International Journal of Game Theory (1995) 24: 37–48

纳什均衡的说明

Two-Person Game Forms Guaranteeing the Stability Against Commitment and Delaying Tactics[1]

Nikolai S. Kukushkin

Computing Center, Russian Academy of Sciences, 40, Vavilova, Moscow 117967 Russian Federation

Summary: The notions of the struggle for leadership or for followership, introduced by H. Moulin for two-person games, are considered for game forms. Necessary and sufficient conditions for a game form never to generate a game with the struggle of either kind are derived. Connections between these properties and the existence of a Nash equilibrium for any preference profile as well as the possibility to select a Nash equilibrium in an incentive compatible way are established.

1 Introduction

Like many other mathematical tools, game theory can be interpreted in real-world terms in various ways. Not the least important among them is to perceive it as a framework for distinguishing, discussing, and occasionally even solving problems facing several agents trying to coordinate their actions in the absence of any external regulating force. From this viewpoint, different equilibrium definitions express conditions for the "viability" of a non-binding agreement betweent the agents.

On the other hand, to reach any kind of agreement, the players have, first of all, to discuss their situation peacefully. As the circumstances may stimulate some players to avoid any talks with the partners, or to use threats or even preemptive actions instead of talks, all the equilibrium concepts may prove irrelevant. Even when the signing of a binding agreement is feasible, the opportunity may be missed for those reasons.

Although this aspect of the problem was recognized and discussed on the early stages of the development of game theory – see Luce and Raiffa (1957), Schelling (1960) – the progress in its mathematical treatment has been very limited. This regrettable fact can only partially be attributed to neglect by researchers: these matters are really hard to formalize.

The situation simplifies considerably when there are just two players, as each player then has one partner instead of several. Indeed, Moulin (1981) quite convincingly formalized Schelling's concepts of the struggle for leadership or followership in two-person games. The exact definitions are given in Section 3 below.

[1] I thank V. Gurvich for a fruitful discussion, H. Moulin for his comment on a draft version of the paper and several helpful suggestions, and an anonymous referee, whose recommendations helped me to enhance the quality of presentation. I have benefited from the hospitality of Universidad de Alicante and Instituto Valenciano de Investigaciones Económicas. Some financial assistance from the Cultural Initiative Foundation and Academy of Natural Sciences is also acknowledged.

0020–7276/95/1/37–48 $ 2.50

The struggle of either kind poses an obstacle to the very process of seeking for a compromise. The struggle for followership can be, in a sense, resolved by extending the strategy sets, see e.g. Moulin (1976), Kukushkin (1991); for the struggle for leadership no such simple solution seems possible. Anyway, it is essential to understand what features of a game promote or suppress the struggle for leadership or followership.

Here the problem is studied on the highest level of generality: in Section 4 are desribed game forms which produce no struggle of either kind for any preferences of the players over the set of outcomes. Interestingly, the absence of the struggle for followership for every preference profile is equivalent to the existence of a Nash equilibrium for every preference profile.

Another complication may arise from incomplete information: if a player is the sole possessor of a piece of information (e.g. about his own preferences), he may see it in his interest to deceive the partner; the mere possibility of the occurrence may cause the latter to refuse to take into account any statement by such a player, and the talks will be deprived of any substance before having started. This problem might be avoided if the players were able to reach a contingent agreement, based on their common prior information, on what to choose for any list of messages about the players' private information; for such an agreement to be self-enforcing, certain incentive compatibility conditions must be satisfied. Non-manipulable Nash equilibrium selection rules were first studied by d'Aspremont and Gérard-Varet (1980), who found a connection between their existence and the absence of the struggle for leadership (however, their interpretation of the latter was rather different from Moulin's).

In Section 5 the incentive compatibility problem is studied under the assumption that the game form is common knowledge while every player's utility is his private information; the results are distinctly different from those of d'Aspremont and Gérard-Varet (1980). Interestingly, the absence of the struggle of both kinds (in Moulin's sense) implies the possibility to select an equilibrium (for every preference profile) in an incentive compatible way.

To sum up, the current results give a complete description of game forms guaranteeing the presence of the above-discussed properties under the assumption of an unrestricted preference domain. Unfortunately, they may be of little or no help when there are grounds to restrict the attention to a more particular game class. From a more technical viewpoint, the paper introduces several apparently new classes of game forms deserving some attention for their strategic properties; see Theorems 1, 3, 4, 5, Examples 4 and 5.

