Abstract

The theory of Mechanism Design intends to find ways to implement social choice functions. That is, to characterize rules such that, for any profile of actual preferences, game solutions support the outcomes of those functions. Maskin, in his seminal 1977 paper showed that game forms provide a natural framework to analyze this problem.

We focus here on game forms in which the strategies are declarations of preferences over the outcomes. These game forms are called direct mechanisms. On the space of this kind of game forms we postulate an operation, that given a direct mechanism provides other mechanisms (not necessarily a single one), by optimizing the preferences of the agents. A fixed point under this operation is shown to be not strategically manipulable by individual agents. We characterize this fixed point in terms of one of the main impossibility theorems in Social Choice theory, Gibbard-Satterthwaite’s, to show that it is dictatorial, i.e. it implements the most preferred outcomes of a single agent.

J.E.L. Codes: C6, D7

1 Introduction

The theory of Mechanism Design¹, which began with Vickrey’s (1961) analysis of second-price auctions, intends to determine games in which the players, by playing according to their own preferences, yield in equilibria the alternatives that a social planner wants. The goals of the planner are completely described by a social choice function that provides, for each profile of preferences, a single

outcome. The reason why a game must be found is that the planner does not know the actual types of the agents. The game provides a way to extract information about them.

While the class of possible games is huge, the Revelation Principle allows to cut it down substantively. This principle indicates that, whatever the kind of game solution that is intended to support the outcomes, if there exists a mechanism that accomplishes this, there exists a direct mechanism that yields the same result. In this mechanism the strategy sets consist just of the possible types that agents can declare if asked directly about that. Therefore, if no such ‘direct’ mechanism implements a given social choice function, no other mechanism does so. This means that the analysis of implementability can be reduced to the space of direct mechanisms without loss of generality.²

Following Maskin (1977), mechanisms are known as game forms, i.e. they are defined by the strategy sets of the agents and an outcome function, that for every profile of strategies yields an element in the space of alternatives. The planner knows only this much. If she knew the full game, including the valuations of the agents over the alternatives, the problem of design would be trivial. In the case of direct mechanisms, if we identify the type of an agent with her preferences, the outcome function should be just the social choice function that the game forms intends to implement.

In this paper we consider a correspondence of the space of direct mechanisms on itself, such that to each direct game form it assigns some others. All of them capture the outcomes that may arise if agents maximize their preferences over their option sets left by the initial mechanism. That is, the ‘options’ that the declared preferences of the agents leave the others to choose from. Fixed points of this correspondence are shown to be non-manipulable, i.e. no single agent, by misrepresenting her preferences, can ensure a more preferred outcome.

In turn, to be non-manipulable means, via Gibbard-Sattertwhite’s theorem, that the direct mechanism is dictatorial. We will prove this claim, following the lines suggested in Barberà and Peleg (1990), making an extensive use of the notion of option set, which becomes the unifying theme of our entire argument.

This paper is organized as follows. Section 2 introduces the main definitions that will be used in the text. Section 3 presents the proofs of our main claims. Finally, section 4 concludes and briefly discusses the prospects for future work.

2 The Setting

We will consider here a society of \( n \) agents, each one endowed with a preference ordering, \( \preceq_i \), over a finite set of alternatives, \( X \), with \( |X| > 2 \). We denote by \( \mathcal{R} \) the class of complete and transitive orderings over \( X \). We denote by \( \tau(\preceq_i) \in X \) the set of top elements of \( \preceq_i \), i.e. for all \( x \in X \), \( x \preceq_i y \) for any \( y \in \tau(\preceq_i) \). If the preference of \( y \) over \( x \) is strict we denote this by \( x \prec y \).

²Despite this, if other considerations, like the complexity of manipulating outcomes are taken into account, non-direct mechanisms may provide better candidates for implementing social choice functions (Mount and Reiter, 2002; Van Zandt, 2007).
A social choice function $f : \mathcal{R}^n \to X$ assigns to each profile of preferences $\preceq = (\preceq_1, \ldots, \preceq_n)$ a single $x \in X$. The range of $f$ is denoted $r_f$.

To implement $f$ a game form can be postulated, namely a $G = (\{S_i\}_{i=1}^n, \hat{f})$, with $\hat{f} : \prod_{i=1}^n S_i \to X$. Given a solution notion, i.e. a class $\mathcal{S} \subseteq \prod_{i=1}^n S_i$, we intend that for every $(s_1, \ldots, s_n) \in \mathcal{S}$, $\hat{f}(s_1, \ldots, s_n) = f(\preceq_1, \ldots, \preceq_n)$.

