Maximality in the Farsighted Stable Set[1]

Debraj Ray
New York University and University of Warwick

and

Rajiv Vohra
Brown University

April 2017

Abstract. The vNM stable set of von Neumann and Morgenstern provides a way of imposing credibility on coalitional deviations that is absent from the simpler notion of coalitional stability defined by the core. It is also amenable to the consideration of farsightedness in considering coalitional stability, as proposed by Harsanyi (1974), and more recently re-formulated by Ray and Vohra (2015) in terms of a farsighted stable set. Farsighted dominance considers a sequence of coalitional moves in which each coalition eventually gains at a 'final', stable allocation. When there are multiple continuation paths leading from an initial coalitional move, however, the farsighted stable set does not consider the possibility that coalitions that come later in the sequence of moves, even though they eventually gain, are expected to make moves that are not *maximal*, in the sense that they may do *even better* by moving elsewhere. This becomes problematic if it makes the initiator worse-off, and it could well make the predictions of the farsighted stable set unreasonable. Dutta and Ray (2015) incorporate maximality by relying on rational expectations regarding the move from one state to another. They show that in simple games the resulting prediction is different from the farsighted stable sets identified in Ray and Vohra (2015). But in addition to maximality they also restrict attention to Markovian expectations, where the expected move from a state is independent of the history. What if expectations can be history dependent? We show that allowing for history dependence makes it possible to support the farsighted stable sets as maximal.

Keywords: stable sets, farsightedness, maximality, history dependent expectations.

JEL classification: C71, D72, D74

[1]Ray acknowledges funding from the National Science Foundation under grant SES-1629370.

1. Introduction

The core is a classical solution concept that seeks payoff profiles that no group, or coalition, can dominate with an allocation that is feasible for the coalition in question. It does not ask if the new allocation itself is threatened in some "credible" way by other coalitions. In this respect — at least at a conceptual level — the solution is too strong, possibly excluding other allocations that would not be "credibly" dominated. A central problem is that the definition of credibility is circular, and concepts such as the bargaining set (Aumann and Maschler 1964), which try to build in an additional "round" of domination, are just not up to the task. But the vNM stable set (von Neumann and Morgenstern, 1944) can indeed be seen as such a theory, because it cuts through that circularity. Say that a payoff profile is dominated by another profile if some coalition prefers the latter profile and can unilaterally implement the piece of the new profile that pertains to it. A set of feasible payoff profiles Z is *stable* if it satisfies two properties:

Internal Stability. If $u \in Z$, it is not dominated by $u' \in Z$.

External Stability. If $u \notin Z$, then there exists $u' \in Z$ which dominates u.

Notice how internal and external stability work in tandem to get around the circularity implicit in the definition of credibility. The set Z is to be viewed as a "standard of behavior." Once accepted, no allocation in the standard can be overturned by another allocation also satisfying the standard. Moreover, allocations within the standard jointly dominate all non-standard allocations. This perspective drives home the idea that the relevant solution concept is not *a* payoff profile, but a *set* of payoff profiles which work in unison. It is a beautiful definition.

Yet, and temporarily setting beauty aside, there are at least three problems with the definition:

1. *Harsanyi critique.* Suppose that u' dominates $u \in Z$, and that u' is in turn dominated by $u'' \in Z$, as required by vnM stability. Then it is true that u' isn't "credible," but so what? What if the coalition that proposes u' only does so to induce u'' in the first place, where it *is* better off? Harsanyi went on to propose a "farsighted version" of vNM stability, one that permits a coalition to anticipate a chain reaction of payoff profiles, and asking for a payoff improvement at the terminal node of this chain (our model below will incorporate this feature).

2. *Ray-Vohra critique.* Ray and Vohra (2015) point to a seemingly innocuous device adopted by von Neumann and Morgenstern. Dominance is defined over entire profiles of payoffs. As described above, profile u' dominates u when some coalition is better off under u' and can implement *its piece* of u' unilaterally. But what about the rest of u', which involves allocations of payoffs to others who have nothing to do with the coalition in question? Who allocates these payoffs, and what incentive to do they have to comply with the stipulated amounts? To this, von Neumann and Morgenstern would answer that it does not matter: the heart of the matter is the domination, and the external payoffs are irrelevant, and only a device for tracking all profiles in a common space. But the point is that once modified along the lines of Harsanyi, the critique *does* matter: the payoffs accruing to others will fundamentally affect the chain reaction that follows. Their determination cannot be finessed.

3. *Maximality problem.* Domination requires that some coalition be better off after the dominating allocation is in place, or (in the case of the Harsanyi critique), once the chain reaction of profiles and counter-profiles has come to a standstill. But that casts new light on yet another issue: "better off" isn't the same as "optimality." This is of no concern to the initial coalition that dominates a proposed profile; after all, if it is credibly better off, the original proposal does not stand and that is that. But it *is* of concern to the sequence of beliefs that shores up the chain reaction predicted to follow: that reaction is supported by the anticipation that the relevant coalitions participating in the chain will also be "better off" doing so. But now "better off" isn't good enough: what if they are *even* better off doing something else, and that something else isn't good for the original deviator? Then the beliefs of the deviator, which rely on the moves of future coalitions, become suspect.

This third problem and a (partial) resolution to it forms the subject of this paper.