2 Basic Definitions

A (two-person) game form G is defined by a set of feasible outcomes A, two strategy sets X_1 and X_2, and a mapping π from $X = X_1 \times X_2$ onto A. In the following we restrict ourselves to finite (non-singleton) sets X_1, X_2, and A, though the main results can be easily extended to the compact-continuous case.

Given a game form G and a pair of utility functions $u_1(a)$, $u_2(a)$ over A, the derivative normal form game $\Gamma(G, u_1, u_2)$ is defined by the strategy sets X_1, X_2, and utility functions $U_i(x) = u_i(\pi(x))$ over X. (The description of preferences with (pre)orderings may seem more general but in fact, as is well known, there is no real difference).

Throughout the whole paper, we will use the convention: $i, j \in \{1, 2\}$, $i \neq j$. For $x_i \in X_i$, the trace $[x_i]$ of the strategy x_i is the set of the outcomes possible when player i uses x_i : $[x_i] = \pi(\{x_i\} \times X_j)$.

A game form G is called tight if for every $i \in \{1, 2\}$, $B \subseteq A$ there exists either $x_i \in X_i$ for which $[x_i] \subseteq B$, or $x_j \in X_j$ for which $[x_j] \cap B = \emptyset$. In the terminology of Moulin and Peleg (1982), G is tight if its (α-)effectivity function is maximal. Gurvich (1988) has shown that the tightness of a game form is necessary and sufficient for the existence of a Nash equilibrium in every derivative game.

A game form G is called dictatorial if there exists a player i such that for every $a \in A$ there exists a strategy $x_i \in X_i$ for which $[x_i] = \{a\}$.

As the opposite case, a game form is called anonymous if $X_1 = X_2$ and π is symmetric.

3 Moulin's Taxonomy

For a given normal form game, define

$$R_i(x_j) = \operatorname*{Argmax}_{x_i \in X_i} U_i(x), \quad \beta_i = \min_{x_j \in X_j} \max_{x_i \in X_i} U_i(x),$$

$$\gamma_i = \sup_{x_i \in X_i} \min_{x_j \in R_j(x_i)} U_i(x), \quad S_i = \max_{x_i \in X_i} \max_{x_j \in R_j(x_i)} U_i(x).$$

(In the compact-continuous case, sup in the definition of γ_i cannot be replaced with max). A strategy bundle $x^{(i)}$ is called a Stackelberg solution for player i if $U_i(x^{(i)}) = S_i$ and $x_j^{(i)} \in R_j(x_i^{(i)})$; generally speaking, it need not be unique.

The underlying interpretation is straightforward: if player j moves after the partner, knowing his decision, he can choose $x_j \in R_j(x_i)$ and be sure that his utility level is not less than β_j; the player i moving first can expect the partner to reply with $x_j \in R_j(x_i)$, so he may evaluate his expected utility level as γ_i or S_i, depending on the degree of optimism (when the preferences are strict, both evaluations coincide).

Following Moulin (1981), we say that there is the struggle for followership in a game if the inequalities

$$U_1(x) \geq \beta_1, \quad U_2(x) \geq \beta_2 \tag{1}$$

are incompatible. This means that claims by both players to the utility levels β_i cannot be satisfied simultaneously (even if signing a binding contract is feasible), so he who manages to put himself into the position of the follower "wins".

As to the struggle for leadership, there are two possible definitions: either to see the struggle when the inequalities

$$U_1(x) \geq S_1,\ U_2(x) \geq S_2 \tag{2}$$

are incompatible (Stackelberg-style definition), or when so are the inequalities

$$U_1(x) \geq \gamma_1,\ U_2(x) \geq \gamma_2 \tag{3}$$

(Germeier-style one). In either case, the interpretation is quite similar to that above. Besides, if every player assumes that his deviation x_i' from a non-binding agreement would result in the partner's choosing $x_j' \in R_j(x_i')$, then the set of potential self-enforcing agreements is described by System (3) – or (2), again depending on the degree of optimism.

Although it is easy to produce an example of a game with the struggle for leadership in the Stackelberg sense and without it in the Germeier sense (as $S_i \geq \gamma_i$, the converse is impossible), Theorem 1 below shows that from our current viewpoint the difference is unessential, so we will keep in mind both definitions simultaneously.

