Rational Expectations and Farsighted Stability[*]

Bhaskar Dutta

University of Warwick and Ashoka University

Rajiv Vohra

Brown University

February 2016

Abstract. In the study of farsighted coalitional behavior, a central role is played by the von Neumann-Morgenstern (1944) stable set and its modification that incorporates farsightedness. Such a modification was first proposed by Harsanyi (1974) and has recently been re-formulated by Ray and Vohra (2015). The farsighted stable set is based on a notion of indirect dominance in which an outcome can be dominated by a chain of coalitional 'moves' in which each coalition that is involved in the sequence *eventually* stands to gain. However, it does not require that each coalition make a *maximal* move, i.e., one that is not Pareto dominated (for the members of the coalition in question) by another. Consequently, when there are multiple continuation paths the farsighted stable set can yield unreasonable predictions. We restrict coalitions to hold *common* rational expectations that incorporate maximality regarding the continuation path from every state. This leads to two related solution concepts: the rational expectations farsighted stable set (REFS) and the strong rational expectations farsighted stable set (SREFS). We apply these concepts to simple games and to pillage games to illustrate the consequences of imposing rational expectations for farsighted stability.

KEYWORDS: stable sets, farsightedness, consistency, maximality, rational expectations, simple games, pillage games.

JEL CLASSIFICATION: C71, D72, D74

[*] We are extremely grateful to Debraj Ray for many fruitful discussions on this subject. Thanks are also due to Jean Jacques Herings and Mert Kimya.

1. Introduction

Theories of coalitional stability are based on the notion of domination or objections by coalitions. A coalition is said to have an objection to the status-quo if it can change the outcome to one in which all its members gain. Perhaps the most widely used solution concept in this literature is the *core*, which is defined as the set of outcomes to which there is no objection. The original formulation of the theory by von Neumann and Morgenstern (1944), however, was concerned with a somewhat more sophisticated equilibrium concept, which they referred to simply as the "solution" but which has since become known as the *vNM stable set*.

A *vNM stable set* consists of outcomes that satisfy two properties: (1) internal stability in the sense that no stable outcome dominates any other stable outcome; (2) external stability in the sense that every outcome not in the stable set is dominated by some stable outcome. It is easy to see that every stable set contains the core. However, as shown by Lucas (1968), it is possible that no stable set exists even if the core is nonempty. There is a large literature on the stable set even though it has been notoriously difficult to work with.[1]

Both the core and the stable set are based on myopic, or one-shot, deviations by coalitions. If a change made by a coalition can be followed by other coalitional moves, then clearly we should require coalitions to be farsighted in their behavior. This is a direction of research that has attracted renewed interest; see, for example, Harsanyi (1974), Aumann and Myerson (1988), Chwe (1994), Bloch (1996), Ray and Vohra (1997, 1999), Xue (1978), Diamantoudi and Xue (2003), Konishi and Ray (2003), Herings et al. (2004, 2009), Ray (2011), Mauleon et al. (2011), Ray and Vohra (2014, 2015), Chander (2015), Kimya (2015). It is by no means obvious how the classical theory should be modified to account for farsightedness, which partly explains the diversity of approaches taken in this literature. Ray and Vohra (2014) distinguish between two principal approaches: (a) the *blocking approach*, which follows traditional cooperative game theory in abstracting away from the details of the negotiation process and relying largely on the specification of what each coalition is able to accomplish on its own, and (b) the *bargaining approach*, which is based on noncooperative coalition bargaining.[2]

This paper is in the tradition of the blocking approach, where farsightedness implies that coalitional decision making is based not on the immediate effect of an initial 'move' but the 'final outcome'.[3] This immediately raises the question of what the 'final outcome' is in a sequence of coalitional moves. In particular, if there is no pre-specified set of 'terminal' states, how do we know that the last step in a sequence of coalition moves is indeed the 'final outcome'? Suppose coalition S^1 replaces x with x^1, and then S^2 replaces x^1 with x^2. If x^2 is the final outcome, farsightedness would require S^1 to compare the utility of x^2 to that of x (and ignore its payoff at x^1). But this argument only works *if* x^2 is known to be the 'final outcome'. What is considered to be a final outcome must, of course, also be stable. Thus, testing the stability of a particular

[1] See Lucas (1992) for a survey.

[2] They also show that both approaches can be included within a more general dynamic model in which payoffs accrue in real-time and and there is an explicit protocol regarding which coalition has the right to move, depending on the history.

[3] In a real-time model what matters is the entire stream of (discounted) payoffs along a sequence of moves.

outcome against a sequence of moves requires us to know which of the other outcomes are stable. This is precisely the kind of circularity that the stable set is very adept at handling, making it a fruitful vehicle for incorporating farsightedness.

The idea of modifying the stable set by allowing for sequences of coalitional moves, with each coalition focused on the 'final outcome', goes back to Harsanyi (1974).[4] One conceptual difficulty with the farsighted stable set, including it's more recent reformulation by Ray and Vohra (2015), is that coalitions involved in a farsighted objection are not required to make the most profitable moves that may be available to them. This is the issue of *maximality*, which we explain in more detail in Sections 2 and 3, along with the issue of *consistency* or *history dependence*. The problem of consistency relates to the possibility that farsighted objections may involve different coalitions holding different expectations about the continuation path from some outcome.

The main aim of this paper is to resolve the maximality and consistency issues while maintaining the parsimony of the blocking approach. We do so by explicitly introducing expectations, held commonly by all agents, regarding the sequence of coalitional moves, if any, from every outcome. This leads us to define two related solution concepts: the rational expectations farsighted stable set (REFS) and the strong rational expectations farsighted stable set (SREFS). We show that although there are a some cases in which farsighted stable sets, or even vNM stable sets, are REFS or SREFS, in general imposing rational expectations can be consequential for farsighted stability.

In Sections 4 and 5, respectively, we apply these concepts to two important economic models: simple games and pillage games. The former have been fruitful in studying voting behavior and possess a rich literature on stable sets. We use simple games to highlight the consistency issue. Pillage games are models of economies where property rights do not exist so that the more "powerful" can capture the assets or wealth of the less powerful. These games cannot be represented as characteristic function games. Jordan (2006) and Acemoglu *et al.* (2008) study farsighted cooperative behavior in these models.[5] In these models we find that the maximality issue makes a crucial difference. Together, these applications illustrate how the imposition of rational expectations can result in predictions that are very different from those of the farsighted stable set.

2. Maximality and Consistency

We consider a general setting, described by an *abstract game*, $(N, X, E, u_i(.))$, where N is the set of players and X is the set of outcomes or states. Let $\mathcal{N}$ denote the set of all subsets of N. The effectivity correspondence, $E : X \times X \mapsto \mathcal{N}$, specifies the coalitions that have the ability to replace a state with another state: for $x, y \in X$, $E(x, y)$ is the (possibly empty) set of coalitions that can replace x with y. Finally, $u_i(x)$ is the utility of player i at state x.

The set of outcomes as well as the effectivity correspondence will depend on the structure of the model being studied. For instance, in a *characteristic function game*, (N, V), there is a set

[4] See Chwe (1994) for a formal definition of the Harsanyi farsighted stable set.

[5] Piccione and Rubinstein (2007) study the analogue of an exchange economy in the "jungle", which is similar to a world without property rights.

of feasible utilities, $V(S)$, for every coalition S.[6] In this case, a state will generally refer to a coalition structure and a corresponding payoff allocation which is feasible and efficient for each of the coalitions in the coalition structure. Historically, however, following von Neumann and Morgenstern (1944), much of the literature has treated the set of states to be the set of *imputations*, the Pareto efficient utility profiles in $V(N)$, and implicitly assumed that $S \in E(x, y)$ iff $y_S \in V(S)$. As explained in Ray and Vohra (2015), and as will become clear below, this turns out to be unsatisfactory for studying farsightedness.

State y *dominates* x if there is $S \in E(x, y)$ such that $u_S(y) \gg u_S(x)$. In this case we also say that (S, y) is an *objection* to x.

The *core* is the set of all states to which there is no objection.

A set $K \subseteq X$ is a *vNM stable set* if it satisfies:

(i) there do not exist $x, y \in K$ such that y dominates x; internal stability.
(ii) for every $x \notin K$, there exists $y \in K$ such that y dominates x; external stability.

For an abstract game, we define farsighted dominance as follows:

State y *farsightedly dominates* x (under E) if there is a sequence $y^0, (y^1, S^1), \ldots, (y^m, S^m)$, with $y^0 = x$ and $y^m = y$, such that for all $k = 1, \ldots m$:

$$S^k \in E(y^{k-1}, y^k)$$

and

$$u(y)_{S^k} \gg u(y^{k-1})_{S^k}.$$

For set $F \subseteq X$ is a *farsighted stable set* if it satisfies:

(i) there do not exist $x, y \in F$ such that y farsightedly dominates x; farsighted internal stability.
(ii) for every $x \notin F$, there exists $y \in F$ such that y farsightedly dominates x; farsighted external stability.

It is important to emphasize that the notion of effectivity is especially delicate in the context of farsightedness. Harsanyi (1974), in defining farsighted dominance for a characteristic function game, maintained the von Neumann-Morgenstern assumption that $S \in E(x, y)$ iff $y_S \in V(S)$.[7] This way of specifying effectivity gives coalition S complete freedom in choosing y_{-S}, the payoffs to outsiders (provided y is an imputation and $y_S \in V(S)$). This is not important for myopic solutions such as the core and the stable set. But in the case of farsighted dominance this is not only conceptually questionable but can significantly alter the nature of the farsighted stable set, as shown by Ray and Vohra (2015). They demonstrate that imposing reasonable restrictions on the effectivity correspondence results in a farsighted stable that is very different from, and

[6] A transferable utility, or TU, characteristic function game will be denoted (N, v), where $V(S) = \{u \in R^S \mid \sum_i u_i \leq v(S)\}$ for all S.

[7] In fact, Harsanyi was following the standard practice of making this part of the dominance condition rather than presenting it through an effectivity correspondence. So it would be more precise to say that this is implicitly what Harsanyi assumed.

arguably more plausible than, that of Harsanyi (1974). We will therefore need to be attentive to this issue when we consider specific models in Sections 4 and 5. Until then, in order to highlight the main concerns of this paper, we shall work in the generality of an abstract game, without any explicit restrictions on the effectivity correspondence.

The farsighted stable set is based on an optimistic view of the coalitions involved in a farsighted objection. A state is dominated if there exists *some* path that leads to a better outcome. Chwe (1994) proposed a farsighted solution concept based on conservative behavior, which is good at identifying states that cannot possibly be considered stable. A set $K \subseteq X$ is *consistent* if

$$K = \{x \in X \mid \text{ for all } y \text{ and } S \text{ with } S \in E(x,y), \text{ there exists } z \in K \text{ such that } z = y \text{ or } z \text{ farsightedly dominates } y \text{ and } u_i(z) \leq u_i(x) \text{ for some } i \in S\}.$$

Thus, any potential move from a point in a consistent set is deterred by *some* farsighted objection that ends in the set. Chwe shows that there exists one such set which contains all other consistent sets, and defines this to be the *largest consistent set* (LCS).

In general, both of these solution concepts are unsatisfactory because optimistic or pessimistic expectations are both ad hoc. Ideally, a solution concept should be based on *optimal* behavior (which may of course turn out to be optimistic or pessimistic in particular examples). The following examples, based on similar ones in Xue (1998), Herings et al. (2004) and Ray and Vohra (2014), illustrate this problem vividly.

EXAMPLE **1.** *The game is depicted in Figure 1, Player 1 is effective in moving from state a to b, while player 2 can replace state b with either c or d, which are both 'terminal' states. The numbers below each state denote the utilities to the players.*

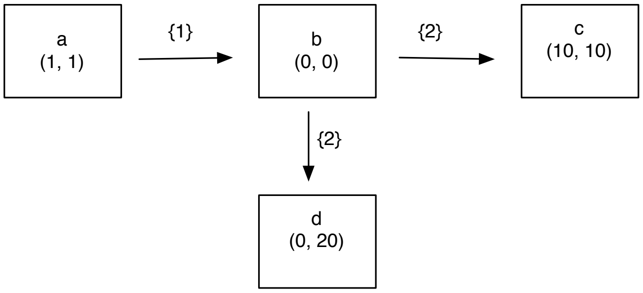

FIGURE 1

Both c and d belong to the farsighted stable set. Since there is a farsighted objection from a to c, the former is not in the farsighted stable set. However, this is based on the expectation that player 2 will choose c instead of d even though 2 prefers d to c. If 2 is expected to move, rationally, to d, the a should be judged to be stable, contrary to the prediction of the farsighted stable set. Note that a belongs to the LCS because of the the possibility that the final outcome is d, so in this example the LCS makes a more reasonable prediction than the farsighted stable set.

