$, we have $U_i(\tilde{\sigma}) > U_i(\sigma)$.

Definition B.4 below formalizes the concept of $CPNE$ as defined by Bernheim, Peleg and Whinston [3]. For convenience, the notion of $CPNE$ is cast in terms of *mixed strategies* (members of $\Sigma$) instead of *strategy profiles* (members of $\prod_{i=1}^{n} \Sigma_i$). We abuse notation sometimes by writing a mixed strategy $\sigma \in \Sigma$ as $(\sigma_S, \sigma_{-S})$, where $\sigma_S \in \Sigma_S$.

Let $\Gamma = (N, (A_i)_{i \in N}, (u_i)_{i \in N})$ be a game in strategic form. Given $\bar{\sigma} \in \Sigma$ and $S \in 2^N \backslash N$, $S \neq \emptyset$, we write $\Gamma / \bar{\sigma}_{-S}$ for the game $(S, (A_i)_{i \in S}, (\bar{u}_i)_{i \in S})$, where for each $i \in S$ and $a_S \in A_S$ we have

$$\bar{u}_i(a_S) = \sum_{\alpha_{-S} \in A_{-S}} \bar{\sigma}_{-S}(\alpha_{-S}) u_i(a_S, \alpha_{-S}).$$

For $S = N$, define $\Gamma / \bar{\sigma}_{-S} = \Gamma$.

The definition of $CPNE$ given by BPW is recursive.

**Definition B.4.** Let $\Gamma = (N, (A_i)_{i\in N}, (u_i)_{i\in N})$ be a game in strategic form.

(i) If $\mid N \mid = 1$, then $\sigma_1 \in \Sigma_1$ is a $CPNE$ if for every $\tilde{\sigma}_1 \in \Sigma_1 : U_1(\sigma_1) \geq U_1(\tilde{\sigma}_1)$.

(ii) Assume that $CPNE$ has been defined for games with fewer than $n$ players, and let $\Gamma$ be a game such that $\mid N \mid = n$.

(a) A mixed strategy $\sigma \in \Sigma$ is *self-enforcing* if for every $S \in 2^N \backslash N$, $S \neq \emptyset$, $\sigma_S$ is a $CPNE$ of $\Gamma/\sigma_{-S}$.

(b) A mixed strategy $\sigma \in \Sigma$ is a $CPNE$ of $\Gamma$ if $\sigma$ is a self-enforcing mixed strategy, and if there is no other self-enforcing mixed strategy $\tilde{\sigma}$ such that for every $i \in N : U_i(\tilde{\sigma}) > U_i(\sigma)$.

For every game in strategic form $\Gamma$ let $CPNE'(\Gamma)$ and $CPNE(\Gamma)$ represent the sets of mixed strategies satisfying, respectively, definitions B.3 and B.4. Also, we denote by $SE(\Gamma)$ the set of all self-enforcing mixed strategies of $\Gamma$. For each $\sigma \in \Sigma$, and each $S \in 2^N$, $S \neq \emptyset$, we write $SED_M^\Gamma(\sigma, S)$ for the set of self-enforcing mixed deviations from $\sigma$ by coalition $S$ in the game $\Gamma$ and we denote by $U_i^\Gamma(\sigma)$ the expected utility of player $i$ given mixed strategy $\sigma$ in the game $\Gamma$. Proposition B.1 can now be stated as follows.

**Proposition B.1.** *For every game $\Gamma$ in strategic form we have $CPNE(\Gamma) = CPNE'(\Gamma)$.*

Before proving the proposition, we establish two lemmas.

**Lemma B.1.** *For each $\Gamma$, each $\bar{\sigma} \in \Sigma$ and each $S \in 2^N$, $S \neq \emptyset$, we have that $\sigma = (\sigma_S, \bar{\sigma}_{-S}) \in SED_M^\Gamma(\bar{\sigma}, S)$ if and only if $\sigma_S \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$.*

**Proof:** We prove the lemma by induction on the number of players in $S$. Let $\Gamma$ be a strategic form game, $\bar{\sigma} \in \Sigma$ and $S \in 2^N$.

(i) If $S = \{i\}$, then $SED_M^\Gamma(\bar{\sigma}, S) = D_M^\Gamma(\bar{\sigma}, S) = \Sigma_S \times \{\bar{\sigma}_{-S}\}$ and $SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S) = D_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S) = \Sigma_S$. Therefore, $\sigma = (\sigma_S, \bar{\sigma}_{-S}) \in SED_M^\Gamma(\bar{\sigma}, S)$ if and only if $\sigma_S \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$.