Classification Theorem: (Moulin, 1981) For any normal form game at least one of Systems (1) and (2) is compatible. If each of the systems is compatible, they are compatible together. Both statements are also true w.r.t. Systems (1) and (3).

There is a connection between Moulin's notions and the question of whether it would be more profitable for either player to be the leader or the follower, see Basar (1973), Gal-Or (1985). However, the similarity should not be overestimated. If there is the struggle for leadership (followership) in a game, then each player would be better of as the leader (the follower). On the other hand, if each of Systems (1) and (2) (or (3)) is compatible, then, according to Classification Theorem, the players have a range of potential mutually advantageous agreements, but all Stackelberg solutions may lie outside the range.

The following three statements are almost obvious: there is no struggle for leadership in any antagonistic game, there is no struggle for followership in a game having a Nash equilibrium, and for an antagonistic game the existence of a saddle-point and the absence of the struggle for followership are equivalent properties. There seems to be no other general statement about either property.

4 SLF and SFF Game Forms

A game form G is called struggle-for-followership-free (SFF) if for every pair of utility functions u_1, u_2 over A, System (1) for the game $\Gamma(G, u_1, u_2)$ is compatible.

A game form G is called struggle-for-leadership-free in the Stackelberg sense (SLF-S) if System (2) is compatible for every derivative game $\Gamma(G, u_1, u_2)$.

A game form G is called struggle-for-leadership-free in the Germeier sense (SLF-G) if System (3) is compatible for every derivative game.

A game form G is said to satisfy the embedded traces condition (ET) if for every $x_1 \in X_1$ and $x_2 \in X_2$ either $[x_1] \subseteq [x_2]$ or vice versa.

Theorem 1: For any game form G the following three statements are equivalent:

(a) G is SLF-S;
(b) G is SLF-G;
(c) G satisfies ET.

Suppose (c) is satisfied and $x^{(i)}$ is a Stackelberg solution for player i in a derivative game. By the ET condition, we may suppose $[x_1^{(1)}] \subseteq [x_2^{(2)}]$, hence $U_1(x^{(2)}) \geq U_1(x^{(1)})$; therefore, $x^{(2)}$ satisfies System (2).

Suppose (c) is violated; then there exist $x_1 \in X_1$, $x_2 \in X_2$, $a, b \in A$ such that $a \in [x_1] \setminus [x_2]$ and $b \in [x_2] \setminus [x_1]$. Define u_1 over A so that $u_1(a) = 1, u_1(b) = 0, u_1(c) < 0$ for all $c \in A \setminus \{a, b\}$, and u_2 so that $u_2(b) = 1, u_2(a) = 0, u_2(c) < 0$ for all $c \in A \setminus \{a, b\}$. Evidently, $\gamma_1 = u_1(a), \gamma_2 = u_2(b)$; so System (3) is incompatible.

Examples: Consider three anonymous 3×3 game forms with three outcomes.

1. $\begin{matrix} b & a & c \\ a & a & a \\ c & a & b \end{matrix}$ 2. $\begin{matrix} b & a & b \\ a & a & a \\ b & a & c \end{matrix}$ 3. $\begin{matrix} b & a & a \\ a & a & a \\ a & a & c \end{matrix}$

In the first two cases the ET condition is satisfied, while in the third one it is not. Note that the effectivity function is the same in all the three examples.

The ET condition is obviously satisfied when $|A| = 2$ or when $[x_i] = A$ for all strategies of one of the players; in particular, it is satisfied for every dictatorial game form. Moreover, one could choose the sets of possible traces for each player, satisfying ET (and the necessary requirement of covering A), in an arbitrary way, and then construct a game form with those traces, repeating the same trace several times when needed. I will not go into details just now; a formal proof for a particular case is given in Theorem 4 below.

Theorem 2: A game form is SFF if and only if it is tight.

In essence, the theorem is proved by Gurvich (1988): For a tight game form, every derivative game has a Nash equilibrium, so there is no struggle for followership. On the other hand, the necessity of tightness in Gurvich (1988) is proved with an antagonistic example (in fact, due to Moulin (1976)), so it simultaneously proves the necessity of tightness for SFF.

As the Gurvich theorem may not be too widely known, I will prove the sufficiency part in detail. For any derivative game, denote $B = \bigcup_{x_2 \in X_2} \pi(R_1(x_2), x_2)$. Now $[x_2] \cap B \neq \emptyset$ for any x_2; as G is tight, there must exist $x_1^\circ \in X_1$ for which $[x_1^\circ] \subseteq B$. Pick $x_2^\circ \in R_2(x_1^\circ)$; obviously, x° satisfies (1).