One reaction to each of the problems described above — and indeed, to all three of them together — is to fully abandon the notion of domination and far-sightedness in a timeless world, and to model each suggested payoff profile as an action in a well-defined game of negotiations, with or without payoffs received *en route*; see, e.g., Konishi and Ray (2001), Gomes and Jehiel (2005) and Ray and Vohra (2014). This alternative viewpoint presupposes a given game form, as in a fully noncooperative model, which includes not just a specification of discounting or shrinking surpluses, but also a well-defined protocol in which Nature chooses an active player or coalition to make a "proposal" to others; see Ray and Vohra (2014) for an extended discussion. Domination, farsightedness and maximality would all be naturally subsumed under a well-defined coalitional game with payoff-maximizing players, with chains of payoff profiles unfolding under some endogenous, commonly anticipated, equilibrium stochastic process.

But, of course, such an approach comes with its own set of problems. The results one obtains are typically protocol-dependent, with the exact game form determining the outcome; see Ray (2008) for a discussion of this point. It is at least worth exploring a looser scenario in which no particular order of player or proposer is cast in stone, where in principle a coalition could come in with a move when it can gain from doing so. Also, the idea is to retain the broad ambience of a negotiation in progress. It is not necessarily that each coalitional move is immediately payoff-relevant; it may simply represent a new conversation or a threat which can be responded to by some other coalition. This is the spirit in which the vNM solution is described, and it is just as reasonable a path to explore.

In a recent paper, Dutta and Vohra (2016) extend the idea of farsighted stability to include maximality. The device they employ is that of an *expectation function*, defined on the current state of the process. The expectation function records the commonly held belief of every player regarding the new state to come (and a coalition that imposes that state) as a function of the current state. Consider the subclass of such functions that *terminate*, in the sense that starting from any state and iteratively applying the expectation function, we eventually arrive at some unchanging state. Think of the union of all such states as the analogue of a stable set.

Dutta and Vohra impose the following requirements:

(a) No stationary state can be profitably disturbed by any coalition, under the presumption that their initial deviation will be followed by iterative application of the expectation function, with a new stationary state being reached. This is analogous to (farsighted) internal stability.

(b) Starting from any non-stationary state, the coalition and move stipulated under the going expectation function indeed *wants* to make that move; that is, after the process has resettled at some stationary state, the coalition in question is better off. This is analogous to (farsighted) external stability.

(c) In addition to (b), no coalition called upon to move at some state strictly prefers to make any *other* move, under the presumption that the expectation function will take over following any such move. This is the new requirement of *maximality.*

The resulting maximal set of Dutta and Vohra therefore incorporates farsightedness, as well as maximality. A notable feature of their analysis is that in the class of simple games — characteristic functions that take the value of 1 for "winning coalitions" and 0 otherwise — their solution differs from that of the farsighted stable set of Ray and Vohra (2015), constructed without a maximality notion in mind.

But Dutta and Vohra impose more than just maximality. Because their expectation function is defined on the domain of *current* states alone, it corresponds to a Markovian notion of coalition formation.[2] It does not answer the question of whether a farsighted stable set in the sense of Ray and Vohra (2015) can indeed be supported as a maximally farsighted stable set, when the expectations function can be taken to be history-dependent.

We should note that history-dependence is no mere technicality. The arrival of the process at a particular state is contextual. How one treats that arrival in a process of negotiation may well depend on just why

[2]A parallel notion in the real-time formulation of Konishi and Ray (2001) is that of a "process of coalition formation," which acts as transition probability on the current state, describing a commonly held probability assessment over future states.

the process got there, much as if one would hesitate, on hearing that a friend has broken a leg, to offer condolences or congratulations. A state with a middling payoff could well be responded to differently if it represents an escape from a bad state, or a deviation from a good state.

The main question we wish to answer is whether, or under what circumstances, it is possible to support a farsighted stable set with history dependent expectations that satisfy the maximality requirement. A result that farsighted stable sets can be supported by such expectations would be extremely useful because in general, constructing expectations that meet the maximality requirement is quite complex. It is simple enough to construct an example of an abstract game in which a farsighted stable set cannot be maximal; see for instance Example 1 in Dutta and Vohra (2016). On the other hand, as we will argue, in many models such as characteristic function games with large cores or simple games, history dependence is often enough to ensure that a farsighted stable set can be supported by maximal expectations.

We already know from Dutta and Vohra (2016), of two classes of games in which maximality holds: (1) characteristic function games that posses a single-payoff farsighted stable set, and (2) constant sun games that possess a certain kind of stable set known as a main simple solution. The former were shown to exist by Ray and Vohra (2015) for games in which the core has a non-empty interior, and the latter were studied extensively by von Neumann and Morgenstern (1944).

This paper shows that the farsighted stable sets identified by Ray and Vohra (2015) for simple games are indeed maximally farsighted stable. We conjecture that this conclusion extends to *all* farsighted stable sets of simple games.[3]

One might argue that the maximality property is achieved for a farsighted stable set simply because of a folk-theorem-like argument: with enough patience (in our case the effective discount factor is indeed 1), *every* set of payoffs can be supported as a stable set. It is easy to see that there is no such folk theorem because the set consisting of all the states is typically not supportable as a (maximal) farsighted stable set. Nevertheless, it will be important to ask what restrictions, if any, are imposed by maximality when history dependence is allowed for.

2. The Farsighted Stable Set: A Quick Review

2.1. **Preliminaries.** A characteristic function game is denoted by (N, V) where $N = \{1, \dots n\}$ is the finite set of players and for each coalition $S \subseteq N$, the set of feasible utility vectors is $V(S) \subseteq R^S$, the S-dimensional Euclidean space with coordinates indexed by the players in S.