EXAMPLE **2.** *This is a modification of Example 1 as shown in Figure 2.*

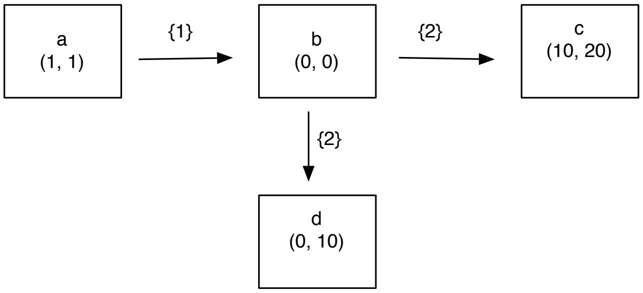

FIGURE 2

Now the optimal move for player 2 is to choose c rather than d. The LCS and farsighted stable set remain unchanged. But now it is the LCS which provides the wrong answer because player 1 should not fear that player 2 will (irrationally) choose d instead of c. In this example, the farsighted stable set makes a more reasonable prediction.

As the previous two examples show, both the LCS and the farsighted stable set suffer from the problem that they do not require coalitions (in these examples, player 2) to make moves that are *maximal* among all profitable moves. (A formal definition of maximality in our framework appears in the next Section).

Another problem that afflicts both the LCS and the farsighted stable set is that they may be based on expectations that are inconsistent in the sense that coalitions move based on different expectations about the continuation to follow. For the LCS this was pointed out by Konishi and Ray (2003). Our next Example illustrates this problem for both the farsighted stable set and the LCS.

EXAMPLE **3.** *This is a three-player game with five states, shown in Figure 3.*

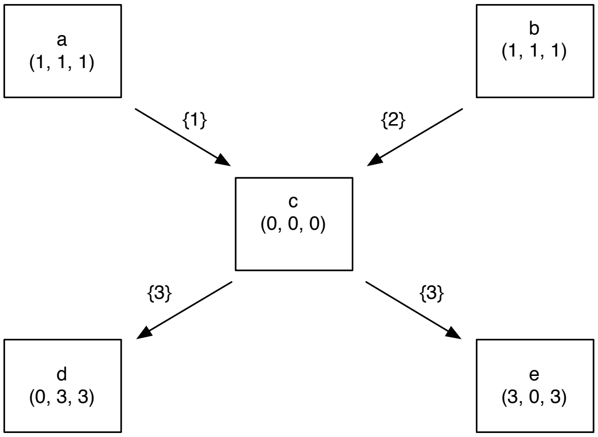

FIGURE 3

In this example, the farsighted stable set is $\{d, e\}$ while the LCS is $\{a, b, d, e\}$. State a is not in the farsighted stable set because of a farsighted objection to e, and b is not in it because of a farsighted objection to d. However, from c player 3 can only move to either d or e. Thus, the exclusion of both a and b from the farsighted stable set is based on inconsistent expectations of the move from c. On the other hand, the inclusion of both a and b in the LCS is also based on inconsistent expectations. The 'right' answer in this example should be that $\{a, d, e\}$ and $\{b, d, e\}$ are two 'stable sets': the former if the expectation is that player 3 will move from c to d and the latter if the expectation is that 3 will move from c to e. Another interpretation of this phenomenon is that expectations under the farsighted stable set or the LCS allow for history dependence: the continuation from c depends on the history. In this sense what we are asking for can be seen as the joint requirement of consistency and history independence. Formally, in a dynamic model this corresponds to the Markovian assumption on processes, which is commonly made in the literature, as in Konishi and Ray (2003).

To define optimal behavior one will need to rely on players having (rational) expectations about the continuation path following any coalition move. In a dynamic setting, such as in Konishi and Ray (2003) or Ray and Vohra (2014), these expectations are specified by a dynamic process of coalitional moves. An *equilibrium process of coalition formation* (EPCF) is a Markovian process in which coalitions take actions that are *maximally profitable* in terms of a value function.[8] One difference in these models is that Ray and Vohra (2014), unlike Konishi and Ray (2003), specify a protocol to explicitly determine the order in which coalitions are called upon to move at each stage. In spirit, though, in both cases the approach for incorporating consistency and rational expectations is similar to ours, even though we seek to accomplish this more directly within the static, blocking approach. Static models most closely related to our approach are Xue (1978) and Kimya (2015).

Xue (1978) argued that to resolve the maximality issue we should consider a stable set defined over *paths* of coalition actions rather than on outcomes. In many cases, such as Example 1 and 2, this can resolve the problem. However, it may push the choice between optimism and pessimism to another level. When a path is tested against a deviation by a coalition, the deviation can itself lead to multiple stable paths and so in evaluating these multiple paths the pessimism/optimism choice resurfaces. This leads him to define the *optimistic stable standard of behavior* and the *conservative stable standard of behavior*. In Example 3, the predictions of these two concepts match the farsighted stable set and the LCS, respectively. We are able to avoid this by considering stability in terms of a given expectation that describes transitions from *every* outcome. In our framework, once a coalition makes a change, there is no further ambiguity about the continuation path. In this respect, our approach is similar to Konishi and Ray (2003), Ray and Vohra (2014) and Kimya (2015), even though these papers propose solution concepts that are not defined in terms of stable sets. In these papers an 'equilibrium path' need not involve all coalitions doing *strictly* better, whereas in our framework a sequence of coalition moves will be a farsighted objection, involving strict improvements; see Kimya (2015) for further discussion.

[8] In Example 3, therefore, the prediction would be that the stable outcomes are either $\{a, d, e\}$ or $\{b, d, e\}$.

At some intuitive level, notions of farsightedness and maximality attempt to bring into coalition games considerations that are similar to backward induction in noncooperative games. The difficulty, of course, is that coalitions games don't typically have the structure of an extensive form that allows for recursion.[9] This connection comes out perhaps most clearly in Kimya's (2015) concept of *equilibrium coalitional behavior* (ECB), which is defined for a model that has the advantage of being directly applicable to extensive form games. See Kimya (2015) for more on the relationship between ECB and our solution concepts.

We should acknowledge that one reason all the examples in this Section are so simple is because they concern abstract games. The skeptical reader might wonder whether issues of maximality or consistency matter in more specific models, e.g., characteristic function games or economic models without externalities. In Sections 4 and 5 it will become clear that the issues highlighted here are indeed of more general importance.

3. Farsightedness with Rational Expectations

Jordan (2006) formulates the idea that farsighted stability can be expressed in terms of commonly held consistent expectations regarding the 'final outcome' from any state.[10] He defines an expectation as a function $\phi : X \to X$ such that for every $x \in X$, $\phi(\phi(x)) = \phi(x)$. A stationary state of ϕ is x such that $\phi(x) = x$. Given a farsighted stable set, Z, it is straightforward to construct an expectation ϕ that is consistent with farsighted dominance and yields Z as the collection of all stationary outcomes. If $x \in Z$, let $\phi(x) = x$. If $x \notin Z$, let $\phi(x) = y$ for some $y \in Z$ that farsightedly dominates x.[11]

In order to deal with the issues discussed in Section 2, we extend Jordan's approach by interpreting an expectation to describe the transition from one state to another, not necessarily the final outcome from a state. In addition, we will also find it important to keep track of the coalition that is expected to make the transition. With this in mind, we define an expectation as a function $F : X \to X \times \mathcal{N}$. For a state $x \in X$, denote $F(x) = (f(x), S(x))$, where $f(x)$ is the state that is expected to follow x and $S(x) \in E(x, f(x))$ is the coalition expected to implement this change. If $f(x) = x$, $S(x) = \emptyset$, signifying the fact that no coalition is expected to change x. A stationary point of F is a state x such that $f(x) = x$. Given an expectation $F(.) = (f(.), S(.))$, let f^k denote the k-fold composition of f. In particular, $f^2(x) = f(f(x))$. With a slight abuse of notation, let $F^k(x) = F(f^{k-1}(x))$.

An expectation is said to be *absorbing* if for every $x \in X$ there exists k such that $f^k(x)$ is stationary. In this case, let $f^*(x) = f^k(x)$ where $f^k(x)$ is stationary.

[9] *Coalition proof Nash equilibrium* in Bernheim, Peleg and Winston (1987) and *equilibrium binding agreements* in Ray and Vohra (1997) are able to make use of recursion by restricting attention to chains of objections in which each coalition is a subset of the previous one.

[10] We discuss Jordan's model of pillage games in Section 4.

[11] This bears some similarity to Harsanyi's (1974) attempt to relate the stationary set of an equilibrium in a noncooperative game to a version of a (farsighted) stable set.

We seek to describe a set of stable outcomes $Z \subseteq X$ that is 'justified' by an expectation in the sense that Z is the set of stationary points of an expectation F that embodies farsighted rationality.

An absorbing expectation F is said to be a *rational expectation* if it has the following properties:

(I) If x is stationary, then from x no coalition is effective in making a profitable move (consistent with F), i.e., there does not exist $T \in E(x,y)$ such that $u_T(f^*(y)) \gg u_T(x)$.

(E) If x is a nonstationary state, then $F(x)$ must prescribe a path that is profitable for all the coalitions that are expected to implement it, i.e., $(x, F(x), F^2(x), \ldots F^k(x))$ is a farsighted objection where $f^k(x) = f^*(x)$.

(M) If x is a nonstationary state, then $F(x)$ must prescribe an optimally profitable path for coalition $S(x)$ in the sense that there does not exist y such that $S(x) \in E(x,y)$ and $u_{S(x)}(f^*(y)) \gg u_{S(x)}(f^*(x))$.

The set of stationary points, $\Sigma(F)$, of a rational expectation F is said to be a *rational expectations farsighted stable set* (REFS).

Conditions (I) and (E) are related but not the same as farsighted internal and external stability (conditions (i) and (ii) in the definition of a farsighted stable set). And the differences can be significant enough to generate very different results, as we will see. Since $\Sigma(F)$ is a set of stationary states, Condition (I) clearly implies that $\Sigma(F)$ satisfies *myopic* internal stability in the traditional sense. It is weaker than farsighted internal stability since it requires internal stability only with respect to those farsighted objections that are *consistent* with the common expectation F.

Condition (E) is stronger than externality stability of Ray and Vohra (2015) because it states that to every $x \notin \Sigma(F)$ there is a farsighted objection (terminating in $\Sigma(F)$) consistent with the common expectation F.

Condition (M) is the *maximality* condition; it is a translation of the corresponding condition of Konishi and Ray (2004) and Ray and Vohra (2014) into our framework.[12] Maximality is the proper expression of optimality if one takes the view that at a nonstationary state x, $S(x)$ is the coalition that has the floor, which gives it sole priority in selecting the transition from x. However, one could entertain models in which, under certain conditions, some other coalition may also have the right to intervene and change course. This motivates the following notion of *strong maximality*:

(M') If x is a nonstationary state, then $F(x)$ must prescribe an optimally profitable path in the sense that no coalition has the power to change course and gain, i.e., there does not exist $T \in E(x,y)$ such that $T \cap S(x) \neq \emptyset$ and $u_T(f^*(y)) \gg u_T(f^*(x))$.

Condition (M') continues to assume that a coalition disjoint from $S(x)$ cannot interfere in the expected move. However, a coalition T which includes some players from $S(x)$ is allowed to change course. This is based on the idea that a move by $S(x)$ requires the unanimous consent

[12] It would require that in Example 1 $f(b) = d$, and in Example 2 $f(b) = c$.

of all its members, which means that another coalition may take the initiative if it can enlist the support of at least one player in $S(x)$.

A expectation F satisfying (I), (E) and (M') is a *strong rational expectation.* The set of stationary points of a strong rational expectation F is said to be a *strong rational expectations farsighted stable set* (SREFS).

Every SREFS is clearly a REFS. We shall therefore attempt to show the existence of a SREFS whenever possible. But, as our next example shows, this is not always possible; condition (M') of SREFS may be too demanding for existence, even though a REFS may exist.

