

14 Mechanism Design

The basic point with mechanism design is that it allows a distinction between the “underlying economic environment” and the “rules of the game”. We will take as given some set of possible “outcomes” (alternatives, allocations) and some preferences over these outcomes, but the “game” is initially unspecified.

There are essentially two distinct ways to apply the mechanism design techniques that we will discuss:

1. Some papers ask: what kind of allocation rules can be “implemented” in the sense that it is possible to construct a game so that the allocation rule under consideration is an equilibrium? This sort of an exercise may be thought of as studying the “feasible set” for some agent or group of agents who decides on the rules of the game. The most powerful results of this kind are results stating that various things are impossible, since then we know that no matter what the actual rules are (in many economic applications it’s not so clear what the right game is...think about bargaining for example) we know that certain things cannot happen.
2. The other sort of exercise asks: which is the best mechanism from the point of view of the mechanism designer. Sometimes, the designer is a profit maximizer, and in this case this is an obviously positive exercise. At other times, the objective function of the mechanism designer is some sort of welfare criterion, in which case the exercise may be thought of as an extension of social welfare maximization, with the distinction that informational asymmetries are explicitly taken into consideration. This is often referred to as “second best” welfare optimization. I try to avoid such language since “second best” often times refers to exercises where some obvious “policy variables” are unavailable for unexplained reasons (i.e., cannot tax this or do that...).

14.1 Definitions

Let

- $N = \{1, \dots, n\}$ be the set of agents.
- Θ^i denote player i 's *type space* (generic element $\theta^i \in \Theta^i$). We let Θ denote $\times_{i=1}^n \Theta^i$.

- A denote the set of possible outcomes (alternatives, allocations)

A mechanism, or game form, is then an “institutional structure” that assigns an alternative (or a probability distribution over alternatives) to actions within this structure. Formally:

Definition 1 A mechanism (game-form) is an object $\langle M, g \rangle$ where,

$M = \times_{i=1}^n M_i$ and M_i is the action space (message space) for player $i \in N$

$g : M \rightarrow \Delta(A)$, where $\Delta(A)$ denotes the set of probability distributions over the set of (physical) outcomes A .

The first thing to realize is that a mechanism together with utility functions over the alternatives and probability distributions over the type space induces a game of incomplete information.

Definition 2 The mechanism $\langle M, g \rangle$, together with preferences $u^i : A \times \Theta \rightarrow R$ for $i = 1, \dots, n$ and a prior distribution $p \in \Delta(\Theta)$ is said to induce a Bayesian game $\{M, \Theta, p, \tilde{u}\}$, where $\tilde{u} = (\tilde{u}^1, \dots, \tilde{u}^n)$ and $\tilde{u}^i : M \times \Theta \rightarrow R$ is defined as

$$\tilde{u}^i(m, \theta) = \sum_{a \in A} u^i(a, \theta) g(a, m)$$

Here $g(a, m)$ denotes the probability that the rule g assigns to alternatives a when agents actions (messages) are $m = (m^1, \dots, m^n)$. As with any other Bayesian game we can of course derive a standard (ex ante) normal form representation $G = (N, S, u)$ by defining $S_i = \{s_i : \Theta_i \rightarrow M_i\}$, $S = \times_i S_i$ and

$$\begin{aligned} u_i(s) &= \sum_{\theta} \tilde{u}^i(s(\theta), \theta) p(\theta) = \\ &= \sum_{\theta_i} \sum_{\theta_{-i}} \left[\sum_{a \in A} u^i(s_i(\theta_i), s_{-i}(\theta_{-i}), \theta) g(a, s_i(\theta_i), s_{-i}(\theta_{-i})) \right] p(\theta_i, \theta_{-i}) \end{aligned}$$

Example 1 (First price auction). In this case $A = \{i, p | i \in N \text{ and } p \in R_+\}$ with interpretation that i is the winner of the auction and p is the price paid. $M_i = R_+$ for each i and $g : M \rightarrow \Delta(A)$ is constructed as follows. For each $m \in M$, let $W = \{i \in N | i \in \arg \max_{j \in N} m_j\}$ and $p(m) = \max_{j \in N} m_j$. Then $g(m) = (g_1(m), g_2(m))$ where $g_1(m) \in \Delta(N)$ is a probability distribution over the set of agents such that

$$g_{1i}(m) = \begin{cases} \frac{1}{|W|} & \text{if } i \in W \\ 0 & \text{otherwise} \end{cases},$$

(i.e., the winner is picked as a random draw from the set of agents with the highest bids). Finally $g_2(m) = p(m)$. Clearly, in analysis we would also have to specify payoffs and priors, but the only point here is that a first price auction is an example of a mechanism.