A transferable utility (TU) game, denoted by (N, v), is one in which each coalition S has a number (its worth), $v(S)$, such that $V(S) = \{v \in R^S \mid \sum_{i \in S} v_i \leq v(S)\}$.

A payoff vector $v \in V(S)$ is *efficient for* S if $v \in \bar{V}(S) \equiv \{v' \in V(S) \mid \text{there is no } v'' \in V(S) \text{ with } v'' > v'\}$.[4]

A *coalition structure* π is a partition of N into coalitions. A *state* x is a coalition structure and a feasible payoff profile; i.e., $x = (u, \pi)$, with u_S feasible and efficient for S, i.e., $u_S \in \bar{V}(S)$, for each $S \in \pi$. For each pair of states x and y, an *effectivity correspondence*, $E(x, y)$ gives the collection of all coalitions — possibly empty — that can change x to y. State y *dominates* state x *under* E if there exists a coalition $S \in E(x, y)$ with $u(y)_S \gg u(x)_S$. For $A \subseteq X$, let

$$\text{dom}_E(A) = \{x \in X \mid x \text{ is dominated by some } y \in A \text{ under } E\}.$$

The *core* of (N, V, E) is

$$C(N, V, E) = X - \text{dom}_E(X).$$

A set $Z \subseteq X$ is a *vNM stable set* of (N, V, E) if

$$Z = X - \text{dom}_E(Z).$$

[3] There may, however, be other maximal expectations supporting sets – ones identified in Dutta and Vohra (2016) – that are not necessarily farsighted stable sets.

[4] We use the convention $\geq, >\gg$ to order vectors in $\mathbb{R}^N$.

These are standard concepts, except that by using the effectivity correspondence, we are explicit about just which states can be induced from some going state. There are good reasons for doing this, as explained in Ray and Vohra (2015). Briefly, von Neumann and Morgenstern focused on imputations, in which a blocking coalition S can essentially choose *any* allocation for *all* the players as long as the projection of that allocation is feasible for S. That confers an unwarranted degree of power on S, which matters very little when the solution concept — such as the core — is myopic, but matters crucially when the solution concept involves *chains* of deviations, as will be the case here. We limit the power of any coalition by imposing two natural restrictions on the effectivity correspondence:

(i) If $T \in E(x, y)$, $S \in \pi(x)$ and $T \cap S = \emptyset$, then $S \in \pi(y)$ and $u(x)_S = u(y)_S$.

(ii) For every state $x \in X$, $T \subseteq N$ and $v \in \bar{V}(T)$, there is $y \in X$ such that $T \in E(x, y)$, $T \in \pi(y)$ and $u(y)_T = v$.

Condition (i) grants coalitional sovereignty to the untouched coalitions: the formation of T cannot influence the membership of coalitions that are entirely unrelated to T in the original coalition structure, nor can it influence the going payoffs to such coalitions. Condition (ii) grants coalitional sovereignty to the deviating coalition: it can choose not to break up, and it can freely choose its *own* payoff allocation from its feasible set.

2.2. **Farsightedness.** State y *farsightedly dominates* x (under E) if there are states $y^0, y^1, \ldots, y^m$ (with $y^0 = x$ and $y^m = y$) and coalitions, $S^1, \ldots, S^m$, such that for all $k = 1, \ldots m$:

$$S^k \in E(y^{k-1}, y^k) \text{ and } u(y)_{S^k} \gg u(y^{k-1})_{S^k}.$$

For $A \subseteq X$, let

$$\text{dom}_E(A) = \{x \in X \mid x \text{ is farsightedly dominated under } E \text{ by some } y \in A\}.$$

A set of states $F \subseteq X$ is a *farsighted stable set* if

$$F = X - \text{dom}_E(F).$$

2.3. **Farsighted Stable Sets in Two Settings.** This section summarizes some results in Ray and Vohra (2015), which form the basis for our analysis here. In Ray and Vohra (2015), we consider two settings. In the first, we look only for single-payoff stable sets; that is, stable sets with just one payoff profile in them. At first sight, this looks like a demanding task, because the job of ensuring that *all* other payoff profiles are farsightedly dominated is left to just *one* distinguished profile. However, it turns out that this can be accomplished in a wide variety of situations that admit what we call a "separable payoff profile."

A payoff profile u is *efficient* if there does not exist another feasible profile u' with $u' > u$. A collection of pairwise disjoint coalitions $\mathcal{T}$ is a *strict subpartition* of N if $N - \cup_{T \in \mathcal{T}} T$ is nonempty. An efficient profile u is *separable* if whenever $u_T \in V(T)$ for every T in some strict subpartition $\mathcal{T}$, then $u_S \in V(S)$ for some $S \subseteq N - \cup_{T \in \mathcal{T}} T$.

Separability has close (but not exact) links to the core. If u is separable, then u must belong to the coalition structure core of (N, V). Conversely, if u belongs to the *interior* of the coalition structure core of (N, v), it must perforce be separable. (See Ray and Vohra, 2015.)

Given a payoff allocation u, let $[u]$ denote the collection of all *states* that are equivalent to u in terms of payoffs, i.e.,

$$[u] = \{y \in X \mid u(y) = u\}.$$

Theorem 1 (Ray and Vohra, 2015). *Consider* (N, V, E) *such that* E *satisfies Conditions* (i) *and* (ii). *Then* $[u]$ *is a single-payoff farsighted stable set if and only if* u *is separable.*

Theorem 1 is a complete characterization of single-payoff farsighted stable sets.

