Chapter 8
Introduction to Part II

8.1 Motivation and summary

In this chapter we first recall the Gibbard-Satterthwaite Theorem and review some of its implications. This is done in Section 8.2. The rest of the chapter is devoted to consideration of the problem of preference distortion as a consequence of manipulation of non-dictatorial voting rules. First we observe that the Gibbard-Satterthwaite Theorem does not tell us whether or not the sincere outcome is obtained after manipulation. It may be the case that strategic voting leads to an equilibrium of which the outcome is the sincere outcome. In that case, the result of voting by a secret ballot would be indistinguishable from that of sincere voting. Indeed, in Section 8.3 we define exactly and strongly consistent social choice functions. Such social choice functions have for each profile of (true) preferences a strong (Nash) equilibrium that yields the sincere outcome. This class of social choice functions is the main topic of Chapters 9–11, in which we present several existence and characterization theorems. In particular, in Chapter 11 we extend some of the results of Chapters 9 and 10 to voting games with a continuum of voters.

In Section 8.4 we very briefly discuss voting on restricted domains for which the manipulation problem is eliminated, and mention a few references in order to direct the reader to some of the main results in this area.

We conclude the chapter with a discussion of equilibrium with threats, following an idea of Pattanaik (1976), which presents another way to obtain the sincere outcome if manipulation is possible. We construct non-dictatorial social choice functions with the property that sincere voting is always an equilibrium with threats.

B. Peleg and H. Peters, *Strategic Social Choice*, Studies in Choice and Welfare,
DOI 10.1007/978-3-642-13875-1_8, © Springer-Verlag Berlin Heidelberg 2010

8.2 The Gibbard-Satterthwaite Theorem and its implications

Let A be a set of m alternatives, $m \geq 3$, and let $N = \{1, \ldots, n\}$ be a set of voters. Let L denote the set of all linear orderings (strict preferences) on A, that is, the set of all transitive, reflexive, antisymmetric and complete binary relations on A. A *social choice function* (SCF) is a map $F : L^N \to A$. An SCF F is *non-manipulable* (or *strategy-proof*) if for each profile of preferences $R^N \in L^N$ the strategy-profile R^N itself is a Nash equilibrium in the game (F, R^N). Thus, if F is *manipulable* (not non-manipulable) then there exist $R_0^N \in L^N$, $i \in N$, and $Q^i \in L$ such that $F(R_0^{N \setminus \{i\}}, Q^i)\, R_0^i\, F(R_0^N)$, and $F(R_0^{N \setminus \{i\}}, Q^i) \neq F(R_0^N)$. In this case, R_0^N is a situation in which player i has an incentive to misrepresent his preference – 'play' Q^i instead of his true preference R_0^i. In slightly different words, non-manipulability of a social choice function means that for every voter (player) reporting (playing) his true preference is a weakly dominant strategy in every situation, i.e., every game (F, R^N).

For a social choice function $F : L^N \to A$ let the *range* of F, A^*, be defined by

$$A^* = \{x \in A \mid \ x = F(R^N) \text{ for some } R^N \in L^N\}.$$

A player $d \in N$ is a *dictator* of F if $F(R^N)\, R^d\, x$ for every $R^N \in L^N$ and $x \in A^*$. The SCF F is *dictatorial* if it has a dictator. A fundamental result of Gibbard (1973) and Satterthwaite (1975) is the following theorem.[1]

Theorem 8.2.1. *If a social choice function F is non-manipulable and $|A^*| \geq 3$, then F is dictatorial.*

Thus, if a non-dictatorial social choice function F has full range ($A = A^*$) then it must be manipulable. This implies that most social choice functions based on voting procedures in every-day use, like choice by plurality voting, Borda count, and approval voting, are manipulable. A natural question is to which extent manipulability of a social choice function is a drawback. After all, a voting game (F, R^N) is just a strategic game like many other every-day games (auctions, oligopolies, etc.) and so 'strategic' behavior is 'all in the game'. Nevertheless, manipulating behavior in strategic voting situations has some disturbing consequences, as pointed out by many authors (e.g., recently, Feldman and Serrano, 2005). Here, we list what we think are its main drawbacks.

First, a specific social choice function may have been adopted because of certain appealing properties, but these may be lost due to manipulation. For example, a social choice function based on plurality voting is Paretian. However, games associated with it may have equilibria resulting in outcomes that are not Pareto optimal. Cf. Feldman and Serrano (2005).