Example 2 (Second price auction). As above, but $g_2(m) = p(m) = \max_{i \notin W} m_i$.

An equilibrium in a game, be it in dominant strategies, Nash equilibrium or any refinement consists of a *strategy profile*, which in this context means a map $s^* : \Theta \rightarrow M$. The *outcome* if $\theta \in \Theta$ is realized is then $g(s^*(\theta))$. Thus, in the case with Θ_i being the set of possible rankings of the alternatives in A the composition $g \circ s^* : \Theta \rightarrow A$ (or $\Delta(A)$) describes the map from individual preferences to physical outcomes, i.e. exactly the same kind of object as a social choice function. The following definition should then be rather natural.

Definition 3 (Weak Implementation) *The mechanism $\langle M, g \rangle$ implements the social choice function $f : \Theta \rightarrow A$ if there is an equilibrium s^* in the game form induced by $\langle M, g \rangle$ such that $f(\theta) = g(s^*(\theta))$ for all θ .*

This weak implementation criterion doesn't rule out some other equilibria s^{**} for which $f(\theta) \neq g(s^{**}(\theta))$. However, below we will show non-existence of mechanisms that cannot “be manipulated” that implement social choice rules and also discuss a few other negative results. Clearly, in this context, the weaker the notion of implementation, the stronger the result.

15 The Revelation Principle

In the literature almost all analysis is (for good reasons) confined to the case of *direct revelation mechanisms*:

Definition 4 *A direct revelation mechanism is a mechanism $\langle M, g \rangle$ where $M_i = \Theta_i$ for all $i \in N$.*

This means that mathematically, the true type space and the set of possible messages is the same. This creates some confusion for some people and it is important to keep in mind that there is a fundamental difference in the *notions* of type space and message space: a type is what an agent is born with, while, even if we consider direct mechanism, there is nothing that guarantees that

an agent actually announces the truth. For clarity we will sometimes use notation $\widehat{\theta}_i$ to denote a reported type.

The following result is extremely important, but often times misunderstood:

Theorem 1 (*Dominant Strategy Revelation Principle*) Fix p, Θ_i and $u^i : A \times \Theta \rightarrow R$ for $i = 1, \dots, n$. Suppose that s^* is a dominant strategy equilibrium in the game induced by $\langle M, g \rangle$. Then truthful reporting of type (i.e. $\widetilde{\theta}_i(\theta_i) = \theta_i$ for all $i \in N$ and $\theta_i \in \Theta_i$) is a dominant strategy equilibrium in the direct revelation mechanism $\langle \Theta, \sigma \rangle$ where $\sigma(a, \theta) = g(a, s^*(\theta))$ for all $a \in A, \theta \in \Theta$.

Proof. The key is to “realize that the result is trivial”, which is to understand how the direct mechanism is constructed and how the payoffs are to be evaluated in the direct mechanism. The strategy profile s^* is a dominant strategy equilibrium if and only if for all $i \in N^1$

$$\sum_{\theta_{-i}} \widetilde{u}_i(s_i^*(\theta_i), s_{-i}(\theta_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) \geq \sum_{\theta_{-i}} \widetilde{u}_i(s_i(\theta_i), s_{-i}(\theta_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) \quad (1)$$

$$\text{for all } s_i(\theta_i) \in M_i, s_{-i} \in S_{-i}, \theta_i \in \Theta_i \quad (2)$$

with the equality holding strict for some $s_{-i} \in S_{-i}$. Now note that we can do the following “computation” for the direct mechanism constructed for every θ_i, θ_{-i} and every conceivable strategy $\widehat{\theta}_{-i} : \Theta_{-i} \rightarrow \Theta_{-i}$

