23 Mechanism Design

Our discussion of social choice has so far ignored strategic considerations. Mechanism design augments social choice with game-theoretic reasoning, and effectively tries to construct games which yield a certain desirable outcome in equilibrium.

23.1 Strategic Manipulation

Again consider the situation of Figure 22.2, where we predicted that the use of plurality would result in the selection of alternative a. This prediction ignores, however, that voters of the third type have an incentive to misrepresent their preferences and claim that they prefer c over b: assuming that ties are broken in favor of c, only a single voter of the third type would have to change its reported preferences in this way to ensure that c is selected instead of a, an outcome this voter prefers. A similar problem exists with STV, where voters of the first type could benefit by pretending that their most preferred alternative is b, with the goal of having this alternative selected instead of their least preferred alternative, c. More generally, we say that SCF f is *manipulable* if there exist $i \in N$, $\succ \in L(A)^n$, and $\succ_i' \in L(A)$ such that $f((\succ_{-i}, \succ_i')) \succ_i f(\succ)$, where $(\succ_{-i}, \succ_i') = (\succ_1, \ldots, \succ_{i-1}, \succ_i', \succ_{i+1}, \succ_n)$ is the preference profile obtained by replacing voter i's preference order in $\succ$ by $\succ_i'$. SCF f is called *strategyproof* if it is not manipulable.

There are two obvious way to achieve strategyproofness: choosing an alternative based on the preferences of a single voter, or ignoring all but two alternatives and using majority rule to choose between these two. The first case corresponds to a dictatorship, the second to an SCF that is not surjective in the sense that some alternatives never get chosen. It turns out that these trivial cases are in fact the only SCFs that are strategyproof. Formally, SCF f is dictatorial if there exists $i \in N$ such that for all $\succ \in L(A)^n$ and $a \in A \setminus \{f(\succ)\}$, $f(\succ) \succ_i a$. SCF f is surjective if for all $a \in A$, there exists $\succ \in L(A)^n$ such that $f(\succ) = a$.

THEOREM 23.1 (Gibbard, 1973; Satterthwaite, 1975). *Consider an SCF $f : L(A)^n \to A$, where $|A| \geqslant 3$. If f is surjective and strategyproof, then it is dictatorial.*

We need two lemmas. The first lemma states that a strategyproof SCF is monotone in the sense that the selected alternative does not change as long as all alternatives ranked below it are still ranked below it for all voters.

LEMMA 23.2. *Let f be a strategyproof SCF, $\succ \in L(A)^n$ with $f(\succ) = a$. Then, $f(\succ') = a$ for every $\succ' \in L(A)^n$ such that for all $i \in N$ and $b \in A \setminus \{a\}$, $a \succ_i' b$ if $a \succ_i b$.*

Proof. We start from $\succ$ and change the preferences of one voter at a time until we get to $\succ'$, showing that the chosen alternative remains the same in every step. Let

$b = f(\succ_1', \succ_{-1})$. By strategyproofness, $a \succeq_1 b$, and thus $a \succeq_1' b$ by assumption. Also by strategyproofness, $b \succeq_1' a$, and thus $a = b$. The claim now follows by repeating the same argument for the remaining voters. □

The second lemma states that the alternative selected by a surjective and strategyproof SCF must be Pareto optimal.

LEMMA 23.3. *Let* f *be a surjective and strategyproof SCF,* $a, b \in A$*, and* $\succ \in L(A)^n$ *such that* $a \succ_i b$ *for all* $i \in N$*. Then,* $f(\succ) \neq b$*.*

Proof. Assume for contradiction that $f(\succ) = b$. By surjectivity, there exists $\succ' \in L(A)^n$ such that $f(\succ') = a$. Let $\succ'' \in L(A)^n$ be a preference profile such that for all $i \in N$

$$a \succ_i'' b \succ_i'' x$$

for all $x \in A \setminus \{a, b\}$. Then, $x \succ_i b$ whenever $x \succ_i'' b$ for some $i \in N$ and $x \in A \setminus \{b\}$, and $x \succ_i' a$ whenever $x \succ_i'' a$ for some $i \in N$ and $x \in A \setminus \{a\}$. Thus, by Lemma 23.2, $f(\succ'') = f(\succ) = b$ and $f(\succ'') = f(\succ') = a$, a contradiction. □

Proof of Theorem 23.1. We first prove the theorem for $n = 2$ and then perform an induction on n.

Let $a, b \in A$ with $a \neq b$ and consider $\succ \in L(A)^2$ such that

$$a \succ_1 b \succ_1 x \qquad \text{and} \qquad b \succ_2 a \succ_2 x$$

for all $x \in A \setminus \{a, b\}$. Then, by Lemma 23.3, $f(\succ) \in \{a, b\}$.

