Journal of Mathematical Economics 10 (1982) 49–65. North-Holland Publishing Company

IMPLEMENTABILITY VIA PROTECTIVE EQUILIBRIA*

Salvador BARBERÁ
Universidad del País Vasco, Bilbao 14, Spain

Bhaskar DUTTA
Indian Statistical Institute, Calcutta, India

We present a notion of non-cooperative strategic equilibrium for games generated by social choice functions, and fully characterize the class of those functions which are directly implementable under this equilibrium concept. Correct preference revelation turns out to always be such an equilibrium for the games generated by this class of implementable functions.

1. Introduction

Social choice functions choose the 'best' outcome corresponding to each profile of individual preferences. Implementing a given social choice function is to make sure that the relationship it establishes between individual preferences and 'best' outcomes always holds. One possible way of implementing a given rule is to first ascertain what individual preferences are, and then find their image. However, since information on an individual's preference is private to the individual, this method will not work unless the individual is motivated to reveal his actual preference.

A particularly clear-cut case arises under rules for which revealing one's own preference is always a dominant strategy. In this case, no individual would ever gain by revealing preferences other than those by which he actually evaluates social outcomes, and such rules would be implementable provided individuals are rational. Moreover, implementation could be decentralized, since computation of one's own preferences does not require any knowledge about the preferences and/or the strategies of others. Unfortunately, the Gibbard/Satterthwaite theorem shows that truthful revelation of preferences is always a dominant strategy only under trivial social choice functions.

Thus, in general, the possibility of implementing a social choice function is subject to a number of qualifications. A social choice function f can be viewed as the outcome function of a *game form* where individual strategies are the possible

*The authors gratefully acknowledge helpful criticism by two referees, which led to a reassessment of the result presented in section 4.

0304–4068/82/0000–0000/$02.75

preference orderings of alternatives.[1] Each specification of a preference profile will determine a pay-off function for each player and thus complete the description of one of the games generated by f. We can analyse games within this class by using different equilibrium concepts, each reflecting our assumptions about the information that each individual has about the others' preferences and about the strategic behaviour of individuals. A social choice function is *implementable via equilibria of type* α iff it is the case that for each game generated by f, the outcomes associated with any of its equilibria of type α coincides with the image under f of the profile determining the pay-offs. The underlying motivation for this definition is that whatever the individuals' preferences might be, such a social choice function would lead to its expected outcomes provided the individuals' strategic behaviour led them to settle at equilibria of type α.

The concept of implementable social choice functions was first introduced by Maskin (1977). Maskin proved that only dictatorial social functions or social choice functions whose range is restricted to only two outcomes are implementable via Nash equilibria. Again, keeping to a non-cooperative framework, and further assuming that every agent has full information about the preference profile, Moulin (1979) has examined the implementation problem using Farquharson's notion of *sophisticated voting*, where the agents mutually anticipate their strategies by successively eliminating dominated strategies.

Both the above specifications of equilibrium behaviour require extreme assumptions about the extent of information possessed by agents. Each agent knows not only the preferences of other individuals, but also their strategic behaviour. In this paper, we make the polar assumption that agents have no information at all about other agents' preferences, and hence about their strategic behaviour. We assume that this leads an agent to use 'protective' strategies of a lexical maximin type, and completely characterise the class of social functions which are implementable when individuals so choose their strategies.

In our way toward a technical characterization we discover an important fact: if a social choice function is directly implementable via protective equilibria, then correct preference revelation by all individuals is always a protective equilibrium. Therefore, we recover some of the interesting features of implementation by dominant strategies: our functions can be implemented in a decentralized way, and truthful revelation of preferences becomes one practical way toward this goal. While the observation that truthful preference revelation is not an objective per se lies at the origin of the implementation literature, truth turns out to be here the (essentially unique) device for decentralized implementation of social decision functions via protective equilibria.

The paper is organized as follows. In section 2, we introduce the basic notation and definitions. Section 3 introduces the concept of protective equilibria and

[1]In general, a game form is an $(n+1)$-tuple $g=(X_1,\ldots,X_n,\pi)$ where X_i is the strategy set of individual i, and π is the outcome function. A social choice function f which is implemented through the game from $g=(\mathscr{P},\mathscr{P},\ldots,\mathscr{P},f)$ is said to be *directly implementable*.

discusses its connection to related concepts. In section 4 we state and prove a theorem relating protective to truthful strategies under social decision functions which are implementable via our equilibrium concept. Section 5 presents some examples which illustrate the structure of such social choice functions, of which section 6 provides a complete characterisation. Section 7 contains some concluding remarks.

2. The notations and basic definitions

Let $A=\{x, y, z, \ldots\}$ be a finite set of alternatives, with cardinality m. Let $I=\{1, 2, \ldots, n\}$ be an initial segment of the integers, whose elements are called individuals. $\mathscr{P}$ is the set of *asymmetric orderings* over A. Elements of $\mathscr{P}$ are represented by $P, P', P_1, P_j, \ldots$, and are called *preferences.*

If $P \in \mathscr{P}$, $Y \subset A$, we say that Y *is bottom for* P iff $(\forall y \in Y)\ (\forall x \in A - Y)\ xPy$.

For $k \in [1, m]$, $P \in \mathscr{P}$, the k-bottom of P denoted by $B(k, P)$ is the unique subset of A which is bottom for P and contains exactly k alternatives.