(ii) Assume Lemma B.1 holds for $\mid S \mid < k$. We show that it holds for $\mid S \mid = k$.

STEP 1: If $\sigma_S \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$ then $\sigma = (\sigma_S, \bar{\sigma}_{-S}) \in SED_M^\Gamma(\bar{\sigma}, S)$.

Let $\sigma = (\sigma_S, \bar{\sigma}_{-S}) \notin SED_M^\Gamma(\bar{\sigma}, S)$. Then there are $R \in 2^S \backslash S$, $R \neq \emptyset$, and $\tilde{\sigma} = (\tilde{\sigma}_R, \sigma_{S\backslash R}, \bar{\sigma}_{-S}) \in SED_M^\Gamma((\sigma_S, \bar{\sigma}_{-S}), R)$ such that for each $i \in R$ we have $U_i^\Gamma(\tilde{\sigma}) > U_i^\Gamma(\sigma)$. Since $\mid R \mid < k$, the induction hypothesis yields $\tilde{\sigma}_R \in SED_M^{\Gamma/(\bar{\sigma}_{-S}, \sigma_{S\backslash R})}(\sigma_R, R)$. Noticing that $\Gamma/(\bar{\sigma}_{-S}, \sigma_{S\backslash R}) \equiv (\Gamma/\bar{\sigma}_{-S})/\sigma_{S\backslash R}$,

it also yields $(\tilde{\sigma}_R, \sigma_{S\setminus R}) \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\sigma_S, R)$. Moreover, for each $i \in R$ we have $U_i^{\Gamma/\bar{\sigma}_{-S}}(\tilde{\sigma}_R, \sigma_{S\setminus R}) = U_i^{\Gamma}(\tilde{\sigma}) > U_i^{\Gamma}(\sigma) = U_i^{\Gamma/\bar{\sigma}_{-S}}(\sigma_S)$. Hence $\sigma_S \notin SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$.

STEP 2: If $(\sigma_S, \bar{\sigma}_{-S}) \in SED_M^{\Gamma}(\bar{\sigma}, S)$ then $\sigma_S \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$.

Let $\sigma_S \notin SED_M^{\Gamma/\bar{\sigma}_{-S}}(\bar{\sigma}_S, S)$. Then there are $R \in 2^S \setminus S$, $R \neq \emptyset$, and $\tilde{\sigma}_S = (\tilde{\sigma}_R, \sigma_{S\setminus R}) \in SED_M^{\Gamma/\bar{\sigma}_{-S}}(\sigma_S, R)$ such that for each $i \in R$ we have $U_i^{\Gamma/\bar{\sigma}_{-S}}(\tilde{\sigma}_S) > U_i^{\Gamma/\bar{\sigma}_{-S}}(\sigma_S)$. The induction hypothesis implies that $\tilde{\sigma}_R \in SED_M^{(\Gamma/\bar{\sigma}_{-S})/\sigma_{S\setminus R}}(\sigma_R, R)$. Since $(\Gamma/\bar{\sigma}_{-S})/\sigma_{S\setminus R} \equiv \Gamma/(\bar{\sigma}_{-S}, \sigma_{S\setminus R})$ it also implies that $\tilde{\sigma} = (\tilde{\sigma}_R, \sigma_{S\setminus R}, \bar{\sigma}_{-S}) \in SED_M^{\Gamma}((\sigma_S, \bar{\sigma}_{-S}), R)$. Furthermore, for each $i \in R$ we have $U_i^{\Gamma}(\tilde{\sigma}) = U_i^{\Gamma/\bar{\sigma}_{-S}}(\tilde{\sigma}_S) > U_i^{\Gamma/\bar{\sigma}_{-S}}(\sigma_S) = U_i^{\Gamma}(\sigma_S, \bar{\sigma}_{-S})$. Therefore $(\sigma_S, \bar{\sigma}_{-S}) \notin SED_M^{\Gamma}(\bar{\sigma}, S)$.□

**Lemma B.2.** *Let $\Gamma$ be a game in strategic form. For each $\sigma \in CPNE'(\Gamma)$ and each $S \in 2^N$, $S \neq \emptyset$, we have $\sigma_S \in CPNE'(\Gamma/\sigma_{-S})$.*

**Proof:** Let $\sigma \in \Sigma$ be such that $\sigma_S \notin CPNE'(\Gamma/\sigma_{-S})$ for some $S \in 2^N$, $S \neq \emptyset$. We show that $\sigma \notin CPNE'(\Gamma)$.