In the second setting, we examine a large class of games that generally do not possess separable allocations, and therefore (by Theorem 1) no single-payoff farsighted stable sets. These are *simple games*: superadditive

TU games with $v(S) = 1$ or $v(S) = 0$ for every coalition S, and if $v(S) = 1$, then $v(N - S) = 0$. Coalition S is a *winning coalition* if $v(S) = 1$, a *losing coalition* if $v(S) = 0$, and a *veto coalition* if its complement is losing, i.e., if $v(N - S) = 0$.[5] A singleton veto coalition — if it exists — will be referred to as a *veto player*; note that every veto player must belong to every winning coalition. The collection of all veto players, also known as the *collegium*, is denoted $S^* = \cap_{S \in \mathcal{W}} S$, where $\mathcal{W}$ is the set of all winning coalitions. A *collegial game* is one in which $S^* \neq \emptyset$. The collegium (and the corresponding game) will be called *oligarchic* if S^* is itself a winning coalition.

As is well known, simple games play an important role in the analysis of political bargaining and institutions; see, e.g., Baron and Ferejohn (1989), Matthews (1989), Winter (1996), Austen-Smith and Banks (1999), Diermeier and Myerson (1999), Cameron (2000), Tsebelis (2002), Gehlbach and Malesky (2010), Diermeier, Egorov and Sonin (2013), and Nunnari (2014). From a more theoretical perspective, simple games are versatile objects: they may or may not possess empty cores, and they may or may not have separable allocations, and we know the exact conditions under when these situations occur. Specifically, oligarchic games have separable allocations, so that Theorem 1 applies, while in any non-oligarchic games there is no separable allocation, so that a single-payoff farsighted stable set cannot exist.

We will continue to assume that the effectivity correspondence satisfies Conditions (i) and (ii). Under these conditions, states with the same winning coalition and the same payoff allocation are essentially equivalent: the payoffs to winning coalitions are unaffected by changes elsewhere. So to simplify the exposition without sacrificing anything of substance, we describe a state x fully by its winning coalition $W(x)$ (if any), ignoring the specific structure elsewhere, and by the payoff allocation $u(x)$ associated with the state. In particular, all outcomes with no winning coalitions are identified with one another, and collectively referred to as the *zero state*.

We provide two results in this second setting. First:

Theorem 2 (Ray and Vohra, 2015, Theorem 3)**.** *Suppose that the game is oligarchic. Under Conditions* (i) *and* (ii)*, F is a farsighted stable set if and only if $F = [u]$, where $u \in C(N, v)$ and $u_{S^*} \gg 0$.*

In oligarchic games, separable allocations exist: these are, in fact, core allocations with strictly positive payoffs to every member of the collegium. So Theorem 1 applies. But Theorem 2 asserts more: single-payoff sets are the *only* farsighted stable sets.

The more interesting case is that of non-oligarchic games. These have non-empty cores, but no separable allocations. So farsighted stable sets must be multi-payoff. Do they exist, and what structure do they have? To answer this question we restrict our environment a bit more. First, we add a mild additional restriction on the effectivity correspondence:

Condition (iii). Suppose $S = W(x)$ and $T \in E(x, y)$. If $S - T$ is a winning coalition, then $S - T = W(y)$ and $u_i(y) \geq u_i(x)$ for all $i \in W(y)$.

That is, if a deviating coalition leaves behind a residual that is in fact a winning coalition, then no player in the residual gets less than she was getting for (after all, the size of the winning coalition is smaller). A deviant subcoalition cannot dictate terms to its residual, if the residual is still winning even after the deviant breaks away.

The second restriction is on the structure of the simple game. Define a *minimal veto coalition* to be a veto coalition such that no strict subset of it is a veto coalition. A veto coalition M is *non-elitist* if it is minimal and $(M - \{i\}) \cup \{j\}$ is also a veto coalition for every $i \in M$ and every $j \notin M$. That is, such a coalition is built on the principle of numbers. It cannot be smaller (it is minimal), but it can replace every individual member with any outsider and still retain its veto power; hence the term non-elitist.

[5] Obviously, a winning coalition is necessarily a veto coalition.

Our restriction is that there exists a non-elitist veto coalition. This is a weak requirement. It does not require symmetry for the game as a whole, and in particular, the existence of a non-elitist veto coalition implies neither the presence or absence of veto players.[6]

In this setting, the next result shows the existence of farsighted stable sets in which the payoffs are what von Neumann and Morgenstern (1944) referred to as *discriminatory sets*. A discriminatory set takes the form $D(K, \mathbf{a}) = \{u \in \triangle \mid u_i = a_i \text{ for } i \in K\}$, where $a_i \geq 0$ for all $i \in K$. The members of K are the "fixed-payoff" players, each receiving a constant amount. The rest of the surplus is divided in an arbitrary way among the remaining agents: the "bargaining players." It turns out that for any minimal winning coalition S, and with some restrictions on $\mathbf{a}$, the fixed payoff vector, discriminatory vNM stable sets of the form $D(N - S, \mathbf{a})$ exist. In particular, the bargaining players are members of the minimal winning coalition.

Theorem 3 (From Ray and Vohra, 2015, Theorems 4 and 5)**.** *Suppose that the game is non-oligarchic. Assume Conditions* (i)–(iii)*, and suppose moreover that there exists a non-elitist veto coalition. Then a farsighted stable set exists, and for any such set,*

M. Then $[D(N - M, \mathbf{a})] \equiv \{x \in X \mid u(x) \in D(N - M, \mathbf{a})\}$ is a farsighted stable set, where

(a) $a_i > 0$ *for all* $i \in N - M$*.*

(b) $a(S^*) \equiv \sum_{i \in S^*} a_i < 1$.[7]

(c) $a(M) \equiv \sum_{j \in M} a_j > \frac{(m-1)}{m}(1 - a(S^*))$*, where m is the cardinality of M.*

Farsighted stable sets have the special structure described in part (b) of the Theorem. They resemble singletons but *only* for the veto players: their payoff cannot change at all. But the remaining surplus can be divided in any way among the non-veto players.