$$\begin{aligned} \sum_{\theta_{-i}} u_i^{DR} \left(\underbrace{\theta_i, \widehat{\theta}_{-i}(\theta_{-i})}_{\text{messages}}, \underbrace{\theta_i, \theta_{-i}}_{\text{true types}} \right) \Pr(\theta_{-i}|\theta_i) &\equiv \sum_{\theta_{-i}} \sum_a \left[u_i(a, \theta) \sigma(a, \theta_i, \widehat{\theta}_{-i}(\theta_{-i})) \right] \Pr(\theta_{-i}|\theta_i) = (3) \\ &= \sum_{\theta_{-i}} \sum_a \left[u_i(a, \theta) g(a, s_i^*(\theta_i), s_{-i}^*(\widehat{\theta}_{-i})) \right] \Pr(\theta_{-i}|\theta_i) \\ &= \sum_{\theta_{-i}} \widetilde{u}_i(s_i^*(\theta_i), s_{-i}^*(\widehat{\theta}_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) , \end{aligned}$$

and similarly for any $\widehat{\theta}_i \in \Theta_i$

$$\begin{aligned} \sum_{\theta_{-i}} u_i^{DR} \left(\underbrace{\widehat{\theta}_i, \widehat{\theta}_{-i}(\theta_{-i})}_{\text{messages}}, \underbrace{\theta_i, \theta_{-i}}_{\text{true types}} \right) \Pr(\theta_{-i}|\theta_i) &\equiv \sum_{\theta_{-i}} \sum_a \left[u_i(a, \theta) \sigma(a, \widehat{\theta}_i, \widehat{\theta}_{-i}(\theta_{-i})) \right] \Pr(\theta_{-i}|\theta_i) (5) \\ &= \sum_{\theta_{-i}} \widetilde{u}_i(s_i^*(\widehat{\theta}_i), s_{-i}^*(\widehat{\theta}_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) . \end{aligned}$$

¹There is a subtle issue about whether dominance should be defined *ex ante* or *ex post* in games of incomplete information. Here we define it *ex post*. This is the stronger version of dominance, but the result holds true no matter which criterion we use. See Fudenberg-Tirole for discussion.

In order for truth telling to be a dominant strategy it must be that

$$\sum_{\theta_{-i}} u_i^{DR} \left(\underbrace{\theta_i, \widehat{\theta}_{-i}(\theta_{-i})}_{\text{messages}}, \underbrace{\theta_i, \theta_{-i}}_{\text{true types}} \right) \Pr(\theta_{-i}|\theta_i) \geq \sum_{\theta_{-i}} u_i^{DR} \left(\underbrace{\widehat{\theta}_i, \widehat{\theta}_{-i}(\theta_{-i})}_{\text{messages}}, \underbrace{\theta_i, \theta_{-i}}_{\text{true types}} \right) \Pr(\theta_{-i}|\theta_i)$$

for all $i \in N$, $\theta_i \in \Theta_i$ and all $\widehat{\theta}_{-i} : \Theta_{-i} \rightarrow \Theta_{-i}$. But using the equalities derived this is equivalent to the condition

$$\sum_{\theta_{-i}} \widetilde{u}_i(s_i^*(\theta_i), s_{-i}(\theta_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) \geq \sum_{\theta_{-i}} \widetilde{u}_i(s_i(\theta_i), s_{-i}(\theta_{-i}), \theta_{-i}, \theta_i) p(\theta_{-i}|\theta_i) \quad (6)$$

$$\text{for all } s_i(\theta_i) \in s_i^*(\Theta_i), \text{ all } s_{-i} : \Theta_{-i} \rightarrow s_{-i}^*(\Theta_{-i}), \theta_i \in \Theta_i \quad (7)$$

where $s_i^*(\Theta_i) = \{m_i \in M_i | \exists \theta_i \in \Theta_i \text{ such that } m_i = s_i^*(\theta_i)\}$ and $s_{-i}^*(\Theta_{-i})$ is similarly defined. Since $s_i^*(\Theta_i) \subset M_i$ and $\{s_{-i} : \Theta_{-i} \rightarrow s_{-i}^*(\Theta_{-i})\} \subset S_{-i}$ this follows directly from the assumption that s^* is a dominant strategy equilibrium. ■

The intuitive explanation is as follows: in the direct mechanism, any configuration of reports gives the same outcomes as some messages in the original mechanism. This means that for any strategy in the direct revelation game there exists some strategy in the original game with the same consequences (converse need not be true). Truth telling for i corresponds with i 's dominant strategy in the original game, so it must be a dominant strategy to tell the truth in the direct mechanism.