Generally speaking, x° need not be a Nash equilibrium; however, it is one when both x_1° and x_2° are unique: Supposing the contrary, define $a = \pi(x^\circ), B' = B \setminus \{a\}$; now $[x_2] \cap B' \neq \emptyset$ for each x_2 (if $a \in \pi(R_1(x_2), x_2)$, then $\pi(x_1^\circ, x_2) \in B'$), so there must

be x_1^* for which $[x_1^*] \subseteq B' \subset B$, and this contradicts the uniqueness of x_1°. For the case when x_1° is not unique (the uniqueness of x_2° is not so important), Gurvich suggests a criterion for choosing the right x_1° ("lexicographic maximin"), heavily dependent on the finiteness assumption. It is still unclear whether his theorem is at all valid in the compact-continuous case.

There seems to be no compact description of the game forms that are both SLF and SFF. However, every such form has the following property.

Call a game form G partially dictatorial if there is a subset $D \subseteq A$ and a player i such that

(i) for each $a \in D$ there exists a strategy x_i with $[x_i] = \{a\}$,
(ii) there exists a strategy x_j with $[x_j] = D$.

In other words, in a partially dictatorial game form one player is entitled to choose any outcome from a fixed subset of A, while his partner is entitled to demand that he choose one of them. A dictatorial game form is obviously partially dictatorial with $D = A$. As the other extreme case, a game form is partially dictatorial if it contains a "cross" (i.e. a pair of strategies satisfying $[x_1] = \{a\} = [x_2]$), in which case each of the players could be dubbed partial dictator – when $|D| > 1$, the dictator is unique.

Theorem 3: If a game form G is SLF and SFF, then it is partially dictatorial.

For any $a \in A$, as G is tight, there exists either x_1 for which $[x_1] = \{a\}$, or x_2 for which $[x_2] \subseteq A \setminus \{a\}$; by the same reasoning, there exists either x_2 for which $[x_2] = \{a\}$, or x_1 for which $[x_1] \subseteq A \setminus \{a\}$. So one of the following statements must be true:

1) $\exists x_1, x_2$ such that $[x_1] = \{a\} = [x_2]$;
2) $\exists i, x_i', x_i''$ such that $[x_i'] = \{a\}, [x_i''] \subseteq A \setminus \{a\}$;
3) $\exists x_1, x_2$ such that $[x_1] \subseteq A \setminus \{a\}, [x_2] \subseteq A \setminus \{a\}$.

Obviously, any two of the statements cannot be simultaneously true w.r.t. any $a \in A$, so A is partitioned into three subsets: A_1, A_2, A_3.

If $A_1 \neq \emptyset$, then G contains a cross and the theorem is proved. Supposing $A_1 = \emptyset$ and applying the tightness condition (in both ways) to the partitioning of A into A_2 and A_3, we see that either there exist x_1, x_2 such that $[x_1] \subseteq A_2$, $[x_2] \subseteq A_2$, or (at least) one of the players has a strategy x_i with $[x_i] \subseteq A_3$.

In the first case, there must be $|A_2| > 1$ ($A_2 = \{a\}$ would imply $a \in A_1$), so the "player i" in Statement 2) is the same for all $a \in A_2$, and player j has a strategy x_j with $[x_j] = A_2$. Thus G is partially dictatorial with $D = A_2$.

In the second case, pick a strategy x_i with a minimal (w.r.t. inclusion) trace among all strategies of both players satisfying $[x_i] \subseteq A_3$, and pick $a \in [x_i]$. As $a \in A_3$, there exists x_j for which $a \notin [x_j]$, hence, by ET, $[x_j] \subset [x_i]$, which contradicts the presumption on $[x_i]$. The theorem is proved.

Example 3 above shows that Theorem 3 cannot be reversed.

We will call a game form perfect (just for want of a better term) if it is anonymous, SLF, and SFF; Examples 1 and 2 present such forms.