For $P, P' \in \mathscr{P}$, $Y \subset A$, we say that P *and* P' *agree on* Y iff $(\forall x, y \in Y)\,[xPy \leftrightarrow xP'y]$. Where $P \in \mathscr{P}$, $r \in [1, m]$, the rth *ranking worst alternative in* P, denoted by $a_r(P)$, is defined by $a_r(P)=\{x \in A/\text{there exist exactly } (r-1) \text{ alternatives } y \in A : xPy\}$.

Let $\mathscr{P}^n$ be the n-fold cartesian product of $\mathscr{P}$. Elements of $\mathscr{P}^n$ are denoted by $\boldsymbol{P}, \boldsymbol{P}', \ldots$ and are called *preference profiles.*

Let $i \in I$. Given a preference profile $\boldsymbol{P}=(P_1, P_2, \ldots, P_{i-1}, P_i, P_{i+1}, \ldots P_n)$, we may denote it by $\boldsymbol{P}=(P_i, P_{-i})$, where $P_{-i}=(P_1, P_2, \ldots, P_{i-1}, P_{i+1}, \ldots, P_n)$. Given $\boldsymbol{P} \in \mathscr{P}^n$ and $P_i' \in \mathscr{P}$, $\boldsymbol{P}/P_i' \stackrel{\text{def}}{=} (P_1, P_2, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n)$; i.e., $\boldsymbol{P}/P_i'$ stands for a profile obtained from $\boldsymbol{P}$ by changing its ith component from P_i to P_i'.

A *social choice function* (SCF) is a function $f : \mathscr{P}^n \to A$.

An SCF f is *non-dictatorial* (ND) iff there does not exist $i \in I$ such that for all $\boldsymbol{P} \in \mathscr{P}^n$, $f(\boldsymbol{P})$ is P_i-maximal in the range of f.

An SCF satisfies the *Pareto Criterion* (P) iff for all $\boldsymbol{P} \in \mathscr{P}^n$, for all $y \in A$, if there exists x such that $xP_i y$ for all $i \in I$, then $f(\boldsymbol{P}) \neq y$.

Given any SCF f, for any $P_i \in \mathscr{P}$ and $x \in A$, we denote by $g_f(x, P_i)$ the set $\{P_{-i} \in \mathscr{P}^{n-1} / f(P_{-i}, P_i)=x\}$. When there is no ambiguity about the SCF, we will simply write $g(x, P_i)$, $g(x, P_i')$, etc.

Let $i \in I$, $P_i, P_i' \in \mathscr{P}$ and $Y \subset A$. We say that P_i *and* P_i' *are* Y*-equivalent for* i *under* f iff $(\forall y \in Y)\ [g_f(y, P_i)=g_f(y, P_i')]$. P_i and P_i' are equivalent for i iff they are A-equivalent.

3. Protective equilibria

The Gibbard–Satterthwaite result is sometimes interpreted to mean that all non-dictatorial SCF's induce strategic misrevelation of preferences. However, what the Gibbard–Satterthwaite result actually tells us is that non-dictatorial SCF's are not *strategy proof*, i.e., truth-telling is not a dominant strategy for all

profiles and all individuals. In order to deduce the negative conclusion that individuals will misreveal their preferences, one has to describe the strategic behaviour of agents when they no longer have dominant strategies. In this section, we specify an informational framework and make a behavioural assumption under which individuals will in fact always use their truthful strategies for a certain class of SCF's.

To be more specific, we assume that no agent has information about other agents' preferences. This obviously implies that agents do not have any information about the strategic choices which are likely to be made by others. Suppose also that all agents are extremely risk-averse. Under these assumptions, agents would choose their strategies so as to 'protect' themselves from the worst eventuality as far as possible. We capture this notion by specifying a type of 'lexical maximin' behaviour.

Let f be a given SCF. All ensuing definitions are relative to f.

For any $i \in I$, given his preference P_i, a strategy P_i^* *protectively dominates* P_i^{**}, denoted by $P_i^* d(P_i) P_i^{**}$, if there exists $k \in [1, m]$ such that

(i) $\quad g(a_k(P_i), P_i^*) \subsetneq g(a_k(P_i), P_i^{**})$,

and

(ii) $\quad g(a_r(P_i), P_i^*) = g(a_r(P_i), P_i^{**}), \qquad \forall r < k.$

Let

$$D(P_i) = \{P_i' \in \mathscr{P} / \sim P_i^* d(P_i) P_i' \text{ for any } P_i^*\}.$$

We say that $\boldsymbol{P} \in \mathscr{P}^n$ is *a protective equilibrium under* $\bar{\boldsymbol{P}}$ iff $(\forall i \in I)\ P_i \in D(\bar{P}_i)$.

A social choice function f is *directly implementable via protective equilibria* (d.i.p.e.) iff $(\forall \boldsymbol{P}, \bar{\boldsymbol{P}} \in \mathscr{P}^n)$ [$\boldsymbol{P}$ is a protective equilibrium under $\bar{\boldsymbol{P}}$] $\rightarrow$ [$f(\boldsymbol{P}) = f(\bar{\boldsymbol{P}})$].