Even easier to explain in words is the revelation principle for (Bayesian) Nash equilibria. Starting from a Nash equilibrium in the game induced by an arbitrary mechanism, construct a direct mechanism so that truth-telling leads to the same allocation as the equilibrium strategy for every $\theta \in \Theta$. Then ask: are there any incentives to lie? The answer is, an almost obvious, no. The reason is that if agent i lies given type θ_i , then the lie generates the same outcome as something that i could have gotten by playing some strategy in the original game, whereas truth-telling corresponds with playing the equilibrium strategy in the original game. For there to be an incentive to lie it is therefore necessary for there to be a profitable deviation in the original game, which contradicts the construction of the direct mechanism. That is:

Theorem 2 (*Nash Equilibrium Revelation Principle*). *Fix p, Θ_i and $u^i : A \times \Theta \rightarrow R$ for all $i \in N$ and suppose that s^* is a (Bayesian) Nash equilibrium in the game induced by $\langle M, g \rangle$. Then truthful reporting of type (i.e. $\widetilde{\theta}_i(\theta_i) = \theta_i$ for all $i \in N$ and $\theta_i \in \Theta_i$) is a Nash equilibrium in the direct revelation mechanism $\langle \Theta, \sigma \rangle$ where $\sigma(a, \theta) = g(a, s^*(\theta))$ for all $a \in A, \theta \in \Theta$.*

Proof. Take an arbitrary $m_i \in s_i^*(\Theta_i)$. For this message, where there exists some type $\hat{\theta}_i$ such that $m_i = m_i^*(\hat{\theta}_i)$, we can write agent i 's interim payoffs as

$$\begin{aligned}
v_i(m_i, s_{-i}^*, \theta_i) &= \sum_{\theta_{-i}} \sum_a u_i(a, \theta) g(a, m_i, s_{-i}^*(\theta_{-i})) \Pr(\theta_{-i}|\theta_i) = \\
&= \sum_{\theta_{-i}} \sum_a u_i(a, \theta) g(a, s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})) \Pr(\theta_{-i}|\theta_i) = \\
&= \left/ \text{using } \sigma(\cdot, (\hat{\theta}_i, \theta_{-i})) = g(\cdot, s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})) \right/ = \\
&= \sum_{\theta_{-i}} \sum_a u_i(a, \theta) \sigma(a, \hat{\theta}_i, \theta_{-i}) \Pr(\theta_{-i}|\theta_i) = \sum_{\theta_{-i}} u_i^{DR}(\hat{\theta}_i, \theta_{-i}, \theta) \Pr(\theta_{-i}|\theta_i)
\end{aligned}$$

If s^* is a Nash equilibrium $v_i(s_i^*(\theta), s_{-i}^*, \theta_i) \geq v_i(m_i, s_{-i}^*, \theta_i)$ for all i, θ_i and $m_i \in M_i$. Hence it follows trivially that $v_i(s_i^*(\theta), s_{-i}^*, \theta_i) \geq v_i(m_i, s_{-i}^*, \theta_i)$ for all i, θ_i and $m_i \in s_i^*(\Theta_i)$, so

$$\begin{aligned}
\sum_{\theta_{-i}} u_i^{DR}(\theta_i, \theta_{-i}, \theta) \Pr(\theta_{-i}|\theta_i) &= v_i(s_i^*(\theta_i), s_{-i}^*, \theta_i) \geq v_i(s_i^*(\hat{\theta}_i), s_{-i}^*, \theta_i) = \\
&= \sum_{\theta_{-i}} u_i^{DR}(\hat{\theta}_i, \theta_{-i}, \theta) \Pr(\theta_{-i}|\theta_i)
\end{aligned}$$

for all $\hat{\theta}_i \in \Theta_i$. Hence truth telling is Nash in direct mechanism. ■

It should be kept in mind what the revelation principle does not say: it is perfectly possible that there are multiple equilibria in a direct revelation mechanism and more elaborate mechanisms can sometimes eliminate the multiplicity. Hence, more complicated mechanisms can play a role if we are interested in *strong implementation*, that is to implement something *uniquely*. It should be noted in this context that the revelation principle often is used to get *negative* results. If something *can not* be an equilibrium in a direct mechanism, then this something *can not* be an equilibrium in any kind of mechanism. Invoked this way, the revelation principle has been used to prove several interesting impossibility and asymptotic impossibility results, some of which we will see later on.

