CHAPTER 9

Introduction to Mechanism Design (for Computer Scientists)

Noam Nisan

Abstract

We give an introduction to the micro-economic field of Mechanism Design slightly biased toward a computer-scientist's point of view.

9.1 Introduction

Mechanism Design is a subfield of economic theory that is rather unique within economics in having an engineering perspective. It is interested in designing economic mechanisms, just like computer scientists are interested in designing algorithms, protocols, or systems. It is best to view the goals of the designed mechanisms in the very abstract terms of *social choice*. A social choice is simply an aggregation of the preferences of the different participants toward a single joint decision. *Mechanism Design* attempts implementing desired social choices in a strategic setting – assuming that the different members of society each act *rationally* in a game theoretic sense. Such strategic design is necessary since usually the preferences of the participants are private.

This high-level abstraction of aggregation of preferences may be seen as a common generalization of a multitude of scenarios in economics as well as in other social settings such as political science. Here are some basic classic examples:

- **Elections:** In political elections each voter has his own preferences between the different candidates, and the outcome of the elections is a single social choice.
- **Markets:** Classical economic theory usually assumes the existence and functioning of a "perfect market." In reality, of course, we have only interactions between people, governed by some protocols. Each participant in such an interaction has his own preferences, but the outcome is a single social choice: the reallocation of goods and money.
- **Auctions:** Generally speaking, the more buyers and sellers there are in a market, the more the situation becomes close to the perfect market scenario. An extreme opposite

case is where there is only a single seller – an auction. The auction rules define the social choice: the identity of the winner.

- **Government policy:** Governments routinely have to make decisions that affect a multitude of people in different ways: Should a certain bridge be built? How much pollution should we allow? How should we regulate some sector? Clearly each citizen has a different set of preferences but a single social choice is made by the government.

As the influence of the Internet grew, it became clear that many scenarios happening there can also be viewed as instances of social choice in strategic settings. The main new ingredient found in the Internet is that it is owned and operated by different parties with different goals and preferences. These preferences, and the behavior they induce, must then be taken