2.4. **Rational Expectations and Maximality.** Dutta and Ray (2016) build on Jordan's notion of an expectation function to describe the transition from one state to another, keeping track of the coalition that causes such a transition. They 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.

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).

---

[6]For instance, the UN Security Council has a collegium made up of the set of five permanent members. But a decision also needs the affirmative vote of 4 of the 10 non-permanent members. So every collection of members with at least 7 non-permanent members forms a veto coalition, and every collection with exactly 7 non-permanent members is a *minimal veto coalition*. Moreover, these coalitions are all non-elitist.

[7]Use the convention that $a(S^*) = 0$ if $S^* = \emptyset$.

Conditions (I) and (E) are related to farsighted internal and external stability. The main difference is that (I) and (E) only consider farsighted objections that are *consistent* with the common expectation F.

They also consider a *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') strengthens (M) by allowing for the possibility that a coalition T which includes some players from $S(x)$ is allowed to change the transition. This is based on the idea that a move by $S(x)$ requires the unanimous consent of all its members, which means that another coalition may seize 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).

An even stronger notion of maximality would allow *any* coalition, even one disjoint from $S(x)$, to interfere in the expected move. Our main result covers this strongest version of maximality.

The definition of an expectation function $F : X \mapsto X \times \mathcal{N}$ implicitly imposes stationarity or history independence of expectations. A *history* is an ordered collection of states as well as coalitions that lead from one state to the next. To allow for history dependence, letting $\mathcal{H}$ denote the set of all histories we consider an expectation to be a function $F : \mathcal{H} \mapsto X \times \mathcal{N}$. With this modification, internal and external stability as well as maximality can now be defined exactly as in Dutta and Vohra (2016). We refer to the corresponding set of stationary points of a strong rational expectation function F as a *history dependent maximal farsighted stable set*.

3. Maximality of farsighted stable sets in simple games

As Dutta and Vohra (2016) show, the farsighted stable sets of Theorem 3 cannot be maximal if expectations are restricted to be Markovian. Our main result is that maximality is restored by allowing for history dependent expectations. Before presenting our general it will be instructive to illustrate this through three-player example of a simple game with one veto player. In this example every farsighted stable set is a discriminatory set, Z in which the veto player (player 1) receives a fixed amount, strictly between 0 and 1:$Z = \{x \in X \mid u(x) \in D(\{1\}, a)\}$, and has the following justification (for external stability):

1. If x is a zero state go anywhere in the interior of Z (a regular state).

2. If $u_1(x) < a$, player 1 goes to the zero state and then anywhere in the interior of Z.

3. If $u_1(x) > a$, 2 and 3 go to the zero state and then to an interval in Z where both gain relative to x.

It is easy to see that the expectations corresponding to these rules are *not Markovian* because rule 3 may require a different transition from the 0 state depending on the previous state. For this reason it is impossible to support Z through rational expectations in the sense of Dutta and Ray (2016); see the paper for further details.

What we will do now is construct a history-dependent expectation function F that will satisfy the above rules and Z will be its set of stationary points. Moreover, F will be defined for all histories and will satisfy the strongest form of maximality.

Let $\hat{x}$ be the state with $W(x) = N$ and $u(\hat{x}) = (a, 0.5(1-a), 0.5(1-a))$. Given a finite history h, consider x the last state in the history. (We assign 0 payoff to every infinite history).

If $x \in Z$, set $f(x) = x$. Otherwise, define F with the following rules:

1.1. Suppose x is a zero state such that the last non-zero state in the history, if any, was x' with $u_1(x') \leq a$. Then $S(x) = N$ and $f(x) = \hat{x}$.

1.2. Suppose x is a zero state such that the last non-zero state in the history was x' with $u_1(x') = a' > a$. Let $u_2(x') = b$ and $u_3(x') = c$. Then $S(x) = N$ and $f(x) = (a, b + 0.5(a' - a), c + 0.5(a' - a))$.

2.1. Suppose x is a non-zero state with $u_1(x) < a$, then $S(x) = \{1\}$ and $f(x)$ is the resulting zero state.

2.2. Suppose x is a non-zero state with $u_1(x) > a$, then $S(x) = \{2, 3\}$ and $f(x)$ is the resulting zero state.

It is easy to see that F applied to any state, with any history, leads to a state in Z in at most two steps through a farsighted objection.

We are now ready to study the consequences for any coalition that makes an effective move that does not conform with F. Suppose there is a history ending at state $x \notin Z$. Suppose T replaces x with y, where $T \in E(x, y)$, and either $T \neq S(x)$ or $y \neq f(x)$. Is it possible for T to gain through such an interference with F by ending up at some state in Z? In other words, is it possible that $u_T(f^*(y)) \gg u_T(f^*(x))$?

First note that T cannot include 1 because $u_1(f^*(z)) = a$ for all z. It cannot include 2 and 3 together because they can't both gain by going elsewhere in Z. Thus $T = \{2\}$ or $T = \{3\}$. The two cases are symmetric so there is no loss of generality in assuming that $T = \{2\}$.