16 The Gibbard-Satterthwaite Theorem

One classic and highly relevant setup is if a “type” is a preference relation over some “social alternatives” given by a set A . We may think of this set A as being some various policies that can be adopted by a society. In principle, A could be Walrasian allocations, but it is crucial that all possible preference orderings are possible for the result we will discuss (and monotonicity is sort of natural in many cases), so it is better to have in mind some discrete set of policies that can

conceivably be ranked in any possible way. A natural example is to let A be the set of potential alternatives for some public office, another is to let A be a list of possible policies where single-peaked preferences seems to be a stretch.

For each i , we let $\Theta^i = \mathcal{R}(A)$, the set of admissible weak orderings over A . If we ignore informational asymmetries we could then imagine some sort supernatural being constructing a rule that would pick an outcome for every possible preference ordering. Such a rule, $f : \Theta \rightarrow A$ is referred to as a *social choice function*. We note that such a rule would say what should happen in any “conceivable society”, whereas the usual normative approach to collective choice problems with complete information about preferences is to take preferences for given.

While this sounds a bit silly, the exercise we will consider simply asks: is it ever possible to find a “somewhat nice” social choice function where agents have no incentive to manipulate their preferences. The answer depends on how big A is: if A has 2 elements, then majority rule “works”. With 3 or more elements, the only way to achieve truth-telling is to ignore all but a single agent. Hence, we know that trying to find some nice general procedure to elicit preferences from agents is a dead end, so we need to consider more specific problems. Notice that if we would have found that any “desirable” $f : \Theta \rightarrow A$ could be implemented by some clever procedure, then game theory would not be needed when doing normative analysis.

We now suppose that we have a social planner with some preferences for society (that depends on individual preferences), so we are essentially forgetting about the problems implied by Arrows theorem, or considering a social welfare function either based on interpersonal comparisons of utility, or one where for instance the independence axiom fails.

The issue now is the following: the planner does not know the true preferences of the agents. We ask if it is possible for the planner to construct a mechanism such that no individual has incentives to misrepresent preferences.

For simplicity we only consider strict preferences. We will also (not necessary but avoids some possible sources of confusion) introduce some additional notation apart from the general notation above and let

$$P_i = P = \{p | p \text{ is a strict ordering on } A\}$$

for all $i \in N$. We let $\times_i P_i = P^n$. Recall that a function $F : P^n \rightarrow P$ is a *social welfare function* and that a function $f : P^n \rightarrow A$ is a *social choice function*. For any $p = (p_1, \dots, p_n) \in P^n$ we use

the notation $p|p'_i$ to denote $(p_1, \dots, p_{i-1}, p'_i, p_{i+1}, \dots, p_n)$. We now introduce some terminology that is common in the literature.

Definition 5 *A social choice function f is manipulable (at $p \in P^n$) if there exists some $i \in N$, $p \in P^n$ and $p'_i \in P$ such that $f(p|p'_i) p_i f(p)$*

Intuitively, the interpretation is simply that a SCF is manipulable if someone has an incentive to misreport preferences at some $p \in P^n$. However, there are some subtle issues that has to be considered when we make the connection to equilibria in games induced by direct revelation mechanisms. We also say that

Definition 6 *A social choice function f is strategy-proof (or incentive compatible) if it is not manipulable at any preferences.*