We restrict ourselves to the case in which each $S_i = \mathcal{R}$. If so, a game form $G$ is called a direct mechanism. The class of direct mechanisms for $n$ agents over the set of alternatives $X$ can be written in simplified form $\mathcal{F} = \{f|f : \mathcal{R}^n \to X\}$.

Notice that if a direct mechanism given by $\hat{f}$ implements a social choice function $f$, for any given $\mathcal{S} \subseteq \mathcal{R}^n$, $\hat{f}(\preceq_1, \ldots, \preceq_n) = f(\preceq_1, \ldots, \preceq_n)$ where $(\preceq_1, \ldots, \preceq_n) \in \mathcal{S}$ while $(\preceq_1, \ldots, \preceq_n)$ is the true profile of preferences. The implementation is said truthful if $(\preceq_1, \ldots, \preceq_n) \in \mathcal{S}$.

Our goal is to characterize direct mechanisms that ensure truthful implementations. In order to reach that let us introduce two notions that will be heavily used in our argumentation:

- For $\preceq_i \in \mathcal{R}$ and $Y \subseteq X$, the choice set of $\preceq_i$ in $Y$ is $C(\preceq_i, Y) = \{y \in Y : x \preceq_i y \text{ for all } x \in Y\}$.
- Given $\hat{f} : \mathcal{R}^n \to X$, and $(\preceq_i, \preceq_{-i}) \in \mathcal{R}^n$, for each $i$ the set of options left to $i$ by $\preceq_{-i}$ is $\sigma_i(\preceq_{-i}) = \{x \in X : \text{there exists } \preceq_i \text{ such that } \hat{f}(\preceq_i, \preceq_{-i}) = x\}$.

Consider the following correspondence, which we call choice-improving:

$$\phi : \mathcal{F} \to \mathcal{F}$$

defined for every $\hat{f} \in \mathcal{F}$ and every profile $\preceq \in \mathcal{R}^n$ as:

$$\phi(\hat{f})[\preceq_i, \preceq_{-i}] = \{x \in X | x \in \bigcap_{i=1}^n C(\preceq_i, \sigma_i(\preceq_{-i}))\}$$

That is, $\phi$ takes a direct mechanism $\hat{f}$ and returns all the direct mechanisms such that their ranges, for any possible profile of preferences, yield the alternatives that maximize the individual preferences over the options left by $\hat{f}$.

The main results of this paper, which involve the behavior of $\phi$ are given in terms of two possible properties of a direct mechanism:

- **non-manipulability**: for all $i$ it does not exists a pair of preferences $\preceq_i, \preceq'_i$ such that and for any $\preceq_{-i}$ $\hat{f}(\preceq_i, \preceq_{-i}) = x$, $\hat{f}(\preceq'_i, \preceq_{-i}) = y$ with $x \neq y$ and $x \preceq_i y$ or $y \preceq'_i x$.
- **non-dictatorship**: there is no $i$ such that for any profile $\preceq = (\preceq_1, \ldots, \preceq_i, \ldots, \preceq_n)$, $\hat{f}(\preceq) \in \tau(\preceq_i)$.

\[3\] We follow closely the presentation in Osborne and Rubinstein (1994). Despite this, we will not go further into the issue of the game theoretic solutions that may support the outcomes. Nevertheless, the interested reader may notice that the non-manipulability condition (see below) can be identified with a dominant strategies equilibrium.
Theorem 1 A direct mechanism \( \hat{f} \) is fixed point of \( \phi \) (\( \hat{f} \in \phi(\hat{f}) \)) if and only if \( \hat{f} \) is dictatorial.

In order to prove this claim we consider two lemmas

Lemma 1 A direct mechanism \( \hat{f} \) is fixed point of \( \phi \) (\( \hat{f} \in \phi(\hat{f}) \)) if and only if \( \hat{f} \) is non-manipulable.

Lemma 2 (Gibbard-Satterthwaite) If a direct mechanism \( \hat{f} \) is non-manipulable it is dictatorial.

3 Fixed-Points and Dictatorship

Proof of Lemma 1: \( \Rightarrow \) Assume that \( \hat{f} \) is a fixed point for \( \phi \). That is, \( \hat{f} \in \phi(\hat{f}) \).
Then \( \hat{f}(\preceq_i, \preceq_{-i}) = \{ x \in X | x \in \cap_{i=1}^n C(\preceq_i, \phi^t_i(\preceq_{-i})) \} \), i.e. for every profile \( \preceq \) and every \( i \), \( \hat{f}(\preceq) \in C(\preceq_i, \phi^t_i(\preceq_{-i})) \). This means that \( \hat{f}(\preceq'_i, \preceq_{-i}) \preceq_i \hat{f}(\preceq)_i \). That is, \( \hat{f} \) is non-manipulable.