Suppose $T = \{2\}$ moves from x to y where $T \in E(x, y)$. We will show that $u_2(f^*(y)) \leq u_2(f^*(x))$. There are two cases to consider.

Case (a). Suppose y is a zero state. Then, according to 1.1 and 1.2, the final outcome will only depend on the history prior to x and $f^*(x) = f^*(y)$. So this cannot allow T to improve upon the prescription according to F.

Case (b). Suppose y is not a zero state. This must mean that $W(y) = \{1, 3\}$ with $u_2(y) = 0$. If $y \in Z$ this cannot be a profitable deviation. Suppose therefore that $y \notin Z$. By monotonicity, $u_1(y) \geq u_1(x)$ and $u_3(y) \geq u_3(x)$. If $u_1(y) < a$, then $u_1(x) < a$ and $f^*(y) = f^*(x) = \hat{x}$, so this cannot be a profitable deviation by T. The only other possibility is that $u_1(y) > a$, in which case $f^*(y)$ splits the surplus equally between 2 and 3:

$$u_3(f^*(y)) = u_3(y) + 0.5(u_1(y) - a) \geq u_3(x) + 0.5(u_1(x) - a). \tag{3.1}$$

If $u_1(x) > a$, then the last expression is $u_3(f^*(x))$, and $u_3(f^*(y)) \geq u_3(f^*(x))$. Since $u_1(f^*(x)) = u_1(f^*(y)) = a$, this implies that $u_2(f^*(y)) \leq u_3(f^*(x))$, as required. If $u_1(x) < a$, then $u_2(f^*(x)) = u_2(\hat{x}) = 0.5(1 - a)$. Since $u_1(y) = 1 - u_3(y)$, it also follows that

$$u_3(f^*(y)) = u_3(y) + 0.5(u_1(y) - a) = 0.5u_3(y) + 0.5(1 - a) \geq 0.5(1 - a), \tag{3.2}$$

and again we have $u_3(f^*(y)) \geq u_3(f^*(x))$, which implies that $u_2(f^*(y)) \leq u_3(f^*(x))$.

These ideas generalize to the class of all simple games satisfying the assumptions of Theorem 3.

Theorem 4. *Consider a farsighted stable set Z identified in Theorem 3, i.e., $Z = [D(N - M, \mathbf{a})] \equiv \{x \in X \mid u(x) \in D(N - M, \mathbf{a})\}$, satisfying properties (a), (b) and (c). There exists a history dependent expectation F that is strong rational and supports Z as a history dependent maximal farsighted stable set.*

We begin by constructing an expectation function F. Let x^0 denote the zero state and partition the set of non-zero states in the complement of Z into the following three sets:

$X^1 = \{x \in X - Z - \{x^0\} \mid u_i(x) < a_i \text{ for some } i \in S^*\}$.

$X^2 = \{x \in X - Z - \{x^0\} - X^1 \mid \sum_{j \in M} u_j(x) < a(M)\}$.

$X^3 = \{x \in X - Z - \{x^0\} - X^1 - X^2\}$.

Note that if $x \in X^3$, then $u_q(x) < a_q$ for some $q \in N - M - S^*$. For $x \in X^3$, we will denote by $r(x)$ the lowest indexed player in M such that $u_{r(x)}(x) \geq u_j(x)$ for all $j \in M$.

The argument for external stability is based on the following features:

1. If x is a zero state, then there is an objection by N from a state in Z.

2. If $x \in X^1$, then there is a farsighted objection starting with a departure by player $i \in S^*$ with $u_i(x) < a_i$, which precipitates the zero state, and from there N moves to some state in Z.

3. If $x \in X^2$, then $W(x) \cap M$ can construct a farsighted objection by leaving $W(x)$, forcing the zero state, followed by a state in Z where all players in $W(x) \cap M$ gain.

4. If $x \in X^3$, a farsighted objection can be constructed by forming a coalition $M - \{r(x)\} \cup \{q\}$. This coalition first moves to the zero state, followed by a move by N to a state in Z in which all members of this coalition gain relative to x.

We will now construct an expectation function F that will satisfy the above rules and Z will be its set of stationary points. Moreover, F will be defined for all histories and will satisfy the strongest form of maximality.

Let $\hat{x}$ be the state with $W(x) = N$, $u_i(\hat{x}) = a_i$ for all $i \in N - M$ and $u_j(\hat{x}) = \frac{a(M)}{m}$ for all $j \in M$.

Given a finite history h, consider x the last state in the history. (We will probably have to assign 0 payoff to every infinite history).

If $x \in Z$, set $f(x) = x$. For $x \notin Z$, define F through the following rules:

1.1. Suppose $x = x^0$ and the last non-zero state in the history, if any, was $x' \in X^1$. Then $S(x) = N$ and $f(x) = \hat{x}$.

1.2. Suppose $x = x^0$ and the last non-zero state in the history was $x' \in X^2$. Then $S(x) = N$ and $f(x) \in Z$ with $u_j(f(x) = u_j(x') + \frac{a(M) - \sum_{j \in M} u_j(x')}{m}$ for all $j \in M$.