[1] For a proof see Chap. 11 in Peters (2008).

Second, manipulation may be objectionable on ethical grounds. Specifically, a manipulating voter may benefit at the expense of others who do vote truthfully.

Third, by manipulating behavior, especially of large groups of voters, the outcome of the voting procedure may be very far from the *sincere outcome*, i.e., the outcome corresponding to the profile of true preferences. Typically, for instance, in a Parliamentary democracy, voters may not vote for a small but favored political party if that party is unlikely to be a member of the government that is formed on the basis of the national election. (More formally, assuming that voters play an equilibrium in some equilibrium correspondence EQ, the actual voting correspondence changes from F to $F \circ EQ$.) On the basis of this argument, Feldman and Serrano (2005) question the legitimacy of the voting outcome if the voting procedure is manipulable.

Fourth, an important consequence of non-manipulability is that it requires only each voter's knowledge of his own preference. Thus, it makes the act of voting simple and reliable. This feature is lost under manipulability. On the other hand, this argument can also be considered to render the possibility of manipulation less harmful, since the cost of manipulation – e.g., to acquire the necessary information about the preferences and voting behavior of others – may prevent voters from actually manipulating.

In the next section we propose a weakening of the non-manipulability condition that takes away many of these drawbacks, since it results in the sincere outcome.

8.3 Exactly and strongly consistent social choice functions

In order to avoid distortion of the voting outcome implied by the Gibbard-Satterthwaite Theorem we shall weaken the non-manipulability requirement in the remainder of this book in such a way that (i) nevertheless the sincere outcome can result and (ii) this outcome can result under a strong stability condition. More precisely, we impose that the sincere outcome is always a strong (Nash) equilibrium outcome of the voting game under consideration. In a strong equilibrium (formally introduced in Definition 5.2.1) not only single players but also coalitions cannot gain by (joint) deviations. This route was first suggested in Peleg (1978a).

Definition 8.3.1. The social choice function $F : L^N \to A$ is *exactly and strongly consistent* (ESC) if for every $R^N \in L^N$ there exists a strong equilibrium Q^N of the game (F, R^N) such that $F(Q^N) = F(R^N)$.

An ESC social choice function trivially exists – take a constant social choice function, assigning a fixed alternative to any preference profile. In interesting cases, however, social choice functions are surjective, so their range is A. If

F is a surjective ESC social choice function then, in particular, the game form F is a strongly consistent representation of the effectivity function E^F associated with (the game form) F (see Chapter 5). This implies, in turn, that $F(Q^N)$ is an element of the core $C(E^F, R^N)$ for every strong equilibrium Q^N of the game (F, R^N) – see Proposition 5.2.6. It follows that the sincere outcome $F(R^N)$ of the voting game (F, R^N) is in the core of (E^F, R^N) and, in particular, Pareto optimal. More generally, Pareto optimality is maintained if we assume that voters play a strong equilibrium.

Moreover, under exact and strong consistency and assuming that voters play a strong equilibrium it is at least possible that the sincere outcome results. If voters have a more or less accurate conjecture about what the sincere outcome is, then a strong equilibrium Q^N resulting in the sincere outcome may become a focal point in the sense of Schelling (1960) – perhaps because of ethical considerations. This may alleviate the objection of political illegitimacy of the voting procedure as mentioned in the preceding section.

As a final note, we mention that if F is a surjective non-dictatorial ESC social choice function, then there exist an $R^N \in L^N$ and a strong equilibrium Q^N of the game (F, R^N) such that $F(Q^N) \neq F(R^N)$. Indeed, if not, then F as a game form *implements* the SCF F in strong equilibrium: all strong equilibria of (F, R^N) result in the outcome $F(R^N)$. This implies in particular that F is Maskin monotonic (see Remark 3.7.3 for the definition of Maskin monotonicity; the implication follows, e.g., by Lemma 6.5.1 in Peleg, 1984). This, in turn, implies that F is dictatorial by the Muller-Satterthwaite Theorem (see Muller and Satterthwaite, 1977). In other words, under non-dictatorship we cannot have that every strong equilibrium always results in the sincere outcome.

8.4 Strategyproofness and restricted preferences

Following the works of Gibbard (1973) and Satterthwaite (1975) there is a large strand of literature trying to avoid the consequences of the Gibbard-Satterthwaite Theorem. Most of this literature concentrates on what is often termed the *universal domain assumption* in this theorem, which means that all (strict) preferences – both as true preferences and as reported preferences – are allowed. Dropping this assumption is usually referred to as the *restricted domain* approach.