Intuitive as these definitions are, is not directly obvious (at least not to me) how to translate this into truth telling being a dominant strategy/Nash equilibrium in the direct revelation game $\langle P^n, f \rangle$. The issue is that on the level of generality we considered in the last section we had that agents utilities were dependent on both the announcements of all agents and the true type vector (think of the oil-well example if you need an example where this makes sense). However, in the context of implementation of social choice functions we have that *each type cares only about the outcomes in A* , so the only way other peoples preferences affect an agent is through the influence on what outcome is implemented. That is

$$u^i(x, p_i, p_{-i}) = u^i(x, p_i, p'_{-i}) \text{ for all } p_{-i}, p'_{-i} \quad (8)$$

Now, if truth telling is a dominant strategy in the direct mechanism $\langle P^n, f \rangle$, then by definition (verify!)

$$\sum_{p_{-i}} u^i(f(p_i, \tilde{p}_{-i}(p_{-i})), p) \Pr(p_{-i}|p_i) \geq \sum_{p_{-i}} u^i(f(\hat{p}_i, \tilde{p}_{-i}(p_{-i})), p) \Pr(p_{-i}|p_i)$$

for each $i \in N$, $p_i \in P$ and $\tilde{p}_{-i} : P^{n-1} \rightarrow P^{n-1}$. Now, since (8) holds we have that for some $v^i : A \times P \rightarrow R$

$$\sum_{p_{-i}} u^i(f(p_i, \tilde{p}_{-i}(p_{-i})), p) \Pr(p_{-i}|p_i) = \sum_{p_{-i}} v^i(f(p_i, \tilde{p}_{-i}(p_{-i})), p_i) \Pr(p_{-i}|p_i)$$

We now claim

Proposition 1 *The following statements are equivalent:*

1. f is strategy-proof
2. Truthful reporting is a dominant strategy equilibrium of the direct revelation game induced by $\langle P^n, f \rangle$
3. Truthful reporting is a Nash equilibrium of the direct revelation game induced by $\langle P^n, f \rangle$ where it is assumed that all agents know the types of all other agents.

Proof. (strategy-proof \Rightarrow truth telling dominant) f strategy-proof $\Leftrightarrow f$ is not manipulable at any $p \in P^n \Leftrightarrow f(p) p_i f(p|p'_i) \forall p \in P^n, \forall i \in N, \forall p'_i \in P$. Now, for each $i \in N$ and $p_i \in P$, let $v_i(x, p_i)$ denote some utility function representation of p_i . Then $f(p) p_i f(p|p'_i) \forall p \in P^n, \forall i \in N, \forall p'_i \in P$ is equivalent to

$$v_i(f(p), p_i) \geq v_i(f(p|p'_i), p_i) \forall p \in P^n, \forall i \in N, \forall p'_i \in P \quad (9)$$

Hence, if we let $\tilde{p}_{-i} : P^{n-1} \rightarrow P^{n-1}$ be an arbitrary strategy-profile for the rest of the agents we have

$$v_i(f(p_i, \tilde{p}_{-1}(p_{-i})), p_i) \geq v_i(f(p'_i, \tilde{p}_{-1}(p_{-i})), p_i)$$

for all $p_{-i} \in P^{n-1}$. Multiplying by $\Pr(p_{-i}|p_i)$ and summing gives

$$\sum_{p_{-i}} v_i(f(p_i, \tilde{p}_{-1}(p_{-i})), p_i) \Pr(p_{-i}|p_i) \geq \sum_{p_{-i}} v_i(f(p'_i, \tilde{p}_{-1}(p_{-i})), p_i) \Pr(p_{-i}|p_i) \quad (10)$$

which means that truth telling is a dominant strategy (note that which particular representation of the ordinal preferences we choose is not an issue).

(truth telling dominant \Rightarrow Nash) By considering strategies such that $\tilde{p}_{-i}(p_{-i}) = p'_{-i}$ we notice that (10) and (9) are equivalent. The result is then obvious, since dominance implies Nash in general.

(truth telling Nash \Rightarrow dominant). Suppose truth telling is a Nash equilibrium in the direct revelation game where all agents know the preferences of all other agents. Then

$$v_i(f(p), p_i) \geq v_i(f(p|p'_i), p_i) \forall p \in P^n, \forall i \in N, \forall p'_i \in P$$

which is exactly the same condition as above.