1.3. Suppose $x = x^0$ and the last non-zero state in the history was $x' \in X^3$. Let $g = \frac{G}{m}$. Since $x' \in X^3$, $u_i(x') \geq a_i$ for all $i \in S^*$, $\sum_{j \in M} u_j(x') \geq a(M)$ and there is $q \in N - S^* - M$ with $u_q(x') < a_q$. In this case, $S(x) = N$ and $f(x) \in Z$ with

$$\begin{aligned}
u_j(f(x)) &= u_j(x') + g \qquad \text{for } j \in M, j \neq r(x') \\
u_{r(x')}(f(x)) &= a(M) - \textstyle\sum_{j \in M, j \neq r(x')} u_j(f(x)) \\
&= G + \tfrac{(m-1)}{m}(1 - a(S^*)) - \textstyle\sum_{j \in M, j \neq r(x')} u_j(x') - \tfrac{(m-1)}{m} g \\
&= g + [\tfrac{(m-1)}{m}(1 - a(S^*)) - \textstyle\sum_{j \in M, j \neq r(x')} u_j(x')].
\end{aligned}$$

Since $u_i(x') \geq a_i$ for all $i \in S^*$, $\frac{(m-1)}{m}(1 - a(S^*) \geq \sum_{j \in M, j \neq r(x')} u_j(x')$. Thus, $u_{r(x')}(f(x)) \geq g$.

2.1. Suppose $x \in X^1$. Then $S(x) = \{i\}$ and $f(x)$ is the zero state. By 1.1, $f(f(x)) = \hat{x}$.

2.2. Suppose $x \in X^2$. Then $S(x) = M$ and $f(x)$ is the zero state. By 1.2, all players in M gain at $f(f(x))$ compared to x.

2.3. Suppose $x \in X^3$. In this case there is $q \in N - S^* - M$ with $u_q(x) < a_q$. Let $S(x) = M - \{r(x)\} \cup \{q'\}$, where q' is the lowest indexed player in $N - S^* - M$ such that $u_{q'}(x) < a_{q'}$, and $f(x)$ is the zero state. By 1.3, all players in $S(x)$ gain in moving from x to $f(f(x))$.

Note that F has been specified such that for every state x, either $F(x)$ does not depend on the history or it depends on the non-zero state that immediately preceded x. Moreover, it leads to Z in at most two steps: for every x, either $f(x) \in Z$ or $f(f(x)) \in Z$. This means that $f^*(x)$, the 'final outcome' from any state x, is $f(f(x))$.

We are now ready to show that it is not possible for any coalition to gain by making a move that does not conform with F. Since Z satisfies farsighted internal stability it follows that no coalition can profit by moving from a state in Z. We therefore only need to check this property for states that are not in Z. To prove Theorem 4 it suffices to prove the following Proposition for the F that we have constructed.

Proposition 1. *Suppose $x \notin Z$ and $T \in E(x, y)$. Then*

(3.3) $$u_j(f^*(y)) \leq u_j(f^*(x)) \text{ for some } j \in T.$$

Proof. Suppose the Proposition is false, i.e., there exists $x \notin Z$, $T \in E(x,y)$ such that

$$u_T(f^*(y)) \gg u_T(f^*(x)).$$

Claim 1. $T \subset M$, $y \neq f(x)$, $x \neq x^0$, $y \neq x^0$.

Proof of Claim 1. If T is to gain by moving from x to y it must be the case that $y \neq f(x)$, for otherwise $f^*(x) = f^*(y)$. Moreover, it cannot include any player in $N - M$ since these players receive a fixed amount at every state in Z. This means that $T \subseteq M$. In fact, it cannot be M because the aggregate payoff to M is constant across all states in Z. Thus, for T to gain it must be the case that $y \neq f(x)$ and T is a strict subset of M.

If $y = x^0$, $f^*(y)$ depends only on x or the history leading up to x, and $f^*((y)) = f^*(x))$, which contradicts the hypothesis. Note that if $x = x^0$, then the fact that $T \subset M$ implies that $y = x^0$, which we have just ruled out. This completes the proof of Claim 1.

Claim 2. $y \notin Z$ and $y \notin X^1$.

Proof of Claim 2. Since $y \neq x^0$, there is a winning coalition at y, $W(y) = W(x) - T$. Of course, $u_T(y) = 0$. If $y \in Z$, then $u_T(f^*(y)) = 0$ and (3.3) holds. If $y \notin Z$, by the monotonicity assumption,

$$u_i(y) \geq u_i(x) \text{ for all } i \in W(y). \tag{3.4}$$

If $u_i(y) < a$ for any $i \in S^*$, this implies that $u_i(x) < a$, i.e., $y \in X^1$. But then we know from 2.1 that $f^*(x) = f^*(y) = \hat{x}$, and (3.3) is satisfied. Thus, we have $u_i(y) \geq a_i$ for all $i \in S^*$, i.e., $y \notin X^1$.

It will be useful to record a couple of other implications of (3.4). First, note that if $i \notin W(y) \cup T$, then $i \notin W(x) \cup T$ and $u_i(x) = u_i(y) = 0$. This, along with (3.4) implies that

$$u_{N-T}(y) \geq u_{N-T}(x), \tag{3.5}$$

which also means that

$$\sum_{j \in M} u_j(y) \leq \sum_{j \in M} u_j(x). \tag{3.6}$$

To complete the proof we will rule out the possibility that y belongs to X^2 or X^3. In doing so we will need to consider several subcases depending on whether $x \in X^1, X^2$ or X^3.

Claim 3. $y \notin X^2$.