For a social choice function F one may consider the set

$$\text{Sp}(F) = \{R^N \in L^N \mid R^N \text{ is a Nash equilibrium of } (F, R^N)\},$$

and then F is strategy-proof on the domain $\text{Sp}(F)$. The implicit assumption here is that the true preference profiles are from this domain, and that may be a strong assumption. Moreover, it may be quite difficult to compute or characterize the domain $\text{Sp}(F)$.

Most results in this area are less ambitious. For instance, it was already known from Black (1948), Arrow (1951, 1963), or Dummet and Farquharson (1961) that generalized majority rule is strategy-proof when both the true and the reported preference profiles are restricted to be single-peaked. (A profile is single-peaked if there exists an ordering of the alternatives along which each individual preference is unimodal.) Blin and Satterthwaite (1976) show that this result no longer holds if the reported preference profiles are allowed to be more general, even if the true preferences are single-peaked. They do this by considering a social choice function that picks the Condorcet winner (i.e., an alternative that beats all other alternatives in pairwise comparison) if there is one, and otherwise picks the alternative with maximal Borda count.[2]

A well-known example of the converse phenomenon is approval voting, which – under some conditions on extensions of preferences from alternatives to sets of alternatives – is strategy-proof when both reported and true preferences are dichotomous: that is, each individual approves of a set B and disproves of the complement (see Brams and Fishburn, 1983). In this case, strategy-proofness is lost if the true preferences can be more refined (see Roy, Peters, and Storcken, 2009).

Most results on restricted domains therefore assume that one and the same restriction applies to both the true and the reported preferences. For instance, Moulin (1980) characterized all Paretian, anonymous and strategy-proof social choice functions for a class of single-peaked preferences on the real line. For an introduction to strategy-proof social choice functions see Barberà (2001).

A different approach to the implication of the Gibbard-Satterthwaite Theorem was initiated by Kelly (1988). Accepting this implication, one may look for social choice functions that are in some way minimally manipulable, e.g., in terms of numbers of manipulable profiles[3]. See Maus, Peters, and Storcken (2007) for a recent overview.

8.5 Equilibrium with threats

In this section, following an idea of Pattanaik (1976), we define equilibrium with threats in game forms. We then show that there exist non-dictatorial social choice functions with the property that sincere voting is always an

[2] The Borda rule attaches a score of m points to an individual's best alternative, $m-1$ points to the second best, ..., and 1 point to the worst alternative. The Borda count is obtained by summing over all individuals. At this introductory level we ignore the possibility of multiple winners.

[3] That is, social choice functions for which $\text{Sp}(F)$ has maximal cardinality.

equilibrium with threats. This provides another way to cope with the negative implication of the Gibbard-Satterthwaite Theorem.

As before let $N = \{1, \ldots, n\}$ be a set of voters and let A be a set of m alternatives, where $n \geq 2$ and $m \geq 3$. Let $\Gamma = (\Sigma^1, \ldots, \Sigma^n; g; A)$ be a game form with surjective outcome function, let $R^N \in L^N$, and let $\sigma \in \Sigma = \prod_{i \in N} \Sigma^i$. A *threat* of a coalition $S \in P_0(N)$ against σ is a strategy profile $\mu^S \in \Sigma^S$ such that

$$g(\mu^S, \sigma^{N \setminus S})\, R^i\, g(\sigma) \text{ for all } i \in S, \text{ and } g(\mu^S, \sigma^{N \setminus S}) \neq g(\sigma).$$

A *counter-threat* to μ^S is a strategy-profile $\mu^{N \setminus S} \in \Sigma^{N \setminus S}$ such that $g(\sigma)\, R^i\, g(\mu^S,\, \mu^{N \setminus S})$ for some $i \in S$. The profile σ is an *equilibrium with threats* in (Γ, R^N) if to each threat against σ there exists a counter-threat. In such a strategy profile, for each deviation of a coalition S that makes all its members strictly better off, there is a deviation by the complement of S such that at least one member of S again prefers the original outcome.

We start our discussion of equilibria with threats with the following observation.