\( \Leftarrow \) Assume that \( \hat{f} \) is non-manipulable. Then, for any \( \preceq \in \mathbb{R}^n \), any \( i \) and any alternative preference \( \preceq'_i \) we have that \( \hat{f}(\preceq'_i, \preceq_{-i}) \preceq_i \hat{f}(\preceq)_i \). That is, \( \hat{f}(\preceq)_i \in C(\preceq_i, \phi^t_i(\preceq_{-i})) \) (obviously \( \hat{f}(\preceq)_i \in o^t_i(\preceq_{-i}) \)) for every \( i \). Then, \( \hat{f}(\preceq)_i \in \cap_{i=1}^n C(\preceq_i, o^t_i(\preceq_{-i})) \), which means that \( \hat{f}(\preceq)_i \in \phi(\hat{f})[\preceq] \), i.e. \( \hat{f} \in \phi(\hat{f}) \).

To prove Lemma 2, we proceed proving five intermediary claims about any non-manipulable direct mechanism \( \hat{f} \):

Claim 1 For any \( \preceq_1, \ldots, \preceq_n \), \( \hat{f}(\preceq_1, \ldots, \preceq_n) \in \cap_i C(\preceq_i, o^t_i(\preceq_{-i})) \).

Proof: Already shown in the proof of Lemma 1. \( \Box \)

Claim 2 For every \( i \) and any \( \preceq_{-i} \), we have that \( C(\preceq_i, r_j) \in \cap_{i \neq j} o^t_j(\preceq_i, \preceq_{-i,j}) \), where \( r_j \) is the range of \( \hat{f} \).

Proof: Suppose not. That is, given \( x = C(\preceq_i, r_j) \) there exists \( j \neq i \) such that \( x \notin o^t_j(\preceq_i, \preceq_{-i,j}) \) for any \( \preceq_{-i} \). Since \( x \in r_j \), there exists \( \hat{x} \in \mathbb{R}^n \) such that \( \hat{f}(\hat{x}) = x \). On the other hand, consider \( y \in \cap_{j \neq i} o^t_j(\preceq_i, \preceq_{-i,j}) \). It follows that \( y \neq x \). Furthermore, \( y = \hat{f}(\hat{x}, \preceq_{-i}) \) since \( y \) belongs to every \( o^t_j(\preceq_i, \preceq_{-i,j}) \) for \( j \neq i \) and therefore each \( \preceq_j \) such that \( y = \hat{f}(\preceq_j, \preceq_{-i,j}) \) must be \( \preceq_j = \hat{x}_j \).

But then, since \( x \in C(\preceq_i, r_j) \), we would have that \( y \preceq_i x \), i.e. \( \hat{f}(\preceq_i, \preceq_{-i}) \preceq_i \hat{f}(\hat{x}, \preceq_{-i}) \). Contradiction, since \( \hat{f} \) is non-manipulable. \( \Box \)
Claim 3 For every $i$ and every $\preceq_{-i}$, if given $\preceq_i, \preceq'_i$, $C(\preceq_i, r_f) = C(\preceq'_i, r_f)$, then for any $j \neq i$, $o_f^j(\preceq_i, \preceq_{-\{i,j\}}) = o_f^j(\preceq'_i, \preceq_{-\{i,j\}})$.

Proof: Suppose that there exists $j$ such that $o_f^j(\preceq_i, \preceq_{-\{i,j\}}) \neq o_f^j(\preceq'_i, \preceq_{-\{i,j\}})$, while $C(\preceq_i, r_f) = C(\preceq'_i, r_f)$. Then, there exists $y \in o_f^j(\preceq_i, \preceq_{-\{i,j\}})$ such that $y \notin o_f^j(\preceq'_i, \preceq_{-\{i,j\}})$.

On the other hand, consider $x = C(\preceq_i, r_f)$. By Claim 2, $x \in o_f^j(\preceq_i, \preceq_{-i,j})$.

By the same token, since also $x = C(\preceq'_i, r_f)$, $x \in o_f^j(\preceq'_i, \preceq_{-i,j})$.

Let now $\tilde{z}_j$ be such that $z \preceq_{-j} \tilde{z}_j y$, for any $z \neq x, z \neq y$. Then, since $\hat{f}$ is non-manipulable and $x, y \in o_f^j(\preceq_i, \preceq_{-\{i,j\}})$, $\hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n) = x$ where $x = C(\tilde{z}_j, o_f^j(\preceq_i, \preceq_{-\{i,j\}}))$.