Since $\sigma_S \notin CPNE'(\Gamma/\sigma_{-S})$, then there are $S' \in 2^S$, $S' \neq \emptyset$, and $\tilde{\sigma}_S = (\tilde{\sigma}_{S'}, \sigma_{S\setminus S'}) \in SED_M^{\Gamma/\sigma_{-S}}(\sigma_S, S')$ such that for each $i \in S'$, we have $U_i^{\Gamma/\sigma_{-S}}(\tilde{\sigma}_S) > U_i^{\Gamma/\sigma_{-S}}(\sigma_S)$. By Lemma B.1, $(\tilde{\sigma}_{S'}, \sigma_{S\setminus S'}) \in SED_M^{\Gamma/\sigma_{-S}}(\sigma_S, S')$ implies that $\tilde{\sigma} = (\tilde{\sigma}_{S'}, \sigma_{S\setminus S'}, \sigma_{-S}) \in SED_M^{\Gamma}(\sigma, S')$. Moreover, for each $i \in S'$ we have $U_i^{\Gamma}(\tilde{\sigma}) = U_i^{\Gamma/\sigma_{-S}}(\tilde{\sigma}_S) > U_i^{\Gamma/\sigma_{-S}}(\sigma_S) = U_i^{\Gamma}(\sigma)$. Hence $\sigma \notin CPNE'(\Gamma)$.□

**Proof of Proposition B.1:** We prove the proposition by induction on the number of players.
(i) If $\mid N \mid = 1$ then Proposition B.1 clearly holds as for each $\sigma_1 \in \Sigma_1$, we have $SED_M^{\Gamma}(\sigma_1, \{1\}) = \Sigma_1$.
(ii) Assume that Proposition B.1 is satisfied for games with fewer than $n$ players. We need to show that it holds for $\Gamma$ with $\mid N \mid = n$.

STEP 1: If $\sigma \in CPNE(\Gamma)$ then $\sigma \in CPNE'(\Gamma)$.

Let $\sigma \notin CPNE'(\Gamma)$. Then there is a $S \in 2^N$, $S \neq \emptyset$, and a $\tilde{\sigma} = (\tilde{\sigma}_S, \sigma_{-S}) \in SED_M^{\Gamma}(\sigma, S)$ such that for each $i \in S$, we have $U_i^{\Gamma}(\tilde{\sigma}) > U_i^{\Gamma}(\sigma)$.

Case (a): $S \neq N$. Since $\tilde{\sigma} \in SED_M^{\Gamma}(\sigma, S)$, by Lemma B.1 $\tilde{\sigma}_S \in SED_M^{\Gamma/\sigma_{-S}}(\sigma_S, S)$. Moreover, $U_i^{\Gamma}(\tilde{\sigma}) = U_i^{\Gamma/\sigma_{-S}}(\tilde{\sigma}_S) > U_i^{\Gamma/\sigma_{-S}}(\sigma_S) = U_i^{\Gamma}(\sigma)$ for each $i \in S$. Hence $\sigma_S \notin CPNE'(\Gamma/\sigma_{-S}) = CPNE(\Gamma/\sigma_{-S})$ where the equality follows from the induction hypothesis and that the game $\Gamma/\sigma_{-S}$ has less than $n$ players. Therefore $\sigma \notin CPNE(\Gamma)$.

Case (b): $S = N$. Assume without loss of generality that $\nexists\, \hat{\sigma} \in SED_M^{\Gamma}(\sigma, N)$ such that for each $i \in N$, $U_i^{\Gamma}(\hat{\sigma}) > U_i^{\Gamma}(\tilde{\sigma})$. Then $\tilde{\sigma} \in CPNE'(\Gamma)$ and so by Lemma B.2, $\sigma_S \in CPNE'(\Gamma/\sigma_{-S}) =$

$CPNE(\Gamma/\sigma_{-S})$ $\forall S \in 2^N \backslash N$, $S \neq \emptyset$, where the equality follows from the induction hypothesis. Thus, $\tilde{\sigma} \in SE(\Gamma)$ and for each $i \in N$, $U_i^\Gamma(\tilde{\sigma}) > U_i^\Gamma(\sigma)$. Therefore, $\sigma \notin CPNE(\Gamma)$.