(truth-telling dominant \Rightarrow strategy proof) Condition above equivalent with strategy proofness. ■

Note what is not true. A Bayesian Nash equilibrium in the direct revelation game where preferences are private information need not be a dominant strategy equilibrium. However, the important point that you have to understand fully is that *if a particular f is implementable by a mechanism $\langle M, g \rangle$ then there exists a direct mechanism that implements f as a truth telling equilibrium and this is true if and only if f is “strategy-proof”.*

We need one more definition before stating the result. The definition is similar to the analogous definition in social choice theory and says that an agent is a *dictator* if she gets her most preferred outcome *among the set of alternatives that actually occurs after some announcement*. The notion of a dictatorial social choice function is then:

Definition 7 *A social choice function is dictatorial if there exists $i \in N$ such that $f(p) p_i x$ for all $p \in P^n$ and all $x \in f(P^n)$, where $f(P^n)$ denotes the direct image of P^n under f .*

The fundamental negative result in implementation theory is:

Theorem 3 *(Gibbard, Satterthwaite) If a social choice function f is strategy-proof and if $f(P^n) \geq 3$, then f is dictatorial.*

Proof. Suppose f is strategy-proof and that $f(P^n) \geq 3$. We will construct a fictitious social welfare function F from f . Then we apply Arrows theorem to conclude that F must be dictatorial (in Arrows sense) and then finally show that f is dictatorial in the sense above:

Lemma 1 *There exists no $p \in P^n, i \in N$ and $p'_i \in P$ such that $f(p) = x \neq y = f(p|p'_i)$, and $x p_i y$ if and only if $x p'_i y$.*

Proof. Suppose $f(p) = x \neq y = f(p|p'_i)$, $x p_i y$ and $x p'_i y$. Then f is manipulable by i at $p|p'_i$, so it is not strategy-proof, which is a contradiction. If, on the other hand, $y p_i x$ and $y p'_i x$, then f is manipulable at p . ■

The meaning of this lemma is that if an agent is “pivotal” (change in i ’s preferences changes social decision), then it must be that the social decision changes in accordance with i ’s preferences.

Lemma 2 *Suppose $B \subset f(P^n)$ and p is such that if $x \in B$ and $y \in A \setminus B$, then $x p_i y$ for all $i \in N$. Then $f(p) \in B$.*

Proof. Suppose the hypotheses of lemma hold, but $f(p) = y \notin B$. For any $x \in B$ the fact that $B \subset f(P^n)$ means that there exists some $p' \in P^n$ such that $f(p') = x$. For $i = 0, 1, \dots, n$ define

$$z_i = f(p|p'_1, \dots, p'_i)$$

where $p|p'_1, \dots, p'_i$ denotes $(p'_1, \dots, p'_i, p_{i+1}, \dots, p_n)$. Note that

$$\begin{aligned} z_0 &= f(p) = y \\ z_n &= f(p') = x \end{aligned}$$

Let $j = \min\{i \in N | z_i \in B\}$, which exists since $z_n \in B$. Since $z_{j-1} \notin B$ and $z_j \in B$ we have that $z_j p_i z_{j-1}$ or $f(p|p'_1, \dots, p'_j) p_i f(p|p'_1, \dots, p'_{j-1})$, which means that f is manipulable by j given preference profile $p'' = (p'_1, \dots, p'_{j-1}, p_j, \dots, p_n)$. ■

The next lemma is technical and can be skipped by anybody who is not very interested in the most general of the general cases. It says essentially that the case with A being infinite will be covered also by the proof for the finite case.

Lemma 3 $|f(P^n)| < \infty$

Proof. If not, there exists $D \subset f(P^n)$ and a function $h : D \rightarrow N$ (that is, any infinite set has a countably infinite subset. Consider the following preferences over $f(P^n)$

for $x, y \in D$, let $x p_i y$ if and only if $h(x) > h(y)$

for $x \in D, z \notin D$, let $x p_i y$

for $z, w \notin D$, let preferences be arbitrary

Let

$$B_n \equiv \{x \in D | h(x) \geq n\}$$

Using the previous lemma $f(p) \in B_n$ for all n (note that we have specified identical rankings for all agents), which implies that $f(p) \in \bigcap_n B_n = \phi$, which is a contradiction. ■