Suppose that $f(\succ) = a$, and let $\succ' \in L(A)^2$ such that

$$a \succ_1' b \succ_1' x \qquad \text{and} \qquad b \succ_2' x \succ_2' a$$

for all $x \in A \setminus \{a, b\}$. Then, $f(\succ') = a$, since $f(\succ') \in \{a, b\}$ by Lemma 23.3 and $f(\succ') \neq b$ by strategyproofness. Lemma 23.2 now implies that f selects alternative a for any preference profile in which voter 1 ranks alternative a first.

By repeating the above analysis for every pair of distinct alternatives in A, we obtain two sets $A_1, A_2 \subseteq A$ such that A_i is the set of alternatives that are selected for every preference profile in which voter $i \in \{1, 2\}$ ranks them first. Let $A_3 = A \setminus (A_1 \cup A_2)$, and observe that $|A_3| \leqslant 1$: otherwise we would have performed the above analysis for two elements in A_3, which would place one of these elements in A_1 or A_2 and thus not in A_3.

Now observe that $|A| \geqslant 3$ and $|A_3| \leqslant 1$, so $|A_1 \cup A_2| \geqslant 2$. Moreover, for $x, y \in A$ with $x \neq y$, it cannot be the case that $x \in A_1$ and $y \in A_2$, because this would lead to a contradiction when voter 1 ranks x first and voter 2 ranks y first. Since $a \in A_1$, it follows that $A_1 \cap A_2 = \emptyset$ and thus that $A_2 = \emptyset$. It finally follows that $A_3 = \emptyset$: otherwise we could repeat the above analysis for $c \in A_3$ and $\succ'' \in L(A)^2$ with

$$c \succ_1'' a \succ_1'' x \qquad \text{and} \qquad a \succ_2'' c \succ_2'' x$$

for all $x \in A \setminus \{a, c\}$, and conclude that $c \in A_1$ or $a \in A_2$, a contradiction. It follows that $A_1 = A$, so voter 1 is a dictator.

Now we assume that the statement of the theorem holds for n voters and prove that it also holds for $n+1$ voters. Consider a surjective and strategyproof SCF $f : L(A)^{n+1} \to A$, and define $g : L(A)^2 \to A$ by letting

$$g(\succ_1, \succ_2) = f(\succ_1, \succ_2, \ldots, \succ_2)$$

for all $\succ_1, \succ_2 \in L(A)$.

Since f is surjective and strategyproof, and by Lemma 23.3, g is surjective as well. Assume for contradiction that g is not strategyproof. By strategyproofness of f, the manipulator must be voter 2, so there must exist $\succ_1, \succ_2, \succ_2' \in L(A)$ and $a, b \in A$ such that $g(\succ_1, \succ_2) = a$, $g(\succ_1, \succ_2') = b$, and $b \succ_2 a$. For $k = 0, \ldots, n$, let $\succ^k = (\succ_1, \succ_2', \ldots, \succ_2', \succ_2, \ldots, \succ_2) \in L(A)^{n+1}$ be the preference profile where k voters have preference order $\succ_2'$ and $n-k$ voters have preference order $\succ_2$, and let $a^k = f(\succ^k)$. Since $a^n = b \succ_2 a = a^0$, it must be the case that $a^{k+1} \succ_2 a^k$ for some k with $0 \leqslant k < n$, which means that f is manipulable, a contradiction. It follows that g is strategyproof, and therefore dictatorial.

If the dictator for g is voter 1, then by Lemma 23.2 voter 1 must also be a dictator for f. Assume instead that the dictator for g is voter 2, and let $h : L(A)^n \to A$ be given by

$$h(\succ_2, \ldots, \succ_{n+1}) = f(\succ_1^*, \succ_2, \ldots, \succ_{n+1})$$

for an arbitrary $\succ_1^* \in L(A)$. Then, h is strategyproof by strategyproofness of f, and surjective because voter 2 is a dictator for g. Therefore, by the induction hypothesis, h is dictatorial.