Thus, in evaluating two strategies P_i^* and P_i^{**}, agent i compares $g(a_1(P_i), P_i^*)$ and $g(a_1(P_i), P_i^{**})$. If, for instance, $g(a_1(P_i), P_i^*)$ is a proper subset of $g(a_1(P_i), P_i^{**})$, then agent i prefers to use P_i^* rather than P_i^{**} since whenever P_i^* gives him his worst outcome (in terms of P_i), P_i^{**} also does the same, but the converse is not true. If the two strategies are $a_1(P_i)$-equivalent, then he compares $g(a_2(P_i), P_i^*)$ and $g(a_2(P_i), P_i^{**})$, and so on. Having eliminated protectively dominated strategies in this manner, agent i is left with the set $D(P_i)$. An SCF is d.i.p.e. iff for all preference profiles $\boldsymbol{P} = (P_1, P_2, \ldots, P_n)$, any choice of strategies $\hat{P}_i$ by all individuals from their corresponding sets $D(P_i)$ would guarantee that $f(\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_n) = f(\boldsymbol{P})$.

The notion of protective equilibria is closely related to Moulin's (1981) concept of prudent behaviour.[2]

[2]See also Barberá's (1980) notion of 'protective stability' which essentially requires that truth-telling be the unique maximin strategy.

For any $i \in I$, given $P_i \in \mathscr{P}$, P_i^* is a *prudent* strategy iff there does not exist P_i^{**} such that for some $k \in [1, m]$ the following holds:

(i) $\quad |g(a_k(P_i), P_i^{**})| < |g(a_k(P_i), P_i^*)|,$

(ii) $\quad |g(a_r(P_i), P_i^{**})| = |g(a_r(P_i), P_i^*)|, \qquad \forall r < k.$

The concept of implementation via prudent strategies and d.i.p.e. are both relevant within a framework where agents are extremely risk-averse and they have no information about other agents' preferences. However, we feel that the assumption of risk-averse agents using 'protectively non-dominated' strategies is more appropriate than that of specifying use of 'prudent' strategies. Implicit in the concept of prudent behaviour seems to be the assumption that an agent i considers all $P_{-i} \in \mathscr{P}^{n-1}$ equally likely. Whether the equi-probability assumption is a good representation of complete ignorance is a debatable question. Note that the concept of 'protective equilibria' is applicable even if agents have no subjective probability distribution about others' strategies.

Of course, $D(P_i)$, the set of protectively non-dominated strategies, contains the set of prudent strategies. This makes d.i.p.e. more discriminatory than implementation through prudent strategies. Indeed, Moulin (1981) remarks that under rules like plurality voting and the Borda count, the only prudent strategy is to tell the truth: this suffices to make them implementable via sets of prudent strategies. However, that these rules are not d.i.p.e. will be evident from our characterisation theorem. The relationship between truth-telling and implementation in our setting will be discussed in the next section.

The assumption that agents have no information about other individuals' preferences is the polar extreme of the assumption of complete information, which is implicit in notions like Farquharson's 'sophisticated' behaviour (i.e., successive elimination of dominated strategies, based on the assumption that other agents are also eliminating their dominated strategies) or the more common Nash equilibrium. However, the difference in these concepts does not end with different knowledge requirements alone. For suppose that agent i knows j's preference ordering P_j, but does not know anything at all about j's strategic behaviour. Even in such circumstances, notions based on maximin behaviour are applicable. (They may become less appealing, however, because even a weak behavioural assumption that all agents use only *non-dominated* strategies would call for a modification of prudent or protective behaviour.) On the other hand, the use of sophisticated strategies entails not only complete information about preferences, but also strong assumptions about strategic behaviour.

The notions of protective, prudent and sophisticated behaviour focus directly on an individual's *choice* of strategies. On the other hand, Nash equilibrium and similar concepts of equilibrium behaviour seem to implicitly assume that a trial and error process goes on until the equilibrium is arrived at.

Sengupta (1978) points out that in two-person, non-zero-sum games, it is possible that the two individuals' maximin strategies may not coincide with their Nash equilibrium strategies, and goes on to argue that even if a strategy n-tuple is not in equilibrium, individuals may still choose these strategies. This line of argument has obvious significance for the problem of implementation, particularly in a non-cooperative context, and seems to make concepts like 'protective equilibria' more interesting.

4. Truthful revelation under implementable functions

This section is devoted to prove that a social choice function is d.i.p.e. if and only if, for all individuals i and all preferences P_i, the set of protectively undominated strategies $D(P_i)$ consists of strategies which are equivalent to P_i (and thus, in particular, always contains P_i). This provides us with an alternative definition of d.i.p.e. functions, to be used in all proofs thereafter.

Besides its technical convenience, we find the result to be interesting on its own. There is no 'a priori' reason why implementation of social choice functions should be achieved through equilibria involving truthful preference revelation. In fact, focusing on the implementability of rules reflects the consideration that correct revelation is not an objective per se, and that what we really care about is the result of strategic behaviour, rather than its correspondence to truth. Our result indicates, however, that the only rules which are implementable in our sense are those which in fact guarantee gruthful revelation by all individuals, under the behavioural and informational assumptions underlying the definition of a protective equilibrium.

Theorem 1. A social choice function f is directly implementable via protective equilibria iff for all $i \in I$, all $P_i \in \mathscr{P}$,

$$D(P_i) = \{P_i^*/P_i^* \text{ and } P_i \text{ are equivalent}\}.$$

Proof. The condition on $D(P_i)$ is obviously sufficient for an SCF to be d.i.p.e. We must prove that it is also necessary. All statements below refer to a given d.i.p.e. SCF, f.

(a) We first show that, if P_i^* and P_i^{**} are not equivalent, then $D(P_i^*) \cap D(P_i^{**}) = \varnothing$. That is, for each P_i, $D(P_i)$ consists of exactly one class of equivalent strategies. Since $\mathscr{P}$ is finite, this also implies that every P_i must belong to some $D(P_i^*)$.