Now we start to construct a (fictitious) social welfare function over $A' = f(P^n)$. Let p be an arbitrary preference profile and let $x_1 = f(p)$. Construct a new profile of rankings p^1 by moving x_1 to the bottom of the ranking for all agents. By Lemma ?? we have that $f(p^1) \in A' \setminus \{x_1\}$. Let $x_2 = f(p^1)$ and continue inductively by letting $x_n = f(p^{n-1})$, where we for each step note

that $x_n \in A' \setminus \{x_1, \dots, x_{n-1}\}$. Since A' is finite this process must end in a finite number of steps (after exactly $|A'|$ steps indeed) and the result is a (strict) ranking $F(p)$ where $x_i F(p) x_j \Leftrightarrow i < j$ (however, this is just how we named the elements in A').

We now want to show that we can restrict the domain of F to be all strict preference orderings over A' and that the resulting social welfare function then satisfies Unanimity and IIA, so it must be dictatorial.

Lemma 4 *If $a, b \in A'$ and $a F(p) b$, then $f(p^*) = a$ where p^* is the preference profile obtained by moving a, b to the top for all i without changing the relative orderings of either a and b or the relative orderings of the other elements.*

Proof. Suppose $a F(p) b$, but $f(p^*) \neq a$. By Lemma 2 this implies $f(p^*) = b$. Now let p' be the profile in the sequence when $F(p)$ is defined such that $f(p') = a$ (must appear at one point). Note that $a F(p) b$ implies that a is moved to the end before b when $F(p)$ is defined so $ap_i b$ if and only if $ap'_i b$ (and by construction iff $ap_i^* b$). Now let

$$z_i = f(p^* | p'_1, \dots, p'_i),$$

so that $z_0 = b$ and $z_n = a$. We also let $k = \min \{i \in N | z_i \neq b\}$. There are now two possibilities:

i) $z_k = c \neq a$. Then let $j = \min \{i > k | z_i \in \{a, b\}\}$. Then

$$\begin{aligned} z_j &= f(p'_1, \dots, p'_j, p_{j+1}^*, \dots, p_n^*) \in \{a, b\} \\ z_{j-1} &= f(p'_1, \dots, p'_{j-1}, p_j^*, \dots, p_n^*) \notin \{a, b\} \end{aligned}$$

and since a or b by construction is at the top of all p_j^* orderings $z_j p_j^* z_{j-1}$, that is

$$f(p'_1, \dots, p'_{j-1}, p_j^*, \dots, p_n^* | p'_j) p_j^* f(p'_1, \dots, p'_{j-1}, p_j^*, \dots, p_n^*),$$

which means that

ii) $z_k = a$. In this case $f(p^* | p'_1, \dots, p'_k) = a \neq b = f(p^* | p'_1, \dots, p'_{k-1})$. This is quickly disposed of by the first lemma since by construction p_k^* and p'_k agrees on choice between a and b . ■

Lemma 5 *F satisfies IIA on A'*

Proof. If not there are two profiles $p, q \in P^n$ such that $ap_i b \Leftrightarrow aq_i b$ for all $i \in N$, but $aF(p)b$ and $bF(q)a$. Form p^* and q^* as in previous lemma by moving a, b to the top without affecting the relative rankings. By Lemma 4 it follows that $f(p^*) = a$ and $f(q^*) = b$. Construct sequence

$$w_i = f(p^* | q_1^*, \dots, q_i^*)$$

so that $w_0 = a$ and $w_n = b$. Continue the proof by an identical argument as in Lemma 4 (i.e. there must be a point where outcome switches from a to b and at this point f is manipulable). ■

Lemma 6 *F satisfies unanimity on A'*

Proof. Follows trivially from Lemma 2, which can be viewed as “set-wise” unanimity. ■

By Arrows theorem, if F satisfies IIA and Unanimity, F is dictatorial on A' , which means that there exists a dictator for f (if you don't understand this last step-go back to construction of F and/or definition of a dictator). ■

Given the fundamental negative implementation result provided by the Gibbard-Satterthwaite we know that we need to go to more specific models and/or weaken the notion of implementation in order to say anything about how agents can be induced to correctly reveal their private information. There is a literature on Nash implementation of social choice functions and more recently some authors have studied various notions of approximate implementation.