Assume without loss of generality that the dictator for h is voter 2, and let $e : L(A)^2 \to A$ be given by

$$e(\succ_1, \succ_2) = f(\succ_1, \succ_2, \succ_3^*, \ldots, \succ_{n+1}^*)$$

for arbitrary $\succ_3^*, \ldots, \succ_{n+1}^* \in L(A)$. Then e is strategyproof and surjective, and hence dictatorial. In fact, the dictator for e must be voter 2, because voter 1 is not a dictator for g and thus cannot be a dictator for e. Since $\succ_i^*$ for $i = 1, 3, \ldots, n+1$ was chosen arbitrarily, it follows that voter 2 is a dictator for f. □

23.2 Implementation of Social Choice Functions

A *mechanism design problem*, or *game form*, is given by a set A of *alternatives* and a set $N = \{1, \ldots, n\}$ of *agents*, each with a set Θ_i of possible *types* and a *utility function* $u_i : A \times \Theta_i \to \mathbb{R}$. Note that a game form and a *type profile* $\theta \in \Theta = \times_{i \in N} \Theta_i$ together induce a normal-form game. A *mechanism* is given by a message space Σ_i for agent i and an outcome function $g : \times_{i \in N} \Sigma_i \to A$. A mechanism is called *direct* if the

agents directly report their type to the mechanism, i.e., if $\Sigma_i = \Theta_i$ for all $i \in N$. The idea is that the agents send messages to the mechanism, providing information about their types, and depending on these messages the mechanism selects an alternative that optimizes some objective. The objective can be encoded by a social choice function.

Mechanism $M = ((\Sigma_i)_{i\in N}, g)$ is said to *implement* SCF $f : \times_{i\in N}\Theta_i \to A$ (in weakly dominant strategies) if there exist functions $s_i : \Theta_i \to \Sigma_i$ for $i \in N$ such that for every $\theta \in \Theta$, $g(s_1(\theta_1), \dots, s_n(\theta_n)) = f(\theta)$, and for all $i \in N$, $\theta_i \in \Theta_i$ and $\sigma \in \Sigma$, $u_i(g(s_i(\theta_i), \sigma_{-i}), \theta_i) \geqslant u_i(g(\sigma), \theta_i)$. An SCF is called *implementable* if it is implemented by some mechanism. A direct mechanism $M = ((\Theta_i)_{i\in N}, g)$ is called dominant strategy incentive compatible, or *strategyproof*, if for all $i \in N$, $\theta \in \Theta$, and $\theta_i' \in \Theta_i$, $u_i(g(\theta), \theta_i) \geqslant u_i(g(\theta_i', \theta_{-i}), \theta_i)$. The profile θ of true types is then also referred to as the truthful equilibrium of the mechanism.

It seems that in principle arbitrarily complicated mechanisms might be required to implement certain social choice functions. The following result implies that we can restrict our attention to strategyproof direct mechanisms.

THEOREM 23.4 (Revelation Principle). *A social choice function is implementable if and only if it is implemented in the truthful equilibrium of a strategyproof direct mechanism.*

Proof. The theorem follows by observing that the direct mechanism can simulate the equilibrium strategies of the agents. Let f be an implementable SCF. Then there exists a mechanism $((\Sigma_i)_{i\in N}, g)$ and functions $s_i : \Sigma_i \to \Theta_i$ for $i \in N$ such that for every $\theta \in \Theta$, $g(s_1(\theta_1), \dots, s_n(\theta_n)) = f(\theta)$, and for every $i \in N$, $\theta_i \in \Theta_i$ and $\sigma \in \Sigma$, $u_i(g(s_i(\theta_i), \sigma_{-i}), \theta_i) \geqslant u_i(g(\sigma), \theta_i)$. Define $h : \Theta \to A$ by letting $h(\theta) = g(s_1(\theta_1), \dots, s_n(\theta_n))$ for all $\theta \in \Theta$. Then, for every $\theta \in \Theta$, $h(\theta) = f(\theta)$, and for all $i \in N$, $\theta \in \Theta$, and $\theta_i' \in \Theta_i$,

$$\begin{aligned} u_i(h(\theta), \theta_i) &= u_i(g(s_1(\theta_1), \dots, s_n(\theta_n)), \theta_i) \\ &\geqslant u_i(g(s_1(\theta_1), \dots, s_{i-1}(\theta_{i-1}), s_i(\theta_i'), s_{i+1}(\theta_{i+1}), \dots, s_n(\theta_n)), \theta_i) \\ &= u_i(h(\theta_i', \theta_{-i}), \theta_i). \end{aligned}$$

This means that (Θ, h) is a strategyproof direct mechanism that implements f, and the claim follows. □

It should be noted that indirect mechanisms can still be useful in certain settings, for example to reduce the amount of information agents have to send to the mechanism, or the amount of computation the mechanism has to carry out.

Theorems 23.1 and 23.4 imply that only dictatorial social choice function are implementable when there are more than two alternatives and utility functions u_i can be arbitrary. In the next lecture we will look at an interesting special case where this impossibility result can be circumvented.