Example 4: Let $a^\circ \in A$ be fixed. Define a game form G as follows: $X_1 = X_2 = A \times \{0, 1, 2, 3\}$, with the generic element $x_i = (a_i, \theta_i)$,

$$\pi(x) = \begin{cases} a^\circ, & \text{if } \theta_1 = 0 \text{ or } \theta_2 = 0, \\ a^\circ, & \text{if } \theta_1 = \theta_2, \\ a_i, & \text{if } \theta_i = \theta_j + 1 \geq 2 \text{ or } \theta_i = 1, \theta_j = 3. \end{cases}$$

It is easily checked that $[x_i] = \{a^\circ\}$ if $\theta_i = 0$ and $[x_i] = A$ otherwise; so the game form is perfect. Note that A need not be finite.

For a finite set A, the perfect game forms allow of something like a complete description (to be more precise, the sets of traces of perfect game forms can be described).

A sequence $T_0 \subset T_1 \subset \ldots \subset T_m$ of subsets of A is called regular if T_0 is a singleton and $T_m = A$.

Theorem 4: An anonymous game form is perfect if and only if the traces of all the strategies of each player form a regular sequence (necessarily the same for each player). Moreover, for any regular sequence $T_0 \subset T_1 \subset \ldots \subset T_m$, there exists a perfect game form in which the traces of the strategies of each player form the sequence.

The first statement follows easily from Theorems 1–3. (The equality $T_m = A$ follows from our assumption that π is onto).

To prove the second one, denote $D_0 = T_0$ and $D_k = T_k \setminus T_{k-1}$, $d_k = |D_k|$ for $k = 1, \ldots, m$; fix a numeration of each set D_k, i.e. a function $n(a)$ with the range $\{0, \ldots, d_k - 1\}$ over each D_k (so $n(a) = 0$ for $a \in D_0 (= \{a\})$). Now we have a function $k(a)$ satisfying $a \in D_{k(a)}$ for every $a \in A$ (as $T_m = A$, $k(a)$ is defined everywhere), and a function $r(k, n)$ (the outcome number n from D_k) satisfying $r(k(a), n(a)) = a$ for every $a \in A$.

Now we choose $X_1 = X_2 = A$ and define π as follows:

$$\pi(a, b) = \begin{cases} a, & \text{if } k(a) < k(b), \\ r(k(a), (n(a) + n(b)) \bmod d_{k(a)}), & \text{if } k(a) = k(b), \\ b, & \text{if } k(a) > k(b). \end{cases}$$

It is easy to see that $[a] = T_{k(a)}$ for every $a \in X_i$. The theorem is proved.

Examples 1 and 2 can be regarded as illustrating the proof of Theorem 4. The proof becomes especially clear when the sequence is maximal, i.e. each $T_k \setminus T_{k-1}$ is a singleton. In this case the sequence defines an ordering of the set A; each player chooses an outcome, and the minimal one is implemented. Moulin (1980) has shown that such games have a dominant strategy equilibrium if all the preferences are single-peaked on A (w.r.t. the ordering defined by T_k); Theorem 4 establishes their properties – SLF and SFF – true for every preferences (see also Theorem 5 below). It seems worth noting that the set A here may be interpreted as the set of feasible cooperation levels (discrete or continuous). Thus we arrive at an optimistic conclusion that there cannot be any need to try and pre-commit oneself to not exceeding a certain level of cooperation.

Whether and in what form Theorem 4 (especially its second statement) could be extended to non-finite A remains unclear. I end up this topic with just one example.

Example 5: Let A be a circle of the radius 1; using polar coordinates, we can describe each outcome with a pair (r, φ), $0 \le r \le 1$, $0 \le \varphi \le 2\pi$, assuming that all $(0, \varphi)$ are the same point. Define a game form over A as follows: $X_1 = X_2 = A$, $\pi((r_1, \varphi_1), (r_2, \varphi_2)) = (\min\{r_1, r_2\}, (\varphi_1 + \varphi_2) \bmod 2\pi)$. It is easy to see that $[(r^\circ, \varphi^\circ)] = \{(r, \varphi) \mid r \le r^\circ\}$; the ET condition is satisfied; the game form contains a cross, i.e. is tight. Therefore, it is perfect.