Lemma 8.5.1. *Let $\Gamma = (\Sigma^1, \ldots, \Sigma^n; g; A)$ be a game form with surjective outcome function, let $R^N \in L^N$, and let $\sigma \in \Sigma$. Then σ is an equilibrium with threats in (Γ, R^N) if and only if $g(\sigma) \in C(E^\Gamma, R^N)$.*

Proof. (i) Suppose that σ is an equilibrium with threats. If $g(\sigma) \notin C(E^\Gamma, R^N)$, then there exists $S \in P_0(N)$ and $B \in E^\Gamma(S)$ such that $B\, R^S\, g(\sigma)$ and $g(\sigma) \notin B$. As $B \in E^\Gamma(S)$ there exists $\mu_0^S \in \Sigma^S$ such that for all $\mu^{N \setminus S} \in \Sigma^{N \setminus S}$ we have $g(\mu_0^S, \mu^{N \setminus S}) \in B$. Thus, μ_0^S is a threat against σ to which there is no counter-threat, a contradiction.

(ii) Assume that $g(\sigma) \in C(E^\Gamma, R^N)$. If σ is not an equilibrium with threats, then there exists $S \in P_0(N)$ and $\mu_0^S \in \Sigma^S$ such that $g(\mu_0^S, \mu^{N \setminus S})\, R^i\, g(\sigma)$ and $g(\mu_0^S, \mu^{N \setminus S}) \neq g(\sigma)$ for all $i \in S$ and $\mu^{N \setminus S} \in \Sigma^{N \setminus S}$. Let $B = \{g(\mu_0^S, \mu^{N \setminus S}) \mid \mu^{N \setminus S} \in \Sigma^{N \setminus S}\}$. Then $B \in E^\Gamma(S)$ and $B\, R^S\, g(\sigma)$ with $g(\sigma) \notin B$, a contradiction. □

We now prove the claim that we made at the beginning of this section. We define a *social choice correspondence* (SCC) as a map $H : L^N \to P_0(A)$. An SCF F can be seen as a special case of an SCC by setting $H_F(R^N) = \{F(R^N)\}$ for each $R^N \in L^N$. Call an SCC H *citizen sovereign* if for each $a \in A$ there is an $R^N \in L^N$ with $H(R^N) = \{a\}$. With a citizen sovereign SCC H we can associate an effectivity function E^H by $E^H(\emptyset) = \emptyset$ and for all $S \in P_0(N)$ and $B \in P_0(A)$,

$B \in E^H(S) \Leftrightarrow$

there is $R^S \in L^S$ with $H(R^S, Q^{N \setminus S}) \subseteq B$ for all $Q^{N \setminus S} \in L^{N \setminus S}$.

Clearly, E^H is superadditive.

Theorem 8.5.2. *There exists a nondictatorial social choice function F with the property that sincere voting is always an equilibrium with threats, i.e., R^N is an equilibrium with threats in the game (F, R^N) for all $R^N \in L^N$.*

Proof. Let E be a(n arbitrary) non-dictatorial, maximal and stable effectivity function. Define $F : L^N \to A$ by choosing $F(R^N) \in C(E, R^N)$ for every $R^N \in L^N$.

Observe that if $B \in P_0(A)$ and $S \in P_0(N)$ with $B \in E(S)$, then for $R^S \in L^S$ with $B\, R^S\, A \setminus B$ we have $C(E, (R^S, Q^{N\setminus S})) \subseteq B$ for all $Q^{N\setminus S} \in L^{N\setminus S}$, so that $B \in E^{C(E,\cdot)}(S)$, where $E^{C(E,\cdot)}$ is the effectivity function associated with the (citizen sovereign) social choice correspondence $C(E, \cdot)$. This implies, in turn, that $B \in E^F(S)$ (see Remark 9.3.3 in the next chapter). Hence, $E(S) \subseteq E^F(S)$ for all $S \in P_0(N)$, and since E is maximal and E^F superadditive, this implies $E = E^F$ by Lemma 9.3.1 in the next chapter.

Thus, $F(R^N) \in C(E^F, R^N)$ for every $R^N \in L^N$, and the proof is complete by Lemma 8.5.1. □

We conclude this section by observing that for an exactly and strongly consistent social choice function F the true preference profile R^N is always an equilibrium with threats of the game (F, R^N). This follows from the remarks in Section 8.3, in particular from the observation that $F(R^N) \in C(E^F, R^N)$, and Lemma 8.5.1.

8.6 Notes and comments

The discussion in this chapter has benefitted from Peleg (1984), besides from the references in the text. Section 8.5 is based on Peleg and Procaccia (2007).