Proof of Claim 3. Suppose not, i.e., $y \in X^2$. It follows from 2.2 and 1.2 that:

$$u_i(f^*(y)) = u_i(y) + \frac{a(M) - \sum_{j \in M} u_j(y)}{m} \text{ for all } i \in M - T. \tag{3.7}$$

and, since $u_j(y) = 0$ for all $j \in T$,

$$u_j(f^*(y)) = \frac{a(M) - \sum_{j \in M} u_j(y)}{m} \text{ for all } j \in T. \tag{3.8}$$

To prove the Claim we consider the following cases:

Case (a): $x \in X^1$. Then $f^*(x) = \hat{x}$ and $u_j(f^*(x)) = \frac{a(M)}{m}$ for all $j \in T$. But then (3.8) implies $u_T(f^*(y)) \leq u_T(f^*(x))$, a contradiction to the hypothesis that $u_T(f^*(y)) \gg u_T(f^*(x))$.

Case (b): $x \in X^2$. From 2.2 and 1.2,

$$u_i(f^*(x)) = u_i(x) + \frac{a(M) - \sum_{j \in M} u_j(x)}{m} \text{ for all } i \in M - T.$$

From (3.5), (3.6) and (3.7), this means that $u_i(f^*(y)) \geq u_i(f^*(x))$ for all $i \in M - T$. Since $\sum_{j \in M} u_j(f^*(y)) = \sum_{j \in M} u_j(f^*(x)) = a(M)$, this contradicts the hypothesis that $u_T(f^*(y)) \gg u_T(f^*(x))$.

Case (c.1): $x \in X^3$ and $r(x) \in M - T$.

We know from 2.3 and 1.3 that

$$u_i(f^*(x)) = u_i(x) + g \text{ for } i \in M - \{r(x)\}$$

and

$$u_{r(x)}(f^*(x)) = g + \frac{(m-1)}{m}(1 - a^*(S)) - \sum_{j \neq r(x)} u_j(x) = g + a(M) - G - \sum_{j \neq r(x)} u_j(x).$$

which implies that

$$\begin{aligned} \sum_{i \in M-T} u_i(f^*(x)) &= \sum_{i \in M-T} u_i(x) + (m-t)g + a(M) - G - \sum_{j \in M} u_j(x) \\ &\leq \sum_{i \in M-T} u_i(x) \end{aligned}$$

where the last inequality follows from the fact that $a(M) \leq \sum_{j \in M} u_j(x))$ and $(m-t)g < G$. From (3.7) we know that $\sum_{i \in M-T} u_i(f^*(y)) \geq \sum_{i \in M-T} u_i(y)$. It now follows from (3.5) that $\sum_{i \in M-T} u_i(f^*(y)) \geq \sum_{i \in M-T} u_i(f^*(x))$. As in case (b), this is a contradiction.

Case (c.2) : $x \in X^3$ and $r(x) \in T$.

Since $u_{r(x)}(f^*(y)) = \frac{a(M) - \sum_{j \in M} u_j(y)}{m}$, $u_{r(x)}(f^*(x)) \geq g$, and by hypothesis $u_T(f^*(y)) \gg u_T(f^*(x))$, we have:

$$\frac{a(M) - \sum_{j \in M} u_j(y)}{m} > g.$$

For $i \in M - T$,

$$u_i(f^*(y)) = u_i(y) + \frac{a(M) - \sum_{j \in M} u_j(y)}{m}$$

while

$$u_i(f^*(x)) = u_i(x) + g.$$

Since, by (3.5), $u_i(y) \geq u_i(x)$ for all $i \in M - T$, this means that $u_i(f^*(y)) > u_i(f^*(x))$ for all $i \in M - T$, which, as in Case (b), contradicts $u_T(f^*(y)) \gg u_T(f^*(x))$.

Claim 4. $y \notin X^3$.

Suppose not. Then, $\sum_{j \in M} u_j(y) \geq a(M)$ and from (3.6) it must also be the case that $\sum_{j \in M} u_j(x) \geq a(M)$, which means that $x \in X^1$ or X^3. Suppose $x \in X^1$. Then for every $j \in T$, $u_j(f^*(x)) = \frac{a(M)}{m}$ and, because $u_T(y) = 0$, $u_j(f^*(y)) = g$. Since $a(M) > G$, $u_j(f^*(y)) < u_j(f^*(x))$ for all $j \in T$, a contradiction to the hypothesis that $u_T(f^*(y)) \gg u_T(f^*(x))$. We can therefore assume that $x \in X^3$. This implies that for every $j \in T$, $u_j(f^*(x)) \geq g$ while $u_j(f^*(y)) = g$, a contradiction.

■

References

Aumann, J. and M. Maschler (1964), "The bargaining set for cooperative games", in Advances in Game Theory (M. Dresher, L. S. Shapley, and A. W. Tucker, Eds.), Annals of Mathematical Studies No. 52, Princeton Univ. Press, Princeton, NJ.

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

Diamantoudi, Effrosyni and Licun Xue (2003), "Farsighted Stability in Hedonic Games," *Social Choice and Welfare*, **21**, 39–61.

Dutta, B. and R. Vohra (2016), "Rational Expectations and Farsighted Stability", forthcoming, *Theoretical Economics.*

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

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

Kimya, Mert (2015),"Equilibrium Coalitional Behavior", mimeo, Brown University.

Konishi, Hideo and Debraj Ray (2003), "Coalition Formation as a Dynamic Process," *Journal of Economic Theory*, **110**, 1–41.

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

Mauleon, Ana, Vincent Vannetelbosch and Wouter Vergote (2011), "von Neumann-Morgenstern farsighted stable sets in two-sided matching," *Theoretical Economics*, **6**, 499–521.

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

Ray, Debraj and Rajiv 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. Peyton Young and Shmuel Zamir, 239–326. North Holland: Elsevier.

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

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

Xue, Licun (1998), "Coalitional Stability under Perfect Foresight," *Economic Theory*, **11**, 603–627.