5 Incentive Compatibility and Stackelberg Solvability

Suppose a game form G is such that every derivative game has a Nash equilibrium (according to Gurvich (1988), this means G is tight). An (undominated) equilibrium selection rule is a mapping $\varepsilon : \mathcal{U} \to X_1 \times X_2$, where $\mathcal{U}$ is the space of pairs of utility functions $\langle u_1(\cdot), u_2(\cdot)\rangle$, such that $\varepsilon(u_1(\cdot), u_2(\cdot))$ is a Nash equilibrium in the derivative game (not Pareto inferior to any other Nash equilibrium). A selection rule ε is called non-manipulable if for every $\langle u_1(\cdot), u_2(\cdot)\rangle \in \mathcal{U}$, $i \in N$, $u_i'(\cdot)$ the following inequality holds:

$$U_i(\varepsilon(u_1(\cdot), u_2(\cdot))) \ge U_i(\varepsilon(u_i'(\cdot), u_j(\cdot))). \tag{4}$$

A selection rule ε is called secure if for every $\langle u_1(\cdot), u_2(\cdot)\rangle \in \mathcal{U}$, $i \in N$, $u_i'(\cdot)$, $x_i \in X_i$ the following inequality holds:

$$U_i(\varepsilon(u_1(\cdot), u_2(\cdot))) \ge U_i(x_i, \varepsilon_j(u_i'(\cdot), u_j(\cdot))) \tag{5}$$

(note that (5) in itself implies that $\varepsilon(u_1(\cdot), u_2(\cdot))$ is a Nash equilibrium). Both definitions have the same underlying interpretation: if each player's preferences are his private information, he cannot gain anything by falsifying his report on them. The difference is that in (5) the information is supposed to be transmitted "in words" (so a player may first misrepresent his utility and then follow his true one), while in (4) it is transmitted "in deeds" – the only way to misrepresent one's preferences is to behave as if the arbitrary invented utility were true; this difference is illustrated in Example 6 below. Certain analogy with Hurwicz's (1981) concept of manipulative equilibrium is present but our framework is quite different.

Theorem 5: For any game form G, the following three properties are equivalent:

(a) G is partially dictatorial;
(b) G possesses a non-manipulable equilibrium selection rule;
(c) G possesses a secure selection rule.

For a partially dictatorial game form, consider a selection rule prescribing player i to choose x_i satisfying $[x_i] = \{a\}$, where a is his best outcome in D (when player i is indifferent between two or more best outcomes, he should choose e.g. what is the best for player j among them), and player j always to choose x_j with $[x_j] = D$. The rule is obviously secure.

To prove the implication $(b) \Rightarrow (a)$, we recall the Gibbard – Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975): Either the selection rule is dictatorial, or its range contains no more than two outcomes. In the first case, suppose player 1 is the dictator and denote D the range of the composition of the selection rule and the projection π; for any $a \in D$ consider the following utilities: $u_1(a) = 1, u_1(b) = 0$ for any $b \neq a, u_2(a) = 0, u_2(b) = 1$ for any $b \neq a$. The outcome a being the only best outcome for player 1, a Nash equilibrium x with $\pi(x) = a$ must be selected; obviously, there must be $[x_1] = \{a\}$. If $D = A$, then G is dictatorial; otherwise, define $u_1(a) = 0$ for any $a \in D, u_1(b) = 1$ for any $b \in A \setminus D$; an equilibrium with $\pi(x) \in D$ must be selected, but such a strategy bundle is an equilibrium only when $[x_2] \subseteq D$, hence $[x_2] = D$.

The case when one of two strategy bundles x and y (with $\pi(x) = a$, $\pi(y) = b$) is always selected produces nothing new. Define $u_1(a) = 1, u_1(b) = 0, u_1(c) = 2$ for all $c \notin \{a, b\}, u_2(b) = 1, u_2(a) = 0, u_2(c) = 2$ for all $c \notin \{a, b\}$, and suppose, without restricting generality, that x is selected; then $[x_1] = \{a\}$. Now define $u_1(b) = 1$, $u_1(a) = 0, u_1(c) = 2$ for all $c \notin \{a, b\}, u_2(a) = 1, u_2(b) = 0, u_2(c) = 2$ for all $c \notin \{a, b\}$; if y is chosen, player 1 is the partial dictator with $D = \{a, b\}$; otherwise, $[x_2] = \{a\}$ and G contains a cross ($D = \{a\}$). (If $a = b$, the reasoning becomes even simpler.)

The implication (c) $\Rightarrow$ (b) is obvious; the theorem is proved.

Remark 1: There is some reason to blame the conditions (4) or (5) for requiring too little. For Examples 1–3 the proof of Theorem 5 suggests always to choose a, the cross, even when the players have the same preferences and a is the worst outcome. Unfortunately, this is a very rare occasion when an undominated Nash equilibrium can be selected in an incentive compatible way (Theorem 6 below).