Suppose $P_i \in D(P_i^*) \cap D(P_i^{**})$. Since P_i^* and P_i^{**} are not equivalent, $\exists P_{-i} \in \mathscr{P}^{n-1}$ such that $f(P_i^*, P_{-i}) = x, f(P_i^{**}, P_{-i}) = y$ and $y \neq x$. Let $\hat{P}_j \in D(P_j)$ for all $j \neq i$. Then, since f is d.i.p.e., $f(\hat{P}_i, \hat{P}_{-i}) = x$ and $f(\hat{P}_i, \hat{P}_{-i}) = y$, which is not possible.

(b) To complete the proof it suffices to show that the one class of equivalent strategies $D(P_i)$ consists of those which are equivalent to P_i, for all P_i. Thus, we suppose there was some P_i^0 such that $P_i^0 \notin D(P_i^0)$, and show that this leads to a contradiction.

Since $P_i^0 \notin D(P_i^0)$, there exists $P_i^1 \in D(P_i^0)$ such that $P_i^1 d(P_i^0) P_i^0$. Hence, there exists $P_{-i}^0 \in \mathscr{P}^{n-1}$, $x, y \in A$, such that

$$f(P_i^0, P_{-i}^0) = x, \tag{1}$$

$$f(P_i^1, P_{-1}^0) = y, \tag{2}$$

$$g(x, P_i^1) \subset g(x, P_i^0). \tag{3}$$

Since P_i^0 and P_i^1 are not equivalent, it cannot be, by (a), that $P_i^1 \in D(P_i^1)$. Thus, there must exist P_i^2 such that $P_i^2 \in D(P_i^1)$, and P_i^2 is not equivalent to P_i^1. By repeated application of this type of argument we can construct a sequence

$$P_i^0, P_i^1, P_i^2, P_i^3, \ldots$$

of elements of $\mathscr{P}$, such that, for all t, $P_i^t \notin D(P_i^t)$ and $P_i^t \in D(P_i^{t-1})$.

Since $\mathscr{P}$ is finite, there must exist integers T, S such that P_i^T and P_i^{T+S} are equivalent. But then, by (a) P_i^{T-h} and P_i^{T+S-h} are also equivalent, for $h = 1, 2, \ldots, T$. In particular, P_i^0 and P_i^S are equivalent.

Let now $\{P_{-i}^1, \ldots, P_{-i}^T\}$ be a sequence of elements in $\mathscr{P}^{n-1}$ such that $P_k^t \in D(P_k^{t-1})$ $(\forall k \neq i)$.

Since f is d.i.p.e.,

$$(1) \rightarrow f(P_i^1, P_{-i}^1) = x, \tag{4}$$

$$(3) \text{ and } (4) \rightarrow f(P_i^0, P_{-i}^1) = x. \tag{5}$$

Since f is d.i.p.e.,

$$(5) \rightarrow f(P_i^1, P_{-i}^2) = x, \tag{6}$$

$$(5) \text{ and } (6) \rightarrow f(P_i^0, P_{-i}^2) = x. \tag{7}$$

In this way, by alternate use of (3) and the fact that f is d.i.p.e., one can show that, for all t in our sequence, $f(P_i^0, P_{-i}^t) = x$, and, in particular,

$$f(P_i^0, P_{-i}^{S-1}) = x. \tag{8}$$

Also, by repeated application of the fact that f is d.i.p.e., from (2) we get $f(P_i^t, P_{-i}^{t-1}) = y$ for all t in our sequence and, in particular,

$$f(P_i^S, P_{-i}^{S-1}) = y. \tag{9}$$

But (8) and (9) contradict the fact that P_i^0 and P_i^S were equivalent.

This completes the proof of Theorem 1.

5. Some examples

The purpose of this section is to illustrate with some simple examples the structure of d.i.p.e. SCF's. The first two are examples of implementable SCF's. Examples 3–5 are examples of SCF's which are not d.i.p.e. These have been chosen so as to illustrate the importance of the necessary conditions described in the next section. Throughout this section, Q represents an arbitrary asymmetric ordering over A. The reader can refer to the property given by Theorem 1 when checking our assertions on the implementability of the functions given in the examples.

Example 1. For any $i \in I$, for any $\alpha \in A$, given $P_i \in \mathscr{P}$,

$$e_i(x) = 0 \quad \text{if } \{x\} \text{ is bottom in } P_i,$$

$$= 1 \quad \text{otherwise.}$$

Then, for any $\boldsymbol{P} \in \mathscr{P}^n$, $f_1(\boldsymbol{P}) = x$, where, for all $y \in A - \{x\}$,

$$\left[\left(\sum_{i\in I} e_i(x) > \sum_{i\in I} e_i(y)\right) \quad \text{or} \quad \left(\sum_{i\in I} e_i(x) = \sum_{i\in I} e_i(y) \text{ and } xQy\right)\right].$$

f_1 is a member of the class of 'approval voting' schemes introduced by Fishburn and Brams (1978). Given an individual's preference P_i, the worst alternative is assigned zero points, while all others are assigned one point. For any profile, f_1 chooses that alternative which gets the highest number of points, while Q is used as a tie-breaking device.

f_1 is d.i.p.e.

Remark 1. f_1 is not Paretian. This reflects the well-known result that noncooperative equilibria need not necessarily be efficient.

We now give an example of an SCF which is d.i.p.e. and Paretian.