But then $\hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n)$ is not in $\preceq_i, \preceq_{-\{i,j\}}$, which contradicts that $\hat{f}$ is non-manipulable. \hfill \Box

Claim 4 For any $\preceq_i$, for every $j \neq i$ either $o_f^j(\preceq_i, \preceq_{-\{i,j\}}) = r_f$. Furthermore, in the latter case, there is no $k \neq j$ such that $o_f^k(\preceq_i, \preceq_{-\{i,k\}}) = r_f$ if $\tau(\preceq_j) \cap \tau(\preceq_k) = \emptyset$ for some $\preceq_j, \preceq_k$.

Proof: Suppose that there exists a $j \neq i$, a preference $\tilde{z}_i$ and $x, y, z \in r_f$ such that $x, y \in o_f^j(\tilde{z}_i, \preceq_{-\{i,j\}})$ while $z \notin o_f^j(\tilde{z}_i, \preceq_{-\{i,j\}})$. Furthermore, assume that $y \tilde{z}_i z$. Now consider $\tilde{z}_j$ such that $w \tilde{z}_j y \tilde{z}_j z$ for all $w \neq z, w \neq y$.

Then, $\hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n) = y$ because $y \in o_f^j(\tilde{z}_i, \preceq_{-\{i,j\}})$ and is more preferred than any other available option in $o_f^j(\tilde{z}_i, \preceq_{-\{i,j\}})$ (z is not in that options set).

But, if $\tilde{z}_i$ verifies that $z \in o_f^j(\tilde{z}_i, \preceq_{-\{i,j\}})$, $\hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n) = z$, since $z$ is the most preferred option under $\tilde{z}_j$. But then $\hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n) \tilde{z}_i \hat{f}(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n)$, contradicting that $\hat{f}$ is non-manipulable.

Finally, suppose that there exist $j, k$, $j \neq k$ such that $o_f^j(\preceq_i, \preceq_{-\{i,j\}}) = r_f = o_f^k(\preceq_i, \preceq_{-\{i,k\}})$, while $\tau(\preceq_j) \cap \tau(\preceq_k) = \emptyset$. This is always possible since $\preceq_j$ and $\preceq_k$ can be freely chosen from $\mathcal{R}$. Then, consider $x \in C(\preceq_j, r_f)$ and $y = C(\preceq_k, r_f)$. By Claim 1, $\hat{f}(x) = x$ and $\hat{f}(x) = y$ while $x \in \tau(\preceq_j)$ and
$y \in \tau(\preceq_k)$. But then $\hat{f}(\preceq) \in \emptyset$. Absurd. □

Finally, the following result closes the argument:

**Claim 5** For any $i$ and any $\preceq \in \mathcal{R}^n$, if either

- for every $j \neq i$, $|o^j_f(\preceq_i, \preceq_{-\{i,j\}})| = 1$,

  or.

- $o^i_f(\preceq_{-i}) = r_f$

then $\hat{f}$ is dictatorial.

**Proof:** If for every $j \neq i$, $o^j_f(\preceq_i, \preceq_{-\{i,j\}})$ is a singleton, recall that by Claim 2, $o^j_f(\preceq_i, \preceq_{-\{i,j\}}) = C(\preceq_i, r_f)$. That is,

$$\hat{f}(\preceq) \in \tau(\preceq_i),$$

i.e. $\hat{f}$ yields the preferences of $i$, for any preferences $i$ may have.

On the other hand, if $o^1_f(\preceq_{-i}) = r_f$, by Claim 1, $\hat{f}(\preceq) \in \cap_k C(\preceq_k, o^k_f(\preceq_-k))$. In particular, $\hat{f}(\preceq) \in C(\preceq_i, r_f)$. That is, $\hat{f}(\preceq) \in \tau(\preceq_i)$. This shows also that $\hat{f}$ is dictatorial. □

Therefore, Claims 4 and 5 show that if $\hat{f}$ is non-manipulable it is dictatorial.

## 4 Discussion

The main goal of this paper has been to characterize, in the class of direct mechanisms, the fixed points of a correspondence that to each such game form assigns other direct game forms that in some form “improve” upon the original one. More precisely, it provides those alternatives that maximize the declared preferences on the sets of options left by the declarations of all other agents. The fixed points are shown to be non-manipulable mechanisms, i.e. such that no individual finds it profitable to declare other preferences than her true ones.

In turn, we have proven the Gibbard-Satterthwaite theorem, following the suggested lines in Barberà and Peleg (1990). It indicates that non-manipulable direct mechanisms are dictatorial. While this is a well-known result, we present the proof here to show the importance of option sets, already present in the definition of the correspondence among direct mechanisms.

Matter for further research is to show whether the mere fact that a direct game form is a fixed point of the correspondence defined in this paper, by simple set-theoretic reasons yields that it is dictatorial, without having to go through Gibbard-Satterthwaite’s theorem.
References