Remark 2: Theorems 3 and 5 show that every SLF and SFF game form possesses a non-manipulable, and even secure, equilibrium selection rule. In principle, there is nothing surprising in a connection between incentive compatibility and stability against commitment tactics: If a selection rule results from an adaptive procedure and player i wishes to create the (wrong) impression that $u_i(a) > u_i(b)$, he has to commit himself to never choosing b when a is attainable. What is more interesting, the ability to commit himself to a fixed reaction rule, generally speaking, gives a player more power than just the ability to commit himself to a fixed strategy, see e.g. Schelling (1960, esp. Figs. 2 and 4), Kukushkin and Morozov (1984, Part 2), while in our current context the contrary is true. It seems futile to speculate on whether or not this should have been expected beforehand.

Naturally, both definitions (4) and (5) can be applied to restricted preference domains (with obvious modifications and under the assumption that each player knows his own utility). The concept of a non-manipulable equilibrium selection rule was introduced and studied (for n-person games) by d'Aspremont and Gérard-Varet (1980), who found a close connection between the existence of such selections and what they called Stackelberg solvability.

For $n = 2$, a normal form game is called Stackelberg solvable if the same outcome is a Stackelberg solution for each player; in other words this means the existence

of a Nash equilibrium satisfying (2), thus implying the absence of the struggle of both kinds. A game form G is called Stackelberg solvable if so is every derivative game.

Theorem 2.1 of d'Aspremont and Gérard-Varet (1980) states that if a game is Stackelberg solvable for every admissible bundle of utility functions, then it allows of a non-manipulable equilibrium selection rule. Theorems 2.2 and 2.3, in a sense, reverse the result: under some additional assumptions, the existence of a non-manipulable equilibrium selection rule implies that the game must be Stackelberg solvable for every admissible bundle of utility functions. Interestingly, the assumptions of both theorems cannot hold for an unrestricted preference domain; the statement of Theorem 2.2 is outright untrue in this case: the game forms of Examples 1–3 are not Stackelberg solvable (this may be shown independently or derived from Theorem 6 below); however, the statement of Theorem 2.3 – the existence of a non-manipulable undominated equilibrium selection rule implies Stackelberg solvability – remains valid.

Theorem 6: For any game form G, the following three properties are equivalent:

(a) G is Stackelberg solvable;
(b) G possesses a non-manipulable undominated equilibrium selection rule;
(c) one (and only one) of the conditions is satisfied:
  (c1) G is dictatorial, or
  (c2) G contains a cross and $|A| = 2$.

The implication (a) $\Rightarrow$ (b) is proved by d'Aspremont and Gérard-Varet (1980, Theorem 2.1). The implication (c) $\Rightarrow$ (a) is straightforward (even without quoting the previous theorems). Prove (b) $\Rightarrow$ (c).

For any $a \in A$ choose $u_i(a) = 1, u_i(b) = 0$ for both i and all $b \neq a$; a strategy bundle is an undominated Nash equilibrium if and only if $\pi(x) = a$. Thus the range of the composition of the selection rule and the projection π is the whole A. Now apply the Gibbard – Satterthwaite theorem: either the selection is dictatorial or $|A| = 2$.

Suppose the first case takes place and the selection is always favorable to player 1. For any $a \in A$ define $u_1(a) = 1, u_1(b) = 0$ for all $b \neq a$; $u_2(a) = 0, u_2(b) = 1$ for all $b \neq a$. A strategy bundle x with $\pi(x) = a$ must be selected, and therefore there must be $[x_1] = \{a\}$. Thus G is dictatorial with player 1 as dictator.

When $|A| = 2$, every tight game form either is dictatorial or contains a cross. The theorem is proved.

Remark 1: The relationship between the statement (a) $\Leftrightarrow$ (c) of Theorem 6 and the Gibbard – Satterthwaite theorem is rather close to that between Theorem 2 and the Gurvich theorem: A Stackelberg solvable game need by no means have a dominant strategy equilibrium, but the difference between the classes of Stackelberg solvable and straightforward game forms is insignificant.

Remark 2: Theorem 4 and examples show that there are various (in particular, anonymous) game forms guaranteeing both the compatibility of System (2) and the existence of a Nash equilibrium. However, Theorem 6 (statement (a) $\Leftrightarrow$ (c)) demonstrates that,

if we require these two conditions to be satisfied by the same outcome, we shall have "almost only" dictatorial game forms at our disposal.