EXAMPLE **4.**

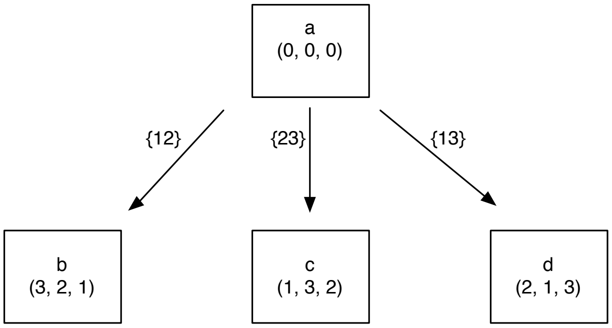

FIGURE 4

There are three REFS in Example 4: $\{b\}$, $\{c\}$ and $\{d\}$. However, none of these can be a SREFS because some player in the coalition that moves from a could do even better by moving with a different player.[13]

In general, even the existence of a REFS is not guaranteed. One example in which is the case is the three-player NTU 'roommate game' which also does not possess a farsighted stable set; see Ray and Vohra (2015).

In general, REFS or SREFS can be different from farsighted stable sets. While this will be the theme of Sections 4 and 5, the following Example illustrates this point.

EXAMPLE **5.** *(A three-player, TU game, (N, v) with one veto player): $N = \{1, 2, 3\}$, $v(\{1, 2\}) = v(\{1, 3\}) = v(N) = 1$ and $v(S) = 0$ for all other S.*

Ray and Vohra (2015) show that, under some mild assumptions on the effectivity correspondence (see conditions (a) and (b) in Section 4), every farsighted stable in this game assigns a fixed

[13] An even stronger notion of maximality, which we will not pursue, is one adopted by Xue (1998). It allows the expected path to be altered by *any* coalition, even one that is disjoint from the coalition that is expected to move. For instance, modify Example 4 so that player 1 is effective in moving from a to b, player 2 from a to c and player 3 from a to d. Now, REFS and SREFS are the same: they consist of b, c and d. But Xue's maximality condition would allow any player to change course when someone else has made a move, which clearly results in a failure of existence.

payoff to the veto player, strictly between 0 and 1, while the remaining surplus is can be divided in any way among players 2 and 3. More precisely, for every $a \in (0, 1)$, there is a farsighted stable set Z_a with the set of payoffs: $\{u \in R^3_+ \mid u_1 = a, u_2 + u_3 = 1 - a\}$; see Figure 5, where the vertices of the simplex denote states at which the entire surplus is allocation to one of the three players.

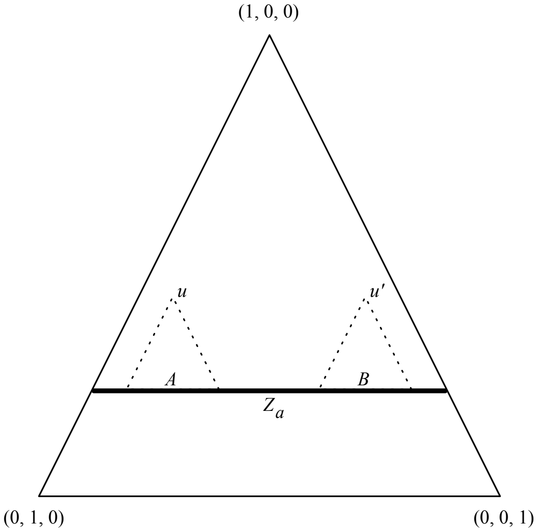

FIGURE 5. A FARSIGHTED STABLE SET, Z_a, IN EXAMPLE 5.

However, no set of the form Z_a can be a REFS because the external stability of Z_a (in the sense of a farsighted stable set) relies on inconsistent expectations. To see this, consider why the allocation u is not in Z_a. There is first an objection by $\{2, 3\}$, resulting in the coalition structure $\{\{1\}, \{2, 3\}\}$ and 0 payoff to all players. Call this state x^0. This is followed by a move by N to a point in A. And u' is not stable because there is first a move by $\{2, 3\}$ to x^0, followed by a move by N to a point in B. In the first case, x^0 is expected to be replaced by a point in A, while in the second case it is expected to be replaced by a point in B. This is precisely the kind of 'inconsistent' expectation that must be ruled out in a REFS or SREFS. In other words, a farsighted stable set in this example cannot be a REFS.

We see in the next Section, where we provide a general result on the existence of a SREFS in a simple game, that SREFS do exist in Example 5, but they are very different from vNM stable sets or farsighted stable sets.

There is one interesting case in which an SREFS (or REFS) coincides with a farsighted stable set.

A set of states Z is a *single-payoff* set if $u(x) = u(y)$ for all $x, y \in Z$.

THEOREM **1.** *If Z is a single-payoff REFS it is a SREFS and a farsighted stable set. Conversely, if Z is a single-payoff farsighted stable, then it is a SREFS.*

It follows from Theorem 2 of Ray and Vohra (2015) that every characteristic function game with a separable allocation[14] possesses a SREFS.

Proof. Suppose Z is a single-payoff REFS. Since all stationary states have the same payoff, Z is clearly also a SREFS and satisfies farsighted internal stability (condition (i) in the definition of a farsighted stable set). As already observed, Condition (E) implies farsighted external stability (condition (ii)). Thus Z is a farsighted stable set.

To prove the second part of the Theorem, consider a single-payoff farsighted stable set X^0. Define X^1 to be the set of states from which there is a farsighted objection to some state in X^0 in a single step. More precisely:

$$X^1 = \{x \in X - X^0 \mid \exists x^0 \in X^0, S \in E(x, x^0) \text{ with } u_S(x^0) \gg u_S(x)\}.$$

Since X^0 is a farsighted stable set, from every $x \in X - X^0$, there is a farsighted objection leading to some $x^0 \in X^0$: $x, (x^1, S^1), \ldots, (x', S'), (x^m, S^m)$. Clearly $x' \in X^1$, which establishes the nonemptiness of X^1. We shall now recursively define subsets of X from which there are farsighted objections leading to X^0 in a minimal number of steps. All of these sets will be disjoint and will cover X. The construction is as follows.

Suppose X^j have been defined for all $j = 1, \ldots, k$. Define X^{k+1} to be the set of all other states from which there is a farsighted objection leading to X^0 such that the first step is a state in X^k:

$$\begin{aligned} X^{k+1} &= \{x \in X - \cup_{j=0}^{k} X^j \mid \text{there is a farsighted objection } x, (x^1, S^1), \ldots, (x^m, S^m), \\ &\quad \text{with } x^1 \in X^k \text{ and } x^m \in X^0\}. \end{aligned}$$

Note that if $X^{k+1} = \emptyset$, then $\cup_{j=0}^{k} X^k = X$. To complete the proof we will construct a function $F : X \to X \times \mathcal{N}$ where $F(x) = (f(x), S(x))$ such that $f(x^0) = x^0$ for every $x^0 \in X^0$, and for every $x \in X^{k+1}$, $f(x) \in X^k$. We know that from $x \in X^{k+1}$ there is a farsighted objection leading to some state in X^0 which proceeds by first moving to a point in X^k. We will choose $f(x)$ as one such point along with a unique coalition that initiates such a farsighted objection. The function F is constructed recursively. For $x \in X^1$, define $f(x) = x^0 \in X^0$ and $S(x)$ to be a coalition that has a one step objection from x to x^0. If there are multiple such coalitions, choose one arbitrarily. This describes a unique transition from X^1 to X^0. Having defined $F : X^j \mapsto X^{j-1} \times \mathcal{N}$ for all $j = 1, \ldots k$, if $X^{k+1} \neq \emptyset$, for $x \in X^{k+1}$ let S^1 be a coalition that has a farsighted objection from x to $x^m \in X^0$, denoted $x, (x^1, S^1), \ldots, (x^m, S^m)$, such that $x^1 \in X^k$. Let $F(x) = (x^1, S^1)$. Note that there may be multiple such farsighted objections. In that case, pick $F(x)$ to be the first element of any such sequence. Proceeding in this way, we have constructed a function $F : X \to X \times \mathcal{N}$ with X^0 as its set of stationary points. It remains to be shown that F is a strong rational expectation.

Since the stationary points of F have the same payoff vector, it trivially satisfies Condition (I) in the definition of a rational expectation.

To prove Condition (E), consider $x \in X - X^0$. Of course, there is some $k \geq 0$ such that $x \in X^{k+1}$. From the construction of F we know that there exists a farsighted objection from x to $x^m \in X^0$, say, $x, (x^1, S^1), (x^2, S^2), \ldots, (x^m, S^m)$, such that $F(x) = (x^1, S^1)$.

[14] A sufficient condition for an allocation to be separable is that it belongs to the interior of the core.

(There is no presumption that $(x^2, S^2) = F^2(x)$, or that $m = k$). Let u^0 denote the (common) payoff corresponding to each of the (single-payoff) states in X^0. Obviously, $u_{S^1}(x^m) = u^0_{S^1} \gg u_{S^1}(x)$. Since $f^k(x) \in X^0$, $u(f^k(x)) = u^0$, which implies that S^1 gains in moving along the path $(x, F(x), F^2(x), \ldots, F^k(x))$. By the same reasoning, S^2 also gains by moving from $f(x)$ along the path $(F^2(x), \ldots, F^k(x))$, and so on for all S^j, $j = 1, \ldots S^k$. Thus, $(x, F(x), F^2(x), \ldots, F^k(x))$ is a farsighted objection from x to $f^k(x) \in X^0$.

To see that Condition (M') is satisfied note that no player, and therefore no coalition, can gain by deviating from the path prescribed by F because any deviation leads to the same payoff vector, u^0. This establishes Condition (M') and completes the proof that F is a strong rational expectation with $\Sigma(F) = Z$. ■

In general, REFS or SREFS can be different from farsighted stable sets, as we have seen in Example 5, and as we will see in more generality in the following Sections.

4. Simple Games

In this Section we study superadditive TU games that have the property that for every coalition S, either $v(S) = 1$, in which case S is said to be a *winning* coalition, or $v(S) = 0$. Moreover, if $v(S) = 1$, then $v(N - S) = 0$. This is the class of proper, simple games (see von Neumann and Morgenstern (1944)). The set of efficient payoff allocations, or imputations, in any such game is the nonnegative n-dimensional unit simplex, $\triangle$. For simplicity, assume that no $i \in N$ is a *dummy player*; that is each i belongs to at least one *minimal* winning coalition.[15]

Let $\mathcal{W}$ denote the set of all winning coalitions. The collection of all veto players, also known as the *collegium*, is denoted $C = \cap_{S \in \mathcal{W}} S$. A *collegial game* is one in which $C \neq \emptyset$. The collegium (and the corresponding game) will be called *oligarchic* if C is itself a winning coalition. Note that in the absence of dummy players an oligarchic game is one in which $C = N$; it is a pure bargaining game, in which the grand coalition is the only winning coalition.

In a simple game, a state x specifies a coalition structure, denoted $\pi(x)$, and an associated payoff, $u(x)$, such that $\sum_{i \in W(x)} u_i(x) = 1$, where $W(x)$ is the winning coalition (if any) in $\pi(x)$. We use X^0 to denote the set of states where no winning coalition forms and so $u_i = 0$ for all i. States in X^0 are called *zero states*.[16]

In this section, for reasons similar to those explained in Ray and Vohra (2015), we make the following assumption regarding the effectivity correspondence.

Assumption 1. *The effectivity correspondence satisfies the following restrictions:*

[15] The proof of our main result in this Section can be easily modified to accommodate the presence of dummy players.

[16] Ray and Vohra (2015) find it convenient to identify a state by the winning coalition, if any, and the payoff allocation. However, in our model a state specifies the entire coalition structure, not just the winning coalition. By allowing for a richer notion of states this allows us some added flexibility in constructing an expectation function.

(a) *For every $x \in X$, $S \subseteq N$ and $u \in R^S_+$ with $\sum_{i\in S} u_i = v(S)$, there is $y \in X$ such that $S \in E(x, y)$ and $u(y)_S = u$.*
(b) *If $S \in E(x, y)$, then $S \in \pi(y)$ and $T - S \in \pi(y)$ for every $T \in \pi(x)$.*
(c) *For all $x, y \in X$, and $T \subset N$, if $(W(x) - T) \in \mathcal{W}$, then $T \in E(x, y)$ only if $u_i(y) \geq u_i(x)$ for all $i \in W(x) - T$.*

Condition (a) states that every coalition can form and divide its worth in any way among its members. Condition (b) states that when a coalition S forms it does not affect any coalition that is disjoint from it, and if it includes some members of a coalition, then the residual remains intact. This is a natural way to describe the immediate change in the coalition structure resulting from the formation of a coalition. Condition (c) requires that if with the formation of T the residual in $W(x)$ remains winning, then the players in $W(x) - T$ cannot lose.[17] It includes the condition that $W(x) \cap T = \emptyset$, then $u(x) = u(y)$. Conditions (b) and (c) should be interpreted as natural restrictions that prevent a coalition from reorganizing the payoffs or coalitional structure of those outside it.