STEP 2: If $\sigma \in CPNE'(\Gamma)$ then $\sigma \in CPNE(\Gamma)$.

Let $\sigma \notin CPNE(\Gamma)$. If $\sigma \notin SE(\Gamma)$, then there is a $S \in 2^N \backslash N$, $S \neq \emptyset$, such that $\sigma_S \notin CPNE(\Gamma/\sigma_{-S}) = CPNE'(\Gamma/\sigma_{-S})$ where the equality follows from the induction hypothesis. But Lemma B.2 and $\sigma_S \notin CPNE'(\Gamma/\sigma_{-S})$ implies that $\sigma \notin CPNE'(\Gamma)$.

If $\sigma \in SE(\Gamma)$, then there is a $\tilde{\sigma} \in SE(\Gamma)$ such that for each $i \in N$, we have $U_i^\Gamma(\tilde{\sigma}) > U_i^\Gamma(\sigma)$. We show that $\tilde{\sigma} \in SED_M^\Gamma(\sigma, N)$, thereby proving that $\sigma \notin CPNE'(\Gamma)$.

Suppose to the contrary that $\tilde{\sigma} \notin SED_M^\Gamma(\sigma, N)$. Since $D_M^\Gamma(\sigma, N) = \Sigma$ (any deviation by the grand coalition is feasible), then there must be a $S \in 2^N \backslash N$, $S \neq \emptyset$, and a $\hat{\sigma} = (\hat{\sigma}_S, \tilde{\sigma}_{-S}) \in SED_M^\Gamma(\tilde{\sigma}, S)$ such that for each $i \in S : U_i^\Gamma(\hat{\sigma}) > U_i^\Gamma(\tilde{\sigma})$. Since $(\hat{\sigma}_S, \tilde{\sigma}_{-S}) \in SED_M^\Gamma(\tilde{\sigma}, S)$, Lemma B.1 yields $\hat{\sigma}_S \in SED_M^{\Gamma/\tilde{\sigma}_{-S}}(\tilde{\sigma}_S, S)$. Moreover, for each $i \in S$ we have $U_i^{\Gamma/\tilde{\sigma}_{-S}}(\hat{\sigma}_S) = U_i^\Gamma(\hat{\sigma}) > U_i^\Gamma(\tilde{\sigma}) = U_i^{\Gamma/\tilde{\sigma}_{-S}}(\tilde{\sigma}_S)$. Hence $\tilde{\sigma}_S \notin CPNE'(\Gamma/\tilde{\sigma}_{-S}) = CPNE(\Gamma/\tilde{\sigma}_{-S})$ where the equality follows from the induction hypothesis. Therefore $\tilde{\sigma} \notin SE(\Gamma)$. This contradiction establishes the proposition.□

# Bibliography


[1] Aumann, R. (1959): Acceptable points in general cooperative $n$-person games, in "*Contributions to the Theory of Games IV*," Princeton Univ. Press, Princeton, N.J., 1959.

[2] Aumann, R. (1974): "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics* **1**, 67-96.

[3] Bernheim, B., B. Peleg, and M. Whinston. (1987): "Coalition-proof Nash equilibria I: concepts," *Journal of Economic Theory* **42**, 1-12.

[4] Einy, E., and B. Peleg. (1992): "Coalition-proof communication equilibria," Hebrew University of Jerusalem discussion paper #9.

[5] Farrell, J. (1987): "Cheap talk, coordination, and entry," *Rand Journal of Economics* **18**: 34-39.

[6] Holmstrom, B. and R. Myerson. (1983): "Efficient and durable decision rules with incomplete information," *Econometrica* **53**, 1799-1819.

[7] Luce, R. and D. Raiffa. (1957): *Games and Decisions*, J. Wiley and Sons, New York.

[8] Moulin, H. (1981): *Game Theory for the Social Sciences*, Studies in Game Theory and Mathematical Economics, N.Y. University Press, N.Y.

[9] Myerson, R. (1991): *Game Theory: Analysis of Conflict,* Harvard Univ. Press, Cambridge, Massachusetts.

[10] Ray, I. (1993): "Coalition-proof correlated equilibrium: a definition," CORE Discussion paper #9353.