Example 2. Let $q=(q_1, q_2, .., q_n)$ and $r=(r_1, \ldots, r_m)$ be n- and m-tuples of positive integers satisfying $\sum_{i=1}^{n} q_i = \sum_{j=1}^{m} r_j - 1$. Denote by rA the set where each alternative $x_j \in A$ is replicated r_j times. For any $\boldsymbol{P} \in \mathscr{P}^n$, let

$$M_1(\boldsymbol{P}) = \{a/a \text{ is amongst the } q_1\text{-worst elements of } rA \text{ according to } P_1\},$$

$$M_i(\boldsymbol{P}) = \left\{a/a \text{ is amongst the } q_i\text{-worst elements of } A - \bigcup_{l<i} M_l(\boldsymbol{P}) \text{ according to } P_i\right\}.$$

Then,

$$f_2(\boldsymbol{P}) = rA - \bigcup_{i \in I} M_i(P).$$

Given that $\sum_{i=1}^{n} q_i = \sum_{j=1}^{m} r_j - 1$, $f_2(\boldsymbol{P})$ is unique for all $\boldsymbol{P}$. Intuitively, each $i \in I$ is given q_i 'tokens' with which to successively veto alternatives, while it takes r_j tokens to veto x_j. The best outcome for any profile is the alternative which is not vetoed.

For every choice of q and r, f_2 is d.i.p.e.

Remark 2. If $m \geqq n$, then f_2 must give *some* individual i veto power over *some* alternative a, in the sense that a is never the outcome for profiles where it is worst in i's preference ordering.

Various forms of 'voting with veto' rules have been extensively studied in the recent literature, following Mueller (1978). Moulin and Peleg (1982) have recently shown the close relation between veto power and implementability via strong Nash equilibria. However, it appears that veto power has no clear connection with implementability via protective equilibria. f_1 shows that veto power is not necessary for d.i.p.e.; the following examples show that veto power is not sufficient either.

Example 3 [Dutta and Pattanaik (1978)]. Let a^* be the Q-maximal element in A. For any $\boldsymbol{P}$, let

$$E(\boldsymbol{P}) = \{x \in A/\text{there does not exist } y \in A : yP_i x \text{ for all } i \in I\}.$$

Then,

$$f_3(\boldsymbol{P}) = a^* \quad \text{if} \quad a^* \in E(\boldsymbol{P})$$

$$= a \quad \text{where } a \text{ is the } Q\text{-maximal element in } E(\boldsymbol{P}) \cap \{x/xP_i a^* \text{ for all } i \in I\}.$$

Thus, f_3 selects a^* if a^* is Pareto-optimal; otherwise it selects the Q-maximal element amongst those elements which Pareto-dominate a^* and which are also Pareto-optimal.

f_3 gives veto power to all individuals over the set $A-\{a^*\}$. However, f_3 is not d.i.p.e. This is shown by the following example. Let

$$A=\{a_1,a_2,a_3,a^*\}, \quad I=\{1,2\},$$

P_1	P'_1	P_2	Q
a_3	a_1	a_2	a^*
a_1	a_3	a_3	a_1
a_2	a_2	a_1	a_2
a^*	a^*	a^*	a_3

Then, $f_3(P_1,P_2)=a_2$, $f_3(P'_1,P_2)=a_1$, and P_1 and P'_1 are a^*-equivalent. Since $a_1P_1a_2$, f_3 is not d.i.p.e.

Remark 3. P'_1 is derived from P_1 by a reshuffling of alternatives above $f_3(P_1,P_2)=a_2$. If f_3 is to be d.i.p.e., such changes should not alter the outcome.

Example 4. Let $|A|=4$, $I=\{1,2,3,4\}$. For any $\boldsymbol{P}\in\mathscr{P}^n$,

(i) If P_1 and P_2 *agree on* $B(2,P_1)$, then $f_4(\boldsymbol{P})=x$, where x is P_3-maximal on $A-B(2,P_1)$.

(ii) If P_1 and P_2 do not agree on $B(2,P_1)$, then $f_4(\boldsymbol{P})=y$, where y is P_4-maximal on $A-B(2,P_1)$.

The following proves that f_4 is not d.i.p.e.:

P_1	P'_1	P_2	P_3	P_4
x	x	x	y	x
y	y	y	x	y
z	w	z	z	z
w	z	w	w	w

Then, $f_4(P_1,P_{-1})=y$, $f_4(P'_1,P_{-1})=x$, where $P_{-1}=(P_2,P_3,P_4)$. Also, P_1 and P'_1 are $\{z,w\}$-equivalent. Clearly, $\sim P_1d(P_1)P'_1$, so that f_4 is not d.i.p.e.

Remark 4. Under f_4, P_1 and P_1' are $B(2, P_1)$-equivalent, $B(2, P_1) = B(2, P_1')$, but the order in which 1 vetoes alternatives changes the outcome. This leads to a violation of d.i.p.e.

Our last example shows violation of another of the necessary conditions for d.i.p.e.

Example 5. Let $A = \{a_1, a_2, a_3\}$. For any $\boldsymbol{P} \in \mathscr{P}^n$,

$$f_5(\boldsymbol{P}) = a_1 \quad \text{when } a_2 = B(1, P_i) \text{ for all } i \in I.$$

$$= a_2 \quad \text{otherwise.}$$

Consider the following $P_1, P_1', P_2 \in \mathscr{P}$:

P_1	P_1'	P_2
a_1	a_1	a_1
a_2	a_3	a_3
a_3	a_2	a_2

Then, $f_5(P_1, P_2) = a_2$, $f_5(P_1', P_2) = a_1$. It is easy to check that $P_1' d(P_1) P_1$. Hence, f_5 is not d.i.p.e.