In an oligarchic game, all states with a strictly positive payoff to all (veto) players is a separable payoff allocation. By Ray and Vohra (2015, Theorem 2), any such payoff allocation, with the coalition N, is a singleton farsighted stable set.[18] By Theorem 1, it is also a SREFS. In the remainder of this Section, therefore, we shall concentrate on non-oligarchic games.

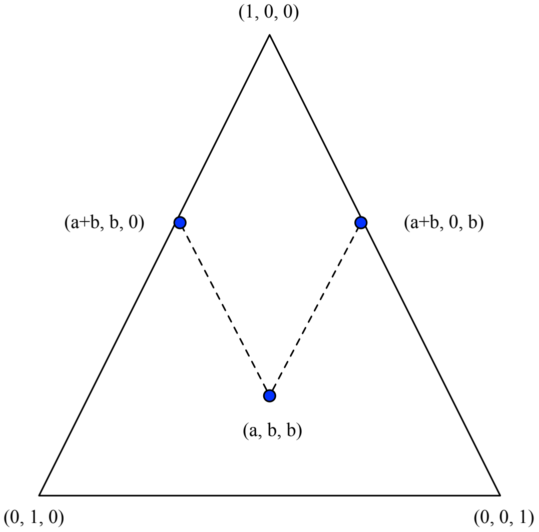

FIGURE 6. SREFS IN EXAMPLE 5.

Ray and Vohra (2015) show that in non-oligarchic games farsighted stable sets have the feature that veto players receive a fixed amount. In Example 5 this results in the payoffs being along

[17] Of course, this also implies that $W(x) - T \in \pi(y)$. Condition (b) goes beyond this because it also applies to residuals that are not winning.

[18] In fact, they show that every farsighted stable set in such a game is of this form.

a horizontal line in Figure 5 or 6. As we argued in Section 3, this relies on non-stationary or inconsistent expectations. It is therefore not surprising that SREFS and REFS can be very different from farsighted stable sets, which can themselves be very different from vNM stable sets. While this will become clear from Theorem 2, it is instructive to illustrate this comparison through Example 5. We will now show that in this Example there is a SREFS, Z, consisting of a finite set of states: $Z = \{(u^1, \pi^1), (u^2, \pi^2), (u^3, \pi^3)\}$, where for some $a \in (0, 1)$ and $b \equiv \frac{(1-a)}{2}$,

- $u^1 = (a, b, b), \pi^1 = \{N\}$.
- $u^2 = (a + b, b, 0), \pi^2 = \{\{1, 2\}, \{3\}\}$.
- $u^3 = (a + b, 0, b), \pi^3 = \{\{1, 3\}, \{2\}\}$.

The three imputations corresponding to Z are shown in Figure 6. To sustain Z as SREFS, we construct an expectation described by the following rules. In what follows, we write $x = (u, \pi)$ and $x^i = (u^i, \pi^i)$, $i = 1, 2, 3$.

(i) For each $x \in Z$, $f(x) = x$.
(ii) If x is such that $u = (0, 0, 0)$, then $S(x) = N$ and $f(x) = x^1$.
(iii) For each $x \notin Z$ such that $u_1 \geq a + b$, $S(x) = \{2, 3\}$ and $f(x) = ((0, 0, 0), \pi)$.
(iv) For each $x \notin Z$ such that $u_1 < a + b$ and $u_2 < b$, $S(x) = \{1, 2\}$ and $f(x) = x^2$.
(v) If x not covered by (i) to (iv) above and $u_1 < a + b$ and $u_3 < b$, $S(x) = \{1, 3\}$ and $f(x) = x^3$.
(vi) Finally, if x is not covered by (i) to (v) above and $u_1 < a$, then $S(x) = \{1\}$ and $f(x) = ((0, 0, 0), \pi)$.

This describes F for all $x \in X$.

Clearly, F satisfies (I) and (E). To check (M'), note that in all states in Z players 2 and 3 get either b or 0. This implies strong maximality in cases (ii) and (iii). The deviations in (iv) and (v) are strongly maximal since 1 gets $a + b$, her highest possible payoff in Z. In case (vi), (M') is satisfied because players 2 and 3 have no reason to move, which means that player 1 does not possess a farsighted objection that could end up at x^2 or x^3.

This completes the demonstration that Z is a SREFS.

In more general non-oligarchic games Ray and Vohra (2015) show that it is possible to construct a farsighted stable set in which veto players, and perhaps some others, receive a fixed payoff while the remainder of the surplus is shared in any arbitrary way among the remaining players. Such sets, known as discriminatory stable sets, also played an important role in vNM stable set theory, though with the rather important difference that in vNM stable sets it is *non-veto players* who received a fixed payoff. In contrast, SREFS do not seem to have the structure of discriminatory stable sets. Instead, in most cases SREFS yield finite payoff sets. We have of course already observed this in Example 5, but this is also the more general conclusion that emerges from the proof of our next result.

Although the notion of a farsighted stable set does not impose maximality, in non-oligarchic simple games this property does seem to hold. In this model, therefore, the difference between farsighted stable sets and SREFS seems to stem from consistency and history independence.

For most simple games we have been able to constructively prove the existence of a SREFS. There is, however, one class of games for which existence has proven to be elusive. We therefore need to make the following assumption.

Assumption 2. *There does not exist a minimal winning coalition with one veto player and two non-veto players.*

Subject to this assumption we are able to construct a SREFS for all collegial games; see the Appendix for a proof of the following result.

THEOREM **2.** *A SREFS exists in every non-oligarchic collegial game satisfying Assumptions 1 and 2.*

Can Assumption 2 can be dispensed with or will this case yield an example in which SREFS does not exist? As of now this question remains open.

Of course, a large class of simple games do not have any veto player, the simplest example being the majority game in which any majority of players constitutes a winning coalition. von Neumann and Morgenstern (1944) identified a class of constant-sum games that have a vNM stable set known as a *main simple solution*. Suppose there is $a \in \Re_+^N$ such that $\sum_{i \in S} a_i = 1$ for every minimal winning coalition S. Define, for each minimal winning coalition S, u^S to be the imputation such that $u_i^S = a_i$ for all $i \in S$ and $u_i^S = 0$ otherwise. If the game is a constant-sum game, then the set of all such imputations is a vNM stable set, known as the main simple solution. For instance, the imputation $(0.5, 0.5, 0)$ and its permutations constitute a main simple solution in the three-person majority game. It can be shown that the set of states corresponding to a main simple solution is a SREFS.

Suppose U is a main simple solution with associated vector $a \in \Re_+^N$. Let $Z(U) = \{x \in X \mid u(x) \in U\}$. We claim that $Z(U)$ is a SREFS. Since U is a vNM stable set, for every $x \notin Z(U)$ there is $S \subseteq N$ and $y \in Z(U)$ such that $S \in E(x, y)$ and $u_S(y) \gg u_S(x)$. For every $x \notin Z(U)$ pick any (S, y) with this property and set $F(x) = (y, S)$. If there are several such (S, y) pick one arbitrarily. For every $x \in Z(U)$, let $f(x) = x$. Clearly, F is an expectation that satisfies (E). Suppose it does not satisfy (I). Then there is $x \in Z(U)$ and $T \in E(x, y)$ such that $u_T(f(y)) \gg u_T(x)$. Let S be the minimal winning coalition such that $u(x) = u^S$. Since $u_i(x) = a_i$ for all $i \in S$, no $i \in S$ can get a higher payoff at any other state in $Z(U)$, which implies that $S \cap T = \emptyset$. Of course, S must be contained in the winning coalition at x. By Assumption 1 (c), $u_S(y) = u_S(x) = a_S$, i.e., $y \in Z(U)$ and therefore $f(y) = y$. But then $u_T(f(y)) \gg u_T(x)$ means that $u_T(y) \gg u_T(x)$, with $y \in Z(U)$, which contradicts the myopic internal stability of U. Thus, F satisfies (I). It clearly satisfies strong maximality, (M'), because if any S gains by moving from x to $y \in Z(U)$, then $u_i(y) = a_i$ for all $i \in S$ and there is no other $y' \in Z(U)$ such that $u_i(y') > u_i(y)$ for any $i \in S$.

5. PILLAGE GAMES

In this Section we apply our solution concepts to pillage games and compare them to the vNM stable set, analyzed by Jordan (2006), as well as the farsighted stable set. In this setting, in contrast to simple games, we find that maximality and strong maximality, rather than consistency, play a crucial role in distinguishing between REFS, SREFS and farsighted stable sets.

In a pillage game, a coalition can appropriate the resources of any other coalition that has less power. Given a set of players $N = \{1, \ldots, n\}$, the set of wealth allocations is $\triangle$, the unit simplex in R^n. We shall consider the class of pillage games in which 'wealth is power': the power of coalition S is is simply its aggregate wealth, $w_S \equiv \sum_{i \in S} w_i$. Given wealth allocations w and w' let $L(w, w') = \{i \in N \mid w'_i < w_i\}$ denote the set of players who lose in moving from w to w'. We define the effectivity correspondence in this model as follows:[19]

(1) $\quad S \in E(w, w')$ if and only if $w_S > w_{L(w,w')}$ and $w_i = w'_i$ for all $i \notin S \cup L(w, w')$.

This expresses the notion that a coalition can pillage another only if its power is strictly greater than that of the victims. Moreover, only the winners' and losers' wealth payoffs can be affected through the act of pillaging. That is, if j is neither amongst those who have been pillaged nor part of the coalition that changes w to w', then $w_j = w'_j$. This last condition rules out a pillaging coalition sharing its spoils with others. While this condition is of no consequence for myopic notions of stability, it becomes important in the context of farsighted stability. As we have remarked earlier, Ray and Vohra's (2015) emphasized that a deviating coalition must not be permitted to affect the distribution of the payoff of outsiders. As we discuss in Example 6 below, a gift can turn out to be hazardous to the recipients – a Trojan horse. We shall therefore assume throughout this Section that the effectivity correspondence is defined by (1).

By way of background, it will be useful to begin with Jordan's analysis of the (myopic) vNM stable set.

A number $a \in [0, 1]$ is said to be *dyadic* if $a = 0$ or $a = 2^{-k}$ for some nonnegative integer k. For every positive integer k let $D_k = \{w \in \triangle \mid \; w_i$ is dyadic for every i and if $w_i > 0$, then $w_i \geq 2^{-k}\}$. The set of all dyadic allocations is $D = \cup_k D_k$. The set of all allocations in which one player captures the entire surplus, D_0, is the set of *tyrannical allocations*. Of course, all such allocations are in the core. It is easy to see that the only other allocations in the core are ones in which two players share the surplus equally. In other words, the core is D_1. Jordan (2006) provides the following characterization of the stable set, which remains

THEOREM **3.** *(Jordan) The unique stable set is D.*

Jordan (2006) illustrates the issue of farsightedness by considering the three-player example in this model, where D consists of the allocations $(1, 0, 0)$, $(0.5, 0.5, 0)$, $(0.5, 0.25, 0.25)$ and all their permutations. From the allocation $(0.5, 0.25, 0.25)$, player 1, by pillaging 2, can achieve the allocation $(0.75, 0, 0.25)$. While the latter is not in the stable set, it allows player 1 to then pillage 3 and achieve the tyrannical allocation $(1, 0, 0)$, which is stable. In other words, $(1, 0, 0)$ is a farsighted objection to $(0.5, 0.25, 0.25)$. Note that if player 3 anticipates the second step in this move, she should not remain neutral when player 1 pillages 2.

Jordan (2006) formalizes this idea by explicitly introducing expectations. He shows that if otherwise neutral players act in accordance with the expected (final) outcome, then the stable set, D, is

[19] Although Jordan (2006) does not explicitly define an effectivity correspondence, our formulation is consistent with his.

indeed farsighted. As Ray and Vohra (2015) point out, this argument can also be made by suitably modifying the notion of a farsighted objection. Assume that *all* players are farsighted, including those who see no change in their payoff in a single step of a farsighted move. With this in mind, say that w' farsightedly dominates w if there is a collection of allocations $w^0, w^1, \dots, w^m$ (with $w^0 = w$ and $w^m = w'$) and a corresponding collection of coalitions, $S^1, \dots, S^m$, such that for all $k = 1, \dots m$:

$$w^{k-1}_{S^k} > w_L(w, w') \text{ where } L(w, w') = \{i \in N \mid w'_i < w_i\}$$

and

$$w'_{S^k} \gg w^{k-1}_{S^k}.$$

This must mean that

$$w_{W(w,w')} > w_{L(w,w')} \text{ where } W(w, w') = \{i \in N \mid w'_i > w_i\}.$$

Thus, w' farsightedly dominates w if and only if it (myopically) dominates w.