At last, the connection between the existence of secure and non-manipulable equilibrium selection rules established in Theorem 5 does not hold under arbitrary restrictions on preferences.

Example 6: Consider a 2×2 bimatrix game where the utility function of player 2 is common knowledge while player 1 has two possible utility functions, and only he knows which is true. Thus we have two possible games:

$$\begin{matrix} (2,1) & (3,0) \\ (0,0) & (1,1) \end{matrix} \quad \text{or} \quad \begin{matrix} (1,1) & (0,0) \\ (3,0) & (2,1) \end{matrix}$$

Each of the games has a unique Nash equilibrium and is Stackelberg solvable; the rule selecting these equilibria is non-manipulable but not secure: if the information about player 1's utility is transmitted "in words", he might profit by sending the false report and then obtaining the utility level 3.

6 Concluding Remarks

1. It is instructive to compare Theorems 1–3 with Moulin's original classification. No normal form game can be characterized by the struggle for both leadership and followership, and all the three remaining classes are "massive" (e.g. have non-empty interiors in the utility functions space). Here, conversely, the most massive class are the game forms ready to produce the struggle of both kinds, while those free from the struggle of either kind are rather singular.

2. For $n > 2$, the problems described in Introduction emerge with no lesser urgency. However, their mathematical treatment is much more difficult. First of all, it becomes natural to consider coalitions. System (1) might be replaced with the condition that β-core is non-empty; to define the struggle for leadership, various versions of "γ-core" might be suggested. However, one cannot hope for a straightforward analogue of Moulin's Classification Theorem: if α-core is empty (which is by no means an exceptional case), all the other cores are also empty.

Certainly, we may restrict ourselves to non-coalitional concepts (Nash equilibrium, etc.), but nature will avenge itself on us. For instance, there seems to be no analogue of the Gurvich Theorem for $n > 2$, while Danilov and Sotskov (1988) have found conditions for the existence of a strong equilibrium for any preference profile (although in terms of effectivity functions and not of game forms).

In fact, something could be done for $n > 2$, but the whole picture still remains rather obscure.

References

d'Aspremont C, Gérard-Varet LA (1980) Stackelberg-solvable games and pre-play communication. J of Economic Theory 23 : 201–217

Basar T (1973) On the relative leadership property of Stackelberg strategies. J of Optimization Theory and Applications 11 : 655–661

Danilov VI, Sotskov AI (1988) Strongly consistent group choice mechanisms. Ekonomika i Matematicheskie Metody 24 : 113–124 (in Russian)

Gal-Or E (1985) First mover and second mover advantages. International Economic Review 26 : 649–653

Gibbard A (1973) Manipulation of voting schemes: a general result. Econometrica 41 : 587–601

Gurvich VA (1988) Equilibrium in pure strategies. Doklady Akademii Nauk SSSR 303 : 789–793 (in Russian; English translation in Soviet Mathematics. Doklady 38 : 597–602)

Hurwicz L (1981) On incentive problems in the design of non-wasteful resource allocation systems. In Assorodobraj-Kula N et al (eds) Studies in Economic Theory and Practice, North-Holland Publ, Amsterdam pp 93–106

Kukushkin NS (1991) Nash equilibria of informational extensions. University of Bielefeld, Institute of Mathematical Economics, Working Paper Nr. 204

Kukushkin NS, Morozov VV (1984) Theory of non-antagonistic games. Moscow State University Press, 2nd ed, Moscow (in Russian)

Luce RD, Raiffa H (1957) Games and decisions. John Wiley and Sons, New York

Moulin H (1976) Prolongements des jeux à deux joueurs de somme nulle. Bulletin de la Société Mathématique de France, Supplémentaire Mémoire N° 45

Moulin H (1980) On strategy proofness and single peakedness. Public Choice 35 : 437–455

Moulin H (1981) Deterrence and cooperation. European Economic Review 15 : 179–193

Moulin H, Peleg B (1982) Cores of effectivity functions and implementation theory. J of Mathematical Economics 10 : 115–145

Satterthwaite MA (1975) Strategy proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J of Economic Theory 10 : 187–217

Schelling TC (1960) The strategy of conflict. Harvard University Press, Cambridge Mass.

Received June 1992
revised version December 1992
final version December 1993