Remark 5. There is an important qualitative difference between f_3, f_4 and f_5. Under f_5, P_1 (and hence all strategies *equivalent* to P_1) does not belong to $D(P_1)$, while for f_3 and f_4, $D(P_i)$ will be 'large' since it will contain P_i as well as strategies which are *not* equivalent to P_i. In a sense, lack of information about others' preferences or their strategic behaviour causes non-implementability of f_3 and f_4.

6. The characterisation

In this section, we provide the necessary and sufficient conditions for a social choice function to be implementable via protective equilibria.

It is useful to recall a condition introduced by Muller and Satterthwaite (1977):

Strong Positive Association (SPA). $(\forall i \in I)(\forall \boldsymbol{P} \in \mathscr{P}^n)(\forall P_i' \in \mathscr{P})[f(\boldsymbol{P}) = x$, and $(\forall z)(xP_i z \rightarrow xP_i' z)] \rightarrow f(\boldsymbol{P}/P_i') = x$.

The following result is well-known in the literature.

Proposition. The following statements are equivalent when $|A| \geqq 3$:

(A) *f is strategy proof.*
(B) *f is implementable via Nash equilibrium.*
(C) *f is implementable via strong Nash equilibrium.*
(D) *f is dictatorial.*
(E) *f satisfies SPA.*

Since a dictatorial SCF is d.i.p.e., SPA is a sufficient condition, although it is clearly not necessary in view of our Examples 1 and 2. However, these equivalence results seem to suggest that conditions like SPA are important. An obvious point of departure is to factorise SPA, and then see what parts are crucial for d.i.p.e.

Condition 1 (Monotonicity). $(\forall i \in I)(\forall \boldsymbol{P} \in \mathscr{P}^n)(\forall P_i' \in \mathscr{P})[(f(\boldsymbol{P}) = x), (P_i \text{ and } P_1' \text{ agree on } A - \{x\}) \text{ and } (\forall z)(xP_i z \to xP_i' z)] \to f(\boldsymbol{P}/P_i') = x$.

Condition 2. $(\forall i \in I)(\forall \boldsymbol{P} \in \mathscr{P}^n)(\forall P_i' \in \mathscr{P})[(f(\boldsymbol{P}) = a_r(P_i)) \text{ and } (B(r, P_i) = B(r, P_i')) \text{ and } (P_i \text{ and } P_i' \text{ agree on } B(r, P_i))] \to f(\boldsymbol{P}/P_i') = f(\boldsymbol{P})$.

Condition 3. $(\forall i \in I)(\forall \boldsymbol{P} \in \mathscr{P}^n)(\forall P_i' \in \mathscr{P})[(f(\boldsymbol{P}) = a_r(P_i)) \text{ and } (B(r, P_i) = B(r, P_i')) \text{ and } (P_i \text{ and } P_i' \text{ agree on } A - B(r, P_i))] \to f(\boldsymbol{P}/P_i') = f(\boldsymbol{P})$.

It is straightforward to show that SPA is equivalent to the conjunction of Conditions 1, 2 and 3. Condition 1 is the familiar 'non-perversity' condition. Condition 2 states that if x is the outcome under a profile, and the *only* change is that the relative ordering among alternatives that an individual i ranked *better than* x is altered, then x should still be the outcome under the new profile. Condition 3 is the 'dual' to Condition 2, and requires that the outcome x should not change if the new profile is obtained from the old by a reshuffling of alternatives below x.

We will retain Conditions 1 and 2, *the conjunction of which will be denoted Upper Strong Positive Association* (USPA). Condition 3 will be replaced by the following conditions:

Lower Conditional Independence (LCI). $(\forall i \in I)(\forall \boldsymbol{P} \in \mathscr{P}^n)(\forall P_i' \in \mathscr{P})[(f(\boldsymbol{P}) = a_{r+1}(P_i)) \text{ and } (B(r, P_i) = B(r, P_i')), \text{ and } (P_i \text{ and } P_i' \text{ are } B(r, P_i)\text{-equivalent and they agree on } A - B(r, P_i))] \to f(\boldsymbol{P}/P_i') = a_{r+1}(P_i)$.

Bottom Equivalence (BE). $(\forall i \in I)(\forall P_i, P_i' \in \mathscr{P})[(P_i \text{ and } P_i' \text{ are } B(k, P_i)\text{-equivalent, but are not } B(k+1, P_i)\text{-equivalent}) \text{ and } (P_i \text{ and } P_i' \text{ agree on } A - B(k, P_i))] \to B(k, P_i) = B(k, P_i')$.

LCI essentially states the following. Suppose Y is a bottom set of P_i and P_i', and they are Y-equivalent. Moreover, these strategies agree on $A-Y$, so that the only difference between P_i and P_i' is a reshuffling of alternatives in Y. Then, LCI requires that P_i and P_i' should be *equivalent*. An implication of LCI is that if an individual can veto two or more alternatives, then the order in which he vetoes the alternatives should not influence the final outcome, so long as he vetoes the same set of alternatives. Although the condition may sound innocuous, we have already seen a violation of LCI — f_4.

BE is somewhat more complicated. It requires that if Y is the maximal bottom set of P_i for which P_i and P_i' are equivalent, then Y is also a bottom set for P_i'. Both f_3 and f_5 violate this condition.

Lemma. *BE, LCI and USPA are logically independent.*

Proof. See the appendix.