In the notion of farsightedness described in the previous paragraph whether a coalition can move from allocation w^{k-1} to w^k depends on the power of the winners and losers at the end of a sequence of moves. Formally, it does not conform to a framework in which the effectivity correspondence specifies which coalition(s) are effective in changing w^{k-1} to w^k, independently of where w^k will end up. For instance, in the three-player example, whether player 1 is effective in changing the allocation $(0.5, 0.25, 0.25)$ to $(0.75, 0, 0.25)$ cannot depend on any further changes that may be expected to take place. What is the farsighted stable set if adopt the effectivity correspondence specified in (1)? As our next result shows, it turns out to be identical to D_1, the core.

THEOREM **4.** *Suppose the effectivity correspondence is defined as in (1). Then the unique farsighted stable set is D_1, the core.*

Proof. Suppose Z is a farsighted stable set. It is obvious that no players have the power to beneficially change a tyrannical allocation since one player has already captured the entire surplus. It must therefore belong to every farsighted stable set (as well as to every REFS). It is easy to see allocations where two players get 0.5 are also stable in this sense. Thus, $D_1 \subseteq Z$.[20]

To complete the proof we will now show that for every $w \notin D_1$ there is a farsighted objection that terminates in D_1. There are two cases:

(i) $w \notin D_1$ is such that $w_i = w_j$ for all i, j such that $w_i > 0, w_j > 0$. This means that there are k players who receive $1/k$, where $k \geq 3$. Suppose two such players, say i and j, pillage a third and share the spoils equally. This increases the power of i and j. If there are any other players remaining with $1/k$, in the next step i and j pillage one such player. This process continues until we arrive at an allocation in D_1 where i and j get 0.5 each. This is clearly a farsighted objection.

(ii) There are i and j such that $w_i > w_j > 0$. Let i' be a player such that $w_{i'} \geq w_i$ for all i. Of course, i' can pillage a player with lower wealth. This results in i' becoming more powerful, and she can now pillage any other player j, with $w_j > 0$, if there is any. Through this process of

[20] The fact that the core is a subset of the farsighted core is a feature of pillage games. In general, it is possible that the core is disjoint from every farsighted stable set or REFS; recall Example 5.

sequential pillaging, i' can achieve the tyrannical allocation in which she has the entire wealth. This describes a farsighted objection, leading from w to a tyrannical allocation in D_1. ■

We now turn to a consideration of SREFS and REFS in this model. Dyadic allocations in which players with positive wealth share equally will play an important role in this analysis. For every nonnegative integer k, let $B_k = \{x \in \triangle \mid x_i = 0 \text{ or } x_i = 2^{-k}, \forall i\}$ and $B = \cup_k B_k$. Note that B_0 is the set of tyrannical allocations and $B_0 \cup B_1 = D_1$.

Our next Example will illustrate a crucial difference between REFS and SREFS in this model.

EXAMPLE **6.** *The pillage game with four players.*

The core, or $D_1 = B_0 \cup B_1$, in this example is the set of all permutations of allocations of the form $(1, 0, 0, 0)$ and $(0.5, 0.5, 0, 0)$. And B_2 consists of the equal-division allocation, $\bar{w} = (0.25, 0.25, 0.25, 0.25)$. As we noted in the proof of Theorem 4, every REFS contains D_1. In fact, D_1 is a REFS and so is B, but only the latter is a SREFS. Formal proofs of these assertions will follow from Theorems 5 and 6 below, but for now we provide a sketch in the four-player case to help understand the differences between REFS and SREFS.

Suppose F is a rational expectation and $Z = \Sigma(F)$ is the associated REFS. Observe that if w is allocation in which exactly two players have positive wealth and one of these is higher than the other, F must be such that the wealthier player pillages the other to achieve a tyrannical allocation. It can also be shown that Z does not contain any allocation in which exactly three players have positive wealth. For our purposes it will be enough to examine the change that F prescribes from the allocation $w' = (0.375, 0.375, 0.25, 0)$. (The analysis for other permutations of this allocation is similar). There are three possibilities, depending on $S(w')$, the coalition that is expected to form at w':

(a) $S(w') = \{1, 2\}$, i.e., players 1 and 2 pillage 3. In this case it is easy to see that $f(w') = f^*(w') = (0.5, 0.5, 0, 0)$;[21]

(b) $S(w') = \{1\}$, $f(w') = (0.625, 0.375, 0, 0)$ and $f^*(w') = (1, 0, 0, 0)$;

(c) $S(w') = \{2\}$ and $f^*(w') = (0, 1, 0, 0)$.

We can now determine whether or not $B_2 = \{\bar{w}\} \subseteq Z$. The only coalitions that have the power to change $\bar{w}$ are all two player coalitions and all three-player coalitions. A three-player coalition cannot be expected to form because, as we noted above, no allocation with three players having positive wealth can be stable (and one of these players will end up with 0). Is it possible that $S(\bar{w})$ consists of two players i and j, who pillage a third and move to a w' (or a permutation thereof)? This will indeed happen if If F conforms to case (a) because then i and j will both end up with 0.5 each. Of course, this implies that $\bar{w} \notin Z$. If, however, F conforms to cases (b) or (c) for every two-player coalition, then $\bar{w} \in Z$; no two player coalition will form because one of its members would eventually get pillaged. In fact, as we will confirm in Theorem 6, there is a rational expectation that supports $B_0 \cup B_1$ as a REFS and also one that supports B as

[21] When 1 and 2 pillage player 3 they must divide w'_3 equally for otherwise, the final outcome would be a tyrannical allocation and one of these players would not have agreed to form $\{1, 2\}$.

a REFS. But there is an important difference between these two cases. A rational expectation that satisfies (a) cannot be a *strong* rational expectation. This is so because one of the players could have done better by refusing to participate in the two player coalition, pillaging player 3 on her own (shifting to case (b) or (c)) and inducing a tyrannical allocation. A strong rational expectation must conform to case (b) or (c). In other words, B is a SREFS while $B_0 \cup B_1$ is not.

This example also illustrates why in defining the effectivity correspondence we ruled out the possibility of unsolicited gifts. Consider the strong rational expectation that supports B: from w', $S(w')$ is either (b) or (c), and $\bar{w}$ is stable. Suppose that from $\bar{w}$, contrary to (1), the effectivity correspondence were to allow players 1 and 2 to pillage player 4 and share the spoils equally with player 3. This leads to $\hat{w} = (1/3, 1/3, 1/3, 0)$ from which, unlike w', it is not possible for 1 or 2 *alone* to engage in any further pillage. But 1 and 2 *together* can pillage player 3 to obtain 0.5 each. This renders $\bar{w}$ unstable, and causes player 3 to eventually get pillaged, all because 1 and 2 were allowed to make her a gift.

We now turn to the general case of an arbitrary number of players.

THEOREM **5.** *Suppose the effectivity correspondence is as in (1). Then B is a SREFS.*

Proof. We construct an expectation F as follows:

(i) Suppose w is such that $w_i > w_j > 0$ for some i, j. Let i' be the lowest indexed player such that $w_{i'} \geq w_i$ for all i, and let j' be the lowest indexed player such that $w_{j'} < w_{i'}$. Then the expectation is that i' will pillage j': $f(w) = w'$ where $w'_{i'} = w_{i'} + w_{j'}$, $w'_{j'} = 0$ and $w_k = w'_k$ for all $k \neq i', j'$, and $S(w) = \{i'\}$. Note that $f^*(w)$ is the tyrannical allocation where i' gets the entire wealth.

(ii) Suppose w is such that all players with positive wealth have the same wealth but this is not 2^{-k} for any integer k. In other words, m players get $1/m$ but $m \neq 2^k$ for any integer k. Let $\hat{k}$ be the largest k such that $2^k < m$. Then $f(w) \in B_{\hat{k}}$, and $S(w)$ is the coalition consisting of the lowest indexed $2^{\hat{k}}$ players getting $1/m$ at w. Note that $S(w)$ has the power to make this move since the total wealth of this coalition at w is $\frac{2^{\hat{k}}}{m} = \frac{1}{2}\frac{2^{\hat{k}+1}}{m} > \frac{1}{2}$.

(iii) For $w \in B$, $f(w) = w$.

We have constructed F such that $\Sigma(F) = B$. It remains to be shown that F satisfies conditions (I), (E) and (M').

Suppose F does not satisfy (I). Then there exists $w \in B_k$ and $S \in E(w, w')$ such that $f^*(w') = w'' \in B_{k'}$ and $w''_S \gg w_S$. The last inequality implies that $k' < k$.

First, suppose that w' is such that $w'_i > w'_j$ for some pair i, j. Then, w'' is a tyrannical allocation, so that $|S| = 1$. But, then $S \notin E(w, w')$ since $w_i = w_j$ if $w_i, w_j > 0$.

So, w' must satisfy $w'_i = w'_j$ if $w'_i, w'_j > 0$. Also, since E satisfies equation (1), $S = \{i | w'_i > w_i\}$. Putting these together, we must have $w' = w''$. That is, $w' \in B_{k'}$.

Since $w'_i > 0$ implies that $w'_i = 2^{-k'}$ and $w_i > 0$ implies that $w_i = 2^{-k}$, this means that $w'_i \geq 2w_i$ for $i \in S$ - those with positive wealth at w' must have at least twice as much as they did at w. Since the added wealth must have been pillaged, those who were pillaged must have had at least as much wealth at w as the pillagers. So,

$$\sum_{\{i \in S\}} w_i \leq \sum_{\{i : w'_i = 0\}} w_i.$$

This implies that $S \notin E(w, w')$.

To see that (E) is satisfied, consider $w \notin B$. If w is covered by Case (i), the only coalition that moves at each step is the singleton consisting of the lowest indexed player with the highest wealth at w. And, at each step, this coalition does better by eventually attaining the tyrannical allocation. Thus (E) holds for w in Case (i). For w covered by Case (ii), $S(w)$ moves in one step to a stationary allocation which is an improvement since it involves equal sharing in a smaller coalition, and Condition (E) is therefore satisfied.

We now turn to Condition (M'). For w in Case (i) maximality is trivially satisfied since the singleton that moves ultimately achieves the tyrannical allocation. Suppose w is covered by Case (ii) and (M') is not satisfied. This means that there is a coalition T with $T \cap S(w) \neq \emptyset$ that does better than $f(w)$. Since all stationary allocations satisfy equal division among players with positive wealth, $|T| < |S(w)|$. Recall that $|S(w)| = 2^{\hat{k}}$, which implies that if $|T| = 2^{k'}$ for some integer $k' < \hat{k}$, T does not have the power to change w. Thus, $T \neq 2^{-k}$ for some positive integer k, in which case the final outcome according to F will result in a smaller coalition and some player in T will get 0. This contradicts the supposition that T can do better than $f(w)$. ■

Since B is a SREFS it is also a REFS. But, as our next result shows, there are several other REFS, including the unique farsighted stable set: $B_0 \cup B_1 = D_1$. In this model, therefore, unlike simple games, the farsighted stable set can be justified on the basis of consistent and rational expectations. It does not, however, meet the strong maximality test.

For any positive integer n, let $k(n)$ be the largest integer such that $2^k \leq n$.

THEOREM **6.** *For any positive integer $k^* \leq k(n)$, $\cup_0^{k^*} B_k$ is a REFS.*

Proof. Choose any positive integer $k^* \leq k(n)$, and define $B(k^*) \equiv \cup_0^{k^*} B_k$.

For any w, define $H(w) = \{i \in N : w_i \geq w_j \ \forall j \in N\}$, and let $\bar{H}(w)$ be the subset of $H(w)$ consisting of the 2^k-lowest ranked players in $H(w)$, where k is the largest integer not exceeding k^* with $2^k \leq |H(w)|$.

Given k^*, F is defined as follows.

(i) If $w \in B(k^*)$, then $f(w) = w$.

(ii) If $w \notin B(k^*)$ and $|H(w)| \geq 2^{k^*}$, let j' be the lowest-indexed agent not in $\bar{H}(w)$ with $w_j > 0$. Such j must exist since $w \notin B(k^*)$. Then, $S(w) = \bar{H}(w)$ and $f(w) = w'$ where

$$w'_i = \begin{cases} w_i + \frac{w_{j'}}{|\bar{H}(w)|}, & \text{if } i \in \bar{H}(w) \\ w_i, & \text{if } i \notin \bar{H}(w) \cup \{j'\} \\ 0, & \text{if } i = j' \end{cases}$$

(iii) If $w \notin B(k^*)$, $|H(w)| < 2^{k^*}$ and there is a pair i, j with $w_i > w_j > 0$, then the lowest-ranked agent $i^* \in H(w)$ pillages the lowest-ranked agent $j^* \notin H(w)$. So, $S(w) = \{i\}$ and $f(w) = w'$ where

$$w'_i = \begin{cases} w_i + w_{j^*}, & \text{if } i = i^* \\ w_i, & \text{if } i \neq i^*, j^* \\ 0, & \text{if } i = j^* \end{cases}$$

(iv) If $w \notin B(k^*)$, $|H(w)| < 2^{k^*}$ and if $H(w) = \{i | w_i > 0\}$, then $S(w) = \bar{H}(w)$ and $f(w) = w'$ where

$$w'_i = \begin{cases} \frac{1}{|\bar{H}(w)|}, & \text{if } i \in \bar{H}(w) \\ 0, & \text{if } i \notin \bar{H}(w) \end{cases}$$

We note that in (iv) above, $|\bar{H}(w)| > |(H(w) - \bar{H}(w))|$ and so $\bar{H}(w)$ can pillage the rest, and hence f is well-defined.

The proof that F satisfies (I) is virtually identical to that in Theorem 5, and we only give a very short sketch of the proof. Again, suppose $w \in B_k$ with $k \leq k^*$ and some S has a farsighted objection ending in $B_{k'}$ where $k' < k$. Then, $|S| \geq 2$ since no singleton has the power to pillage anyone at w. But, then the first move from w must be to some w' which is an equal allocation - any unequal allocation terminates in a tyrannical allocation. Just as before, w' itself must be a stationary allocation, and then S cannot have the power to pillage the remaining players.

We now check (M). In cases (ii) and (iv), each $i \in \bar{H}(w)$ ends up getting $\frac{1}{|\bar{H}(w)|}$. There cannot be a better deviation. In case (iii), the agent initiating the deviation ends getting 1. Hence, (M) is satisfied.

Finally, it is easy to check that (E) is satisfied. In each of cases (ii) to (iv), $S(w)$ has a farsighted objection culminating in some allocation in $B(k^*)$.

This completes the proof of the theorem. ■

We close this section with a discussion of Acemoglu *et al.* (2008). They study a model of political coalition formation in which the power of each player is exogenously given. For each $i \in N$, $\gamma_i > 0$ denotes i's political power. The power of coalition S is $\gamma_S = \sum_{i \in S} \gamma_i$. Coalition $S \subseteq T$ is *winning in* T if $\gamma_S > \alpha\gamma_T$, where $\alpha \in [0.5, 1)$. Denote by $\mathcal{W}(T)$ the set of subsets of T that are winning in T. If such a coalition exercises its power, it captures the entire surplus and becomes *the ruling coalition*. The other players are eliminated and play no further role. However, the ruling coalition may itself be subject to a new round of power grab from within.

The distribution of wealth is determined through an exogenous rule that depends only on the identity of the ruling coalition. Assume that for every player it is better to be in a ruling coalition that not. Moreover, it is better to be in a ruling coalition with lower aggregate power. As Acemoglu *et al.* (2008) point out, a particular example of such a rule, which shall adopt for the sake of simplicity, is the following:

$$w_i(S) = \begin{cases} \gamma_i/\gamma_S & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases}$$

A state can now be defined as the ruling coalition: at state S, the ruling coalition is S and the wealth distribution is $w(S)$. The set of states is therefore $\mathcal{N}$. Winning coalitions are the ones effective in changing a state:

$$S \in E(T, S) \text{ if and only if } S \text{ is winning in } T. \tag{2}$$

This means that if a change occurs, the new ruling coalition is smaller. If such a change is expected to lead to a further change, then a winning coalition will choose *not* to exercise its power. This is a result of the fact that any further change must leave some member(s) of the original winning coalition with a payoff of 0. In other words, if there is a farsighted objection $S, S^1, \ldots S^m$ leading from S to S^m it must be the case that $m = 1$; farsighted dominance is equivalent to dominance.[22] Harsanyi (1974) refers to a farsighted dominance relation with this property as *trivial* and points out that if this property holds for every farsighted dominance of one state over another, then the vNM stable set is equivalent to the farsighted stable set.

Another feature of this model that makes it very tractable is that objections can only come from subsets of the ruling coalition (internal blocking). This makes it possible to construct a stable set recursively; see Ray and Vohra (2014). Of particular interest in these models is the stability of N, or the stable state(s) starting from N. We illustrate this through the following example.

EXAMPLE **7.** *(Four-player example of the Acemoglu et al. (2008) model). Suppose* $N = \{1, 2, 3, 4\}$, $\gamma = (2, 4, 6, 8)$ *and* $\alpha = 0.5$.

A vNM stable set can be constructed as follows. Any ruling coalition consisting of one individual clearly belongs to the stable set (it is in the core). Any ruling coalition consisting of two players is not in the stable set because the more powerful player will eliminate the weaker one; there is an objection leading to a stable state. Next, consider the three-player coalitions. The coalition $\{1, 2, 4\}$ is not stable because player 4 has enough power to eliminate the other two. Let the collection of the other three-player coalitions be denoted $\mathcal{S} = \{\{1, 3, 4\}, \{2, 3, 4\}, \{1, 2, 3\}\}$. It is easy to see that no coalition in $\mathcal{S}$ is threatened by a single powerful player. In each instance, two of the players have enough power to eliminate the third, but the resulting outcome is not stable, as we have just noted. This means that all coalitions in $\mathcal{S}$ are stable, and N is not; it will be replaced by one of these three-player coalitions. Thus, the stable set consists of singletons and the collection $\mathcal{S}$. In fact, this is also a farsighted stable set because farsighted dominance is equivalent to myopic dominance in this model. To verify this directly in this Example, it is only necessary to establish the farsighted internal stability for states in $\mathcal{S}$. While two players could

[22] Recall that in Jordan's pillage game a farsighted objection could last several steps, although at each step it would be the same coalition making the change. This difference stems from the fact that in the Acemoglu *et al.* model only the winning coalition survives to the next stage; there are no neutral players.

eliminate a third, this cannot result in a farsighted objection ending in a stable state because the weaker of the two will get eliminated at the next stage. We conclude that N is not stable and will replaced be one of the coalitions in $\mathcal{S}$.[23]

Recall that farsighted internal stability is stronger than condition (I) of REFS and myopic external stability is stronger than condition (E) of REFS. Consequently, the equivalence of farsighted dominance and myopic dominance also yields equivalence with REFS. More precisely, consider a set of states Z that is a stable set as well as a farsighted stable set. For every state $S \in Z$, let $F(S) = S$. (Because $X = \mathcal{N}$, we abuse notation slightly to consider F as a function from $\mathcal{N} \to \mathcal{N}$). For $S \notin Z$ define $F(S)$ to be a subcoalition of S that dominates it (myopically) and belongs to Z. If there are several such coalitions, pick one arbitrarily. By construction, F satisfies (E). It satisfies (I) because Z is farsighted stable set. Finally, it satisfies (M) because for every nonstationary state, S, it prescribes a move by coalition $F(S)$ that ends with $F(S)$. Since this is the *only* profitable move available to $F(S)$ it is trivially maximal.

Acemoglu *et al.* (2008) provide an axiomatic characterization of a solution to this model and also show that it coincides with the subgame perfect equilibria of a noncooperative model of coalition formation. Their solution is a refinement of REFS, or the stable set. This difference turns out to hinge on the difference between REFS and SREFS. In fact, in this model SREFS refines REFS precisely to the Acemoglu *et al.* solution.[24] This can be illustrated through Example 7. As explained above, in constructing a rational expectation F we have the freedom to choose $F(N)$ to be any one of the three coalitions in $\mathcal{S}$. In particular, we could define $F(N) = \{2, 3, 4\}$. But players 3 and 4 could do better by forming $\{1, 3, 4\}$; recall that the payoff to a player is higher in a coalition with lower aggregate power. In other words, F does not satisfy (M'). In fact, strong maximality in this model, not just in Example 7, reduces to the condition that if S is not a stationary state, then $F(S)$ has the *lowest* aggregate power among all stable coalitions that are winning in S:

$$\text{If } S \notin \Sigma(F), \text{ then } F(S) \in \text{argmin}_{T \in \Sigma(F) \cap \mathcal{W}^*(S)} \gamma_T, \text{ where } \mathcal{W}^*(S) \text{ denotes } \mathcal{W}(S) - S.$$

If γ is generic in the sense that $\gamma_S \neq \gamma_T$ for any S, T, $S \neq T$, then clearly $F(S)$ is unique for every S. And the unique strong rational expectation can be computed recursively as follows. Of course, $F(S) = S$ if $|S| = 1$. Suppose $F(S)$ has been defined for all S such that $|S| < k$. Then for S with $|S| = k$,

$$F(S) = \begin{cases} \text{argmin}_{T \in \Sigma(F) \cap \mathcal{W}^*(S)} \gamma_T & \text{if } \Sigma(F) \cap \mathcal{W}^*(S) \neq \emptyset \\ S & \text{otherwise} \end{cases}$$

$F(S)$ is the same as $\phi(S)$, the *ultimate ruling coalition* for player set S in Acemoglu *et al.* (2008).

[23] The singletons are also stable, but none of those states are reachable from N.

[24] Ray and Vohra (2014) show that their notion of an EPCF yields the same predictions as REFS but an appropriately chosen protocol sharpens the equilibria to coincide with the Acemoglu *et al.* solution. See also Kimya (2015).

REFERENCES

Acemoglu, D., Egorov, G., and K.SONIN (2008), "Coalition Formation in Non-democracies", *Review of Economic Studies*, **75**, 987–1009.

AUMANN, R. and R. MYERSON (1988), "Endogenous Formation of Links Between Players and of Coalitions, An Application of the Shapley Value," in *The Shapley Value: Essays in Honor of Lloyd Shapley*, A. Roth, ed., 175–191. Cambridge: Cambridge University Press.

BERNHEIM, D., PELEG, B. and M. WHINSTON (1987), "Coalition-Proof Nash Equilibria. I. Concepts," *Journal of Economic Theory*, **42**, 1–12.

CHWE, M. (1994), "Farsighted Coalitional Stability," *Journal of Economic Theory*, **63**, 299–325.

BLOCH, F. (1996), "Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division," *Games and Economic Behavior*, **14**, 90–123.

CHANDER, P. (2015), "An Infinitely Farsighted Stable Set", mimeo, Jindal Global University.

DIAMANTOUDI, E. AND L. XUE (2003), "Farsighted Stability in Hedonic Games," *Social Choice and Welfare*, **21**, 39–61.

GREENBERG, J. (1990), *The Theory of Social Situations*, Cambridge, MA: Cambridge University Press.

HARSANYI, J. (1974), "An Equilibrium-Point Interpretation of Stable Sets and a Proposed Alternative Definition," *Management Science*, **20**, 1472–1495.

HERINGS, P., A. MAULEON, AND V. VANNETELBOSCH (2004), "Rationalizability for Social Environments," *Games and Economic Behavior*, **49**, 135–156.

——— (2009),"Farsightedly Stable Networks," *Games and Economic Behavior*, **67**, 526–541.

JORDAN, J. (2006), "Pillage and Property," *Journal of Economic Theory* **131**, 26–44.

KIMYA, M. (2015),"Equilibrium Coalitional Behavior", mimeo, Brown University.

KONISHI, H. AND D. RAY (2003), "Coalition Formation as a Dynamic Process," *Journal of Economic Theory*, **110**, 1–41.

LUCAS, W. (1968), "A Game with No Solution," *Bulletin of the American Mathematical Society*, **74**, 237–239.

——— (1992), "von Neumann-Morgenstern Stable Sets," in *Handbook of Game Theory, Volume 1*, ed. by R. J. Aumann, and S. Hart, 543–590. North Holland: Elsevier.

MAULEON, A. AND V. VANNETELBOSCH (2004), "Farsightedness and Cautiousness in Coalition Formation Games with Positive Spillovers," *Theory and Decision*, **56**, 291–324.

MAULEON, A., V. VANNETELBOSCH AND W. VERGOTE (2011), "von Neumann-Morgenstern farsighted stable sets in two-sided matching," *Theoretical Economics*, **6**, 499–521.

PICCIONE, M., AND A. RUBINSTEIN (2007), "Equilibrium in the Jungle", *The Economic Journal*, **117**, 883–896.

RAY, D. (2007), *A Game-Theoretic Perspective on Coalition Formation*, Oxford University Press.