Theorem 2. *An SCF f is directly implementable via protective equilibria iff it satisfies USPA, LCI and BE.*

Proof. *Necessity*: (i) Suppose f violates USPA. Then, there exists $i\in I$, $\boldsymbol{P}\in\mathscr{P}^n$, $P_i'\in\mathscr{P}$ such that

(a) $f(\boldsymbol{P})=x=a_r(P_i)$,

(b) $B(r-1,P_i)=B(r-1,P_i')$,

(c) P_i and P_i' agree on $B(r-1,P_i)$,

(d) $f(\boldsymbol{P}/P_i')=y\neq x$.

(A) Suppose yP_ix. We want to show that $\sim P_id(P_i)P_i'$. Suppose $P_id(P_i)P_i'$. Given yP_ix, (a) and (d) imply that for some $k<r$, P_i and P_i' are (1) $B(k-1,P_i)$-equivalent, *and* (2) $g(a_k(P_i),P_i)\subsetneq g(a_k(P_i),P_i')$. However, (b) and (c) imply that P_i and P_i' agree on $B(k-1,P_i)$, $B(k-1,P_i)=B(k-1,P_i')$, and $a_k(P_i)=a_k(P_i')$. From (1) and (2), it follows that $P_id(P_i')P_i'$, so that f is not d.i.p.e.

(B) If xP_iy, then in an analogous manner, it can be shown that $\sim P_i'd(P_i')P_i$. Hence, USPA is necessary for d.i.p.e.

(ii) Suppose f violates LCI. Then, there exist $i\in I$, $\boldsymbol{P}\in\mathscr{P}^n$, $P_i'\in\mathscr{P}$ such that $f(\boldsymbol{P})=a_{r+1}(P_i)\equiv x$, $B(r,P_i)=B(r,P_i')$, P_i and P_i' agree on $A-B(r,P_i)$ and they are $B(r,P_i)$-equivalent, but $f(\boldsymbol{P}/P_i')=y\neq x$.

Since P_i and P_i' are $B(r,P_i)$-equivalent, it must be true that yP_ix. Hence, $g(x,P_i)\not\subset g(x,P_i')$. Since $x=a_{r+1}(P_i)$ and $g(a_k(P_i),P_i)=g(a_k(P_i),P_i')$ $(\forall k\leqq r)$, this shows that $\sim P_id(P_i)P_i'$. Since P_i and P_i' are *not* equivalent, this shows that f is not d.i.p.e.

Hence, LCI is necessary for d.i.p.e.

(iii) If BE does not hold, then there exist $i \in I$, $P_i, P_i' \in \mathscr{P}$ such that

(a) P_i and P_i' are $B(k, P_i)$-equivalent,

(b) P_i and P_i' are not $B(k+1, P_i)$-equivalent,

(c) P_i and P_i' agree on $A - B(k, P_i)$,

(d) $B(k, P_i) \neq B(k, P_i')$.

Let $x = a_{k+1}(P_i)$. From (c) and (d) it follows that $x \in B(k, P_i')$. Since P_i and P_i' are not $\{x\}$-equivalent, there are two possibilities:

Case (1). There exists P^*_{-i} such that $f(P_i, P^*_{-i}) = x \neq f(P^*_{-i}, P_i') = y$.

Case (2). There exists P^{**}_{-i} such that $f(P_i', P^{**}_{-i}) = x \neq f(P^{**}_{-i}, P_i) = y$.

Suppose Case (1) holds. Since P_i and P_i' are $B(k, P_i)$-equivalent, we must have yP_ix. Hence $\sim P_i d(P_i) P_i'$.

Suppose Case (2) occurs. Let $x = a_l(P_i')$. From (a) and (c), it follows that $yP_i'x$ *and* P_i and P_i' are $B(l-1, P_i')$-equivalent. Hence, $\sim P_i' d(P_i') P_i$.

Thus, in either case, f is not d.i.p.e. and BE is necessary for d.i.p.e.

Sufficiency. Suppose P_i and P_i' are not equivalent. Let $a_r(P_i)$ be such that P_i and P_i' are *not* $\{a_r(P_i)\}$-equivalent, but are $B(r-1, P_i)$-equivalent. Construct P_i^* such that

(a) $B(r-1, P_i) = B(r-1, P_i^*)$,

(b) P_i and P_i^* agree on $B(r-1, P_i)$,

(c) P_i' and P_i^* agree on $A - B(r-1, P_i)$.

Given USPA, (a) and (b) imply that P_i and P_i^* are $B(r-1, P_i)$-equivalent. Since P_i and P_i' are $B(r-1, P_i)$-equivalent, we must have P_i^* and P_i' also $B(r-1, P_i)$-equivalent. Since P_i' and P_i^* agree on $A - B(r-1, P_i)$, by BE, we have $B(r-1, P_i) = B(r-1, P_i')$ or P_i' and P_i^* are $\{a_r(P_i)\}$-equivalent. In the latter case by USPA we derive $g(a_r(P_i), P_i) \subset g(a_r(P_i), P_i')$. In the former case we set P_i^{**} such that

$B(r-1, P_i) = B(r-1, P_i^{**})$,

P_i and P_i^{**} agree on $A - B(r-1, P_i)$,

P_i and P_i^{**} agree on $B(r-1, P_i)$.