RAY, D. and R. VOHRA (1997), "Equilibrium Binding Agreements," *Journal of Economic Theory*, **73**, 30–78.

——— (1999), "A Theory of Endogenous Coalition Structures," *Games and Economic Behavior*, **26**, 286–336.

——— (2014), "Coalition Formation," in *Handbook of Game Theory, Volume 4*, ed. by H. P. Young and S. Zamir, 239–326. North Holland: Elsevier.

——— (2015), "The Farsighted Stable Set", *Econometrica*, **83**, 977–1011.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Xue, L. (1998), “Coalitional Stability under Perfect Foresight,” *Economic Theory*, **11**, 603–627.

Appendix

Our proof of Theorem 2 will make use of the following lemma.

Lemma **1.** *There exists a positive number $d < 1/|C|$, a vector $b \in R_+^{N-C}$ and a non-empty collection of coalitions $\mathcal{J}$ in $N - C$ such that:*

(1) For every $J \in \mathcal{J}$, $C \cup J$ is a minimal winning coalition, $b_J \gg 0$ and $\sum_{j \in J} b_j = \epsilon \equiv 1 - d|C|$.

(2) There does not exist a winning coalition $C \cup T'$ such that $\sum_{j \in T'} b_j < \epsilon$.

(3) $1 > \sum_{j \in N-C} b_j > \epsilon$.

Proof. Let $\mathcal{J}' = \{J \subset N - C \mid C \cup J \text{ is a minimal winning coalition}\}$. Without loss of generality, assume that the coalitions in $\mathcal{J}' = \{J^1, \ldots, J^K\}$ are ranked in non-decreasing order of cardinality, so $|J^k| \leq |J^{k+1}|$ for all $k = 1, \ldots, K-1$.

Choose $\epsilon \in (0,1)$ such that $\epsilon < \frac{|J^1|}{|N-C|}$ and let $d \equiv \frac{1-\epsilon}{|C|}$. We will now construct an algorithm which will yield b and $\mathcal{J} \subseteq \mathcal{J}'$ satisfying the desired properties.

Let J^f, $f > 1$, be the first coalition in $\mathcal{J}'$ which has a non-empty intersection with some J^k, $k < f$. If no such f exists, then set $f = K+1$, and $J^{K+1} = \emptyset$.

Step 1: For all $k < f$ let $b_i = \frac{\epsilon}{|J^k|}$ for all $i \in J^k$. Note that for all $k, k' < f$, $J^k \cap J^{k'} = \emptyset$, so this construction is well-defined. Clearly, b and J^k satisfiy Condition (1) of the Lemma for all $k < f$. Moreover, since the coalitions in $\mathcal{J}'$ are in non-decreasing order of cardinality:

(L.1) for any $k < k' < f$, $i \in J^k$ and $j \in J^{k'}$, we have $\bar{b} = \frac{\epsilon}{|J^1|} \geq b_i \geq b_j > 0$.

Define $A^1 = \cup_{k=1}^{f-1} J^k$. Then, for every $i \in A^1$, b_i as defined above will be the “terminal“ value. For $i \notin A^1$ we will construct b_i iteratively.

For every $k \geq f$, we will recursively define non-negative numbers t_i^1 for all $i \in J^k - A^1$ as follows. Suppose t_j^1 have been defined for all $j \in G^k \equiv \cup_{j=1}^{k-1} J^j - A^1$. If $J^k - (G^k \cup A^1) \neq \emptyset$, for every $i \in J^k - (G^k \cup A^1)$ let

$$t_i^1 = \max\left[0, \frac{\epsilon - \sum_{j \in J^k \cap A^1} b_j - \sum_{j \in J^k \cap G^k} t_j^1}{|J^k - (G^k \cup A^1)|}\right]$$

Since there are no dummy players, every $i \in N - C$ belongs to at least one coalition in $\mathcal{J}'$, and has therefore been assigned a non-negative number, b_i or t_i^1. It follows from (L.1) that for all $i \notin A^1$, $t_i^1 \leq \bar{b}$.

Let $\mathcal{J}^1 = \{J \in \mathcal{J}' | \sum_{i \in J \cap A^1} b_i + \sum_{i \in J - A^1} t_i^1 < \epsilon\}$.

If $\mathcal{J}^1 = \emptyset$, terminate the algorithm, set $b_i = t_i^1$ for all $i \notin A^1$ and go to Step 3.

Suppose $\mathcal{J}^1 \neq \emptyset$. Since $J \in \mathcal{J}^1$ implies that $|J| \geq |J^k|$ for all $k < f$ and $J \in \mathcal{J}^1$, it follows from (L.1) that for every $J \in \mathcal{J}^1$, $J - A^1 \neq \emptyset$. Let J^{k_1} be a coalition in $\mathcal{J}^1$ which maximizes $\frac{\epsilon - \sum_{j \in J^k \cap A^1} b_j}{|J^k - A^1|}$, ties being broken arbitrarily. For each $i \in J^{k_1} - A^1$, set

$$b_i = \frac{\epsilon - \sum_{j \in J^{k_1} \cap A^1} b_j}{|J^{k_1} - A^1|}$$

Since $J^{k_1} - A^1 \neq \emptyset$, $A^2 \equiv A^1 \cup J^{k_1}$ is a strict superset of A^1.

Step 2: Since some of the components of the t vector have been increased, the remaining components may not be feasible. So, now repeat step 1 with A^2 replacing A^1.

For $k \geq f$, $k \neq k_1$, we will recursively define t_i^2 recursively as follows. Suppose t_i^2 have been defined for all $j \in G'^k \equiv \cup_{j=1}^{k-1} J^j - A^2$. If $J^k - (G'^k - A^2) \neq \emptyset$, for every $i \in J^k - (G'^k \cup A^2)$ let

$$t_i^2 = \max\left[0, \frac{\epsilon - \sum_{j \in J^k \cap A^2} b_j - \sum_{j \in J^k \cap G'^k} t_j^2}{|J^k - A^2|}\right]$$

As in Step 1, it follows from (L.1) that $t_i^2 \leq \bar{b}$ for all $i \notin A^2$. Let $\mathcal{J}^2 = \{J \in \mathcal{J} | \sum_{i \in J \cap A^2} b_i + \sum_{i \in (J - A^2)} t_i^2 < \epsilon\}$. By construction, $J^{k_1} \notin \mathcal{J}^2$. If $\mathcal{J}^2$ is empty, terminate the algorithm with $b_i = t_i^2$ for all $i \in N - (C \cup A^2)$ and move to Step 3. Otherwise, choose the coalition J^{k_2} which maximizes $\frac{\epsilon - \sum_{j \in J^k \cap A^2} b_j}{|J^k - A^2|}$ in this set. For each $i \in J^{k_2} - A^2$, set

$$b_i = \frac{\epsilon - \sum_{j \in J^{k_2} \cap A^2} b_j}{|J^{k_2} - A^2|}$$

Claim: $J^{k_2} - A^2$ is non-empty.

Suppose not. Since $J^{k_2} \in \mathcal{J}^2$, and by hypothesis, $J^{k_2} \subseteq A^2$, $\sum_{i \in J^{k_2} \cap A^2} b_i = \sum_{i \in J^{k_2} \cap A^1} b_i + \sum_{i \in S} t_i^1 < \epsilon$. Recall that for $i \in S \subseteq J^{k_1} - A^1$, $t_i^1 = \frac{\epsilon - \sum_{i \in A^1 \cap J^{k_1}} b_i}{|J^{k_1} - A^1|}$. But this means that

$$\frac{\epsilon - \sum_{i \in J^{k_2} \cap A^1} b_i}{|S|} > \frac{\epsilon - \sum_{i \in J^{k_1} \cap A^1} b_i}{|J^{k_1} - A^1|},$$

which contradicts the choice of J^{k_1}. Hence the claim is true, and $A^3 \equiv A^2 \cup J^{k_2}$ is a strict superset of A^2.

Since the sets A^k are strictly increasing over stages, the algorithm terminates.

Step 3. At this stage we have constructed b such that $b_i \leq \bar{b}$ for all $i \in N - C$, and for all $J \in \mathcal{J}'$, $d|C| + \sum_{i \in J} b_j \geq 1$. Clearly, then Condition (2) of the Lemma holds. Define $\mathcal{J} = \{J \in \mathcal{J}' \mid b_J \gg 0 \text{ and } \sum_{i \in J} b_i = \epsilon\}$, which is non-empty because $J^k \in \mathcal{J}$ for all $k < f$. Of course, $\mathcal{J}$ satisfies Condition (1).

Since $b_i \leq \bar{b}$ for all $i \in N - C$, $\sum_{i \in N-C} b_i \leq |N - C|\bar{b} = |N - C|\frac{\epsilon}{|J^1|} < 1$. To establish Condition (3) of the Lemma it remains to be shown that $\sum_{j \in N-C} b_j > \epsilon$. Recall that for every $J \in \mathcal{J}$, $\sum_{i \in J} b_i = \epsilon$, which implies that this condition clearly holds if $f > 2$. Suppose $f = 2$, i.e., $J^1 \cap J^2 \neq \emptyset$. Since J^1 and J^2 are minimal winning coalitions, neither is a subset of the other. By construction, $b_i > 0$ for all $i \in J^1 \cup J^2$. This together with the fact that $\sum_{i \in J^1} b_1 = \epsilon$ implies that $\sum_{i \in N-C} b_i \geq \sum_{i \in J^1 \cup J^2} b_i > \epsilon$.

■

Proof of Theorem 2. We will construct a SREFS for two distinct cases. The first is the case, as in Example 5, where minimal winning coalitions consist of C and any one of the non-veto players. For the second case, in which there is at least one minimal winning coalition with two or more non-veto players, our construction will rely on Lemma 1 and Assumption 2.

Case 1: $C \cup \{j\} \in \mathcal{W}$ for all $j \notin C$.

Let $a > 0$ and $b > 0$ be such that $|C|a + |N - C|b = 1$. Define $\hat{x}$ so that $\pi(\hat{x}) = N$ and

$$u_i(\hat{x}) = \begin{cases} a & \text{if } i \in C \\ b & \text{if } i \notin C \end{cases}$$

For every $j \notin C$, let X^j be the set of all states x in which the winning coalition contains $C \cup \{j\}$ and the payoff vector has the property that $u_i(x) \geq a$ for all $i \in C$, $u_j(x) = b$ and $u_k(x) = 0$ for all $k \notin C \cup \{j\}$.

We claim that $Z = \cup_{j \notin C} X^j \cup \{\hat{x}\}$ is a SREFS: it is the set of stationary points of a strong rational expectation F satisfying the following properties.[25]

(1.1) For $x \in Z$, $f(x) = x$.

(1.2) For $x \in X^0$, $S(x) = N$ and $f(x) = \hat{x}$.

The remainder of the rules for F relate to $x \in X - Z - X^0$:

(1.3) For x such that $u_j(x) < b$ for all $j \in N - C$, $S(x) = N - C$ and $f(x) \in X^0$.

(1.4) For x such that $u_j(x) \geq b$ for all $j \in N - C$, since $x \notin Z$, there must be a veto player i for whom $u_i(x) < a$. Let $\{i'\}$ be the lowest indexed player of this kind. In this case we have $S(x) = \{i'\}$ and $f(x) \in X^0$.

(1.5) For x such that $u_j(x) \geq b$ for some non-veto player but not all, let j' be the lowest indexed non-veto player such that $u_{j'}(x) < b$. There are now two distinct cases for describing F.

(1.5.a) Suppose $s(x) \equiv 1 - b - \sum_{i \in C} \max\{a, u_i(x)\} > 0$, then $S(x) = C \cup \{j'\}$, $u_i(f(x)) = \max\{a, u_i(x)\} + \frac{s}{|C|}$ and $u_{j'}(f(x)) = b$. Note that $f(x) \in X^{j'}$.

[25] In what follows it will be understood that $\pi(f(x))$ is the immediate change in $\pi(x)$ resulting from the formation of $S(x)$, as formalized in Assumption 1 (b).

(1.5.b) Suppose $s(x) \leq 0$. In this case, unlike the previous one, it is not possible to construct an objection that leads to Z in one step. However, it must be the case that there is a veto player i for whom $u_i(x) < a$. Otherwise, $s(x) = 0$, which contradicts the supposition that $x \notin Z$. In this case, as in (4), $S(x) = \{i'\}$ and $f(x) \in X^0$.

For $x \notin Z$, $f^*(x) = \hat{x}$ in all cases except (1.5.a). It is easy to see that in every instance all the players in $S(x)$ prefer $f^*(x)$ to x which means that F satisfies (E).

Since all non-veto players weakly prefer $\hat{x}$ to any other state in Z, and a winning coalition coalition must include at least one such player, it follows that (M') is satisfied in case (1.2). The same reasoning applies to case (1.3). In case (1.4) player i' cannot construct another farsighted objection as none of the non-veto players are interested in moving, which implies that this move is strongly maximal. In case (1.5.a) there is no state in Z that all the veto players prefer to $f(x)$ and so (M') holds. Finally, in case (1.5.b) it is clear that player i' cannot construct an objection leading to any other state in Z. Thus, in all cases (M') is satisfied.

To see that F satisfies (I), consider a possible farsighted move from $\hat{x}$ that ends at some $x \in X^j$. Since all non-veto players weakly prefer $\hat{x}$ to x, none of them can be part of the first move from $\hat{x}$. But then the first move must lead to a state in X^0, which only results in returning to $\hat{x}$, making it impossible for the initiating coalition to gain. Next, consider a move from $x \in X^j$ that ends up at $\hat{x}$. Since all players in $C \cup \{j\}$ weakly prefer x to $\hat{x}$, the first move must come from non-veto players other than j. But then Assumption 1 (c) on the effectivity correspondence implies that the payoff remains $u(x)$ through the entire sequence of moves, contradicting the supposition that the state eventually becomes $\hat{x}$. Finally, consider the possibility of a farsighted move that leads from $x^j \in X^j$ to some $x^k \in X^k$, $k \neq j$. Again, by Assumption 1 (c), the first coalition in such a move cannot be from a coalition that is disjoint from $C \cup \{j\}$. Moreover, such a coalition cannot include j, since j prefers x^j to x^k. It cannot include C since it's impossible for all $i \in C$ to gain in moving from x^j to x^k. The only remaining possibility is that includes some strict subset of C. But that leads to a state in X^0, from which the final outcome is $\hat{x}$, not x^k.

This completes the the proof of Case 1.

Case 2: There exists a minimal winning coalition $C \cup J$ such that $J \subset N - C$ and $|J| \geq 2$.

Let d, b and $\mathcal{J}$ be as in Lemma 1. Define $a \equiv \frac{1-\sum_{i \in N-C} b_i}{|C|}$. By Condition (3) of lemma 1, $a \in (0, d)$.

Let $S^* = \cup_{J \in \mathcal{J}} J$, and let $N^* = C \cup S^*$. Since $C \cup J$ is a winning coalition for every $J \in \mathcal{J}$, clearly N^* is a winning coalition.

Define $\hat{X}$ to consist of all states x such that $W(x) \supseteq N^*$ and

$$u_i(x) = \begin{cases} a & \text{if } i \in C \\ b_i & \text{if } i \notin C \end{cases}$$

Corresponding to each $J^k \in \mathcal{J}$, define X^k as the set of states in which the winning coalition contains $C \cup J^k$ and the payoff vector corresponding to $x^k \in X^k$ is

$$u_i(x^k) = \begin{cases} d & \text{if } i \in C \\ b_i & \text{if } i \in J^k \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{X}$ be the set of all (nonzero) states x such that

$$u_i(x) = \begin{cases} d & \text{if } i \in C \\ b_i & \text{if } i \in W(x) - C \\ 0 & \text{otherwise} \end{cases}$$

Of course, corresponding to every $J^k \in \mathcal{J}$, $X^k \subseteq \bar{X}$. However, $\bar{X}$ may also include a state x with $W(x) = C \cup K$ and $K \notin \mathcal{J}$ because b_K is not *strictly* positive.

We claim that $Z = \bar{X} \cup \hat{X}$ is a SREFS.

To prove this we will construct a rational expectations function F with the following properties: (a) $f(x) = x$ for all $x \in Z$, (b) for $x \in X - Z - X^0$, $f(x) \in X^0$, and (c) for $x \in X^0$, $f(x) \in Z$, depending on the nature of $\pi(x)$.

Let $T = \{i \in N - C \mid C \cup \{i\} \in \mathcal{W}\}$. Note that if $i \in T$, then $\{i\} \in A^1$ as constructed in the proof of Lemma 1. Moreover, given that we are considering Case 2, A^1 also includes at least one coalition J such that $|J| \geq 2$. This means that T is a strict subset of S^*. It is also easy to see from the proof of Lemma 1 that $\mathcal{J}$ includes at least two distinct coalitions, which implies that $|S^*| \geq 3$.

To describe the transition from zero states, we partition X^0 into *three* disjoint sets. X^0_1 is the set of all zero states in which the coalition structure contains precisely *one* two-player coalition consisting of one player from C and the other from S^*. X^0_2 is the set of zero states containing precisely *two* two-player coalitions: $\{i, j\}$ and $\{i', k\}$ such that $i, i' \in C$, $j \in T \subset S^*$ and $k \in S^* - T$. The set of all other zero states is denoted X^0_3.

For each $k \in S^*$ pick a unique $J(k) \in \mathcal{J}$ that contains k. The existence of such $J(k)$ follows from condition (1) of Lemma 1. With some abuse of notation, let X^k refer to the set of states in which the winning coalition is $C \cup J(k)$, the payoff is d for all $i \in C$ and b_j for all $j \in J(k)$.

We now provide a complete description of F.

(2.1) For $x \in Z$, $f(x) = x$.

(2.2) For $x \in X^0_3$, $S(x) = N^*$ and $f(x) \in \hat{X}$.

(2.3) For $x \in X^0_1$, let (i, k) with $i \in C$ and $k \in S^*$ be the unique two-player coalition of this form in $\pi(x)$. Let $S(x) = C \cup J(k)$, and $f(x) \in X^k$.

(2.4) For $x \in X^0_2$, let (i, k) with $i \in C$ and $k \in S^* - T$ be the unique pair of this form in $\pi(x)$. Let $S(x) = C \cup J(k)$ and $f(x) \in X^k$.

The remainder of the rules for F relate to $x \in X - Z - X^0$.

(2.5) There is $k \in S^*$ such that $u_k(x) < b_k$ and $i \in C$ such that $u_i(x) < d$.

There are three subcases to consider:

(2.5.1) Either $|C| \neq 2$ or $W(x) = C \cup J$ where $|J| \geq 3$.

Let H be the two-player coalition consisting of the lowest ranked player $i \in C$ with $u_i(x) < d$ and the lowest ranked player $k \in S^*$ such that $u_k(x) < b_k$. Define $S(x) = H$. If H is a winning coalition, let $f(x) \in X^k$. Otherwise, $x' = f(x) \in X^0$. Now, since $|C| \neq 2$ or $|J| \geq 3$, invoking Assumption 1 (b), H is the unique coalition in $\pi(x')$ consisting of one player from C and another from S^*. Thus, $x' \in X_1^0$ and $f(x') \in X^k$ with $u_i(f(x')) = d > u_i(x)$ and $u_k(f(x')) = b_k > u_k(x)$. Thus, H has a farsighted objection to x that leads to a state in X^k.

(2.5.2) $|C| = 2$ and $W(x) = C \cup J$ where $|J| = 2$.

Since $|S^*| \geq 3$ and $|J| = 2$, there exists $k \in S^* - J$. Since $k \notin W(x)$, $u_k(x) = 0$. Without loss of generality, let k be the lowest ranked player in $S^* - J$. Let $S(x) = H = \{i, k\}$. Clearly, for $x' = f(x)$, H is the unique coalition in $\pi(x')$ with one player from C and another from S^*. Thus, $x' \in X_1^0$ and, as in the previous paragraph, $f^*(x) \in X^k$.

(2.5.3) $|C| = 2$ and $W(x) = C \cup \{j\}$ for some $j \in T$.

Recall that T is a strict subset of S^*, i.e., there exists $k \in S^* - T$. Of course $k \neq j$ and $u_k(x) = 0 < b_k$. Let i be the lowest ranked player in C such that $u_i(x) < d$ and let $S(x) = H = \{i, k\}$. Note that $x' = f(x) \in X_2^0$, with H as the unique two-player coalition in $\pi(x')$ with one player from C and another from $S^* - T$. This means that $f(x') \in X^k$.

(2.6) Suppose $x \notin Z$ is such that there is $k \in S^*$ with $u_k(x) < b_k$ and $u_i(x) \geq d$ for all $i \in C$. We now define $S(x)$ to ensure that $f(x) \in X_3^0$.

Since $u_i(x) \geq d$ for all $i \in C$, we must have $R \equiv \{j \in W(x) - C \mid u_j(x) < b_j\} \neq \emptyset$. Otherwise, by (2) of Lemma 1, $u_i(x) = d$ for all $i \in C$ and $u_j(x) = b_j$ for all $j \in W(x) - C$, which means that $x \in \bar{X}$, a contradiction to the hypothesis that $x \notin Z$. Let R' be a minimal subset of R such that $W(x) - R'$ is not winning; i.e., $W(x) - R'$ is not winning but the addition of any single player, j, from R' would make $W(x) - (R' - \{j\})$ a winning coalition. By Assumption 2, this winning coalition cannot be of the form $\{i, j, k\}$ where i is a veto player and j and k are non-veto players. This must mean that $W(x) - R'$ cannot consist of precisely one veto player and one non-veto player. Moreover, since $W(x) - R' \notin \mathcal{W}$, with $S(x) = R'$, $f(x) \in X_3^0$.

(2.7) If x is such that $u_j(x) \geq b_j$ for all $j \in S^*$, since $x \notin Z$, there must be a veto player i for whom $u_i(x) < a$. Let $\{i'\}$ be the lowest indexed player of this kind and $S(x) = \{i'\}$. Since $|S^*| \geq 3$, by Assumption 1 (b), $f(x) \in X_3^0$.

This completes the description of F. Note that in each case that $x \notin Z$, the expectation leads in at most two steps to a state in Z. It is also easy to see that all players in $S(x)$ strictly gain in moving from x to $f^*(x)$. Thus F satisfies (E).

To show that F satisfies (M') it is useful to note that a non-veto player $j \in N - C$ receives either b_j or 0 in Z while a veto player i receives either a or $d > a$. In cases (2.3), (2.4) and (2.5), the move from x is initiated by a coalition of the form $\{i, k\}$ with $i \in C$ and $k \notin C$ leading either directly or in two-steps to a state in X^k. Since neither i nor k prefer any other state in Z, strong maximality holds in all these cases. In case (2.2), $S(x)$ includes all non-veto players and every such j receives b_j at $f^*(x) = f(x)$. Thus, none of them can do better as part of some other objecting coalition, implying that (M') is satisfied. The same argument also applies to case (2.6). In case (2.5), non-veto players have no reason to join any deviating coalition, and so any objection coalition must be a subset of C. Since $|S^*| \geq 3$, Assumption 1 (b) implies that non-veto players can only move to X_3^0 (and then to $\hat{X}$). Thus, the move by $\{i'\}$ is strongly maximal.

Finally, we show that F satisfies (I).

Take any $\bar{x} \in \bar{X}$ and $\hat{x} \in \hat{X}$. Then, all $i \in W(\bar{x})$ weakly prefer $\bar{x}$ to $\hat{x}$. So, if there is a farsighted objection from $\bar{x}$ to $\hat{x}$, and K is the first coalition to move, then $K \subseteq N - W(\bar{x})$. From Assumption 1 (c) it follows that if $K \in E(\bar{x}, x)$, then $u_i(x) \geq u_i(\bar{x})$ for all $i \in W(\bar{x})$. Repeated application of this argument rules out any farsighted objection. Notice that an identical argument ensures that there cannot be a farsighted objection from a state in $\bar{X}$ to another state in $\bar{X}$. Obviously there cannot be a farsighted objection from a state in $\hat{X}$ to another state in $\hat{X}$ since the payoff vector for all states in $\hat{X}$ is unique.

Finally, consider the possibility of a farsighted objection from $\hat{x}$ to $\bar{x}$, where $\hat{x} \in \hat{X}$ and $\bar{x} \in \bar{X}$. All members of $N - C$ weakly prefer $\hat{x}$ to $\bar{x}$. So, the first deviation must come from some subset of C. Since $|S^*| \geq 3$, Assumption 1 (b) implies that this leads to a state x in X_3^0, and $f(x) = \hat{x}$. This establishes that Z satisfies (I) and completes the proof that Z is a SREFS.

If we allowed a deviating coalition to reorganize the coalition structure of its residuals, even in zero states, it would be a lot easier to handle all these cases in which we carefully account for $f(x)$ to be in X_1^0 etc. In fact, we could even do without assumption 2. But if we drop Assumption 1 (b), then internal stability would be a problem.