Then by USPA, P_i' and P_i^{**} are $B(r-1, P_i)$ equivalent and, hence by LCI, $g(a_r(P_i), P_i) = g(a_r(P_i), P_i^{**})$. By USPA again, $g(a_r(P_i), P_i^{**}) \subset g(a_r(P_i), P_i')$. Hence, if P_i and P_i' are not $\{a_r(P_i)\}$-equivalent then $P_i d(P_i) P_i'$.

This completes the proof of the theorem.

7. Conclusion

In this paper, we have presented a notion of non-cooperative strategic equilibrium-protective equilibrium-relevant for risk-averse agents without any information about other individuals' preferences. If individuals do not have any information about other agents' strategic choices they have to choose their strategies based solely on knowledge about their own preferences and about the rules of the game. If a social choice function is to be implementable via protective equilibrium, then the set of protectively non-dominated strategies must coincide with the set of strategies equivalent to truth-telling. Under most of the well-studied SCF's like Borda count, plurality voting or SCF's based on majority voting, strategies not equivalent to truth-telling would also turn out to be protectively non-dominated. In particular, these rules would all violate USPA — one of the necessary conditions for direct implementability via protective equilibrium. However, given our assumptions, there does exist a non-empty class of social choice functions (containing some forms of 'voting by veto'), under which individuals would always use their truthful strategies.

Of course, this result is based on the extreme assumption of complete ignorance about other individual's preferences. Much of the literature on implementation makes the polar assumption of full information. It seems pointless to argue about the relative merits of these assumptions. Surely, 'reality' lies in between. We hope, however, that the present result clarifies the type of framework in which possibility results on non-manipulability and/or implementability can be proved. It would also seem worthwhile to introduce strategic equilibrium concepts based on 'partial information'. A particularly interesting framework would seem to be one where agents choose their strategies, given full knowledge of other agents' *preferences*, but only partial knowledge of their *strategic behaviour*.

Appendix

We prove the Lemma by several examples.

Lemma. BE, LCI and USPA are logically independent.

Proof. We show that the following statements can be true:

(i) USPA is satisfied, but neither LCI nor BE.

(ii) LCI is satisfied, but neither USPA nor BE.
(iii) BE is statisfied, but neither USPA nor LCI.
(iv) USPA and LCI are satisfied, but not BE.
(v) USPA and BE are satisfied, but not LCI.
(vi) LCI and BE are satisfied, but not USPA.

Let Q be an arbitrary asymmetric ordering over A.

(i) Let $A=\{a_1, a_2, a_3, a_4, a_5\}$, $I=\{1,2,3\}$. For any $\boldsymbol{P}$,
  (a) If P_1 and Q agree on $B(2, P_1)$, then $f(\boldsymbol{P})=\{x/x$ is the Q-maximal element in $A-B(2,P_1)-B(1,P_2)\}$.
  (b) If P_1 and Q do not agree on $B(2, P_1)$, then $f(P)=\{x/x$ is the Q-maximal element in $A-B(2,P_1)-B(1,P_3)\}$.

(ii) f_3 of section 4 satisfies LCI, but violates USPA and BE.

(iii) Let $A=\{a_1, a_2, a_3, a_4, a_5\}$, $I=\{1,2,3\}$. For any $\boldsymbol{P}$,
  (a) If P_1 and Q agree on $B(2,P_1)$, then $f(\boldsymbol{P})=\{x/x$ is P_2-minimal on $A-B(2,P_1)\}$.
  (b) If P_1 and Q agree on $B(2,P_1)$, then $f(\boldsymbol{P})=\{x/x$ is P_3-minimal on $A-B(2,P_1)\}$.

Thus, the order in which individual 1 vetoes two alternatives determines whether 2 or 3 is an 'anti-dictator' over the remaining pair of alternatives.
This violates USPA and LCI, but satisfies BE.

(iv) f_5 of section 4 satisfies USPA and LCI, but violates BE.

(v) f_4 of section 4 illustrates this case.

(vi) For all $\boldsymbol{P}$, let $f(\boldsymbol{P})=a_1(P_1)$, i.e., individual 1 is an 'anti-dictator'.

This satisfies LCI and BE, but violates USPA.

References

Barberá, S., 1980, Stable voting schemes, Journal of Economic Theory 28, 267–274.
Dutta, B. and P.K. Pattanaik, 1978, On nicely consistent voting schemes, Econometrica 46, 163–170.
Farquharson, R., Theory of voting (Yale University Press, New Haven, CT).
Fishburn, P.C. and S. Brams, 1978, Approval voting, American Political Science Review 72, 831–847.
Gibbard, A., 1973, Manipulation of voting schemes: A general result, Econometrica 41, 587–601.
Maskin, E., 1977, Nash equilibrium and welfare optimality, Mathematics of Operations Research, forthcoming.
Moulin, H., 1979, Dominance-solvable voting schemes, Econometrica 47, 1337–1351.
Moulin, H., 1980, Voting with proportional veto power, Econometrica, forthcoming.
Moulin, H., 1981, Prudence versus sophistication in voting strategy, Journal of Economic Theory 3, 398–412.

Moulin, H. and B. Peleg, 1982, Stability and implementation of effectivity functions, Journal of Mathematical Economics.
Muller, E. and M.A. Satterthwaite, 1977, The equivalence of strong positive association and strategy proofness, Journal of Economic Theory 14, 412–418.
Mueller, D.C., 1978, Voting by veto, Journal of Public Economics 10, 57–75.
Satterthwaite, M.A., 1975, Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions, Journal of Economic Theory 10, 187–217.
Sengupta, M., 1978, On a difficulty in the analysis of strategic voting, Econometrica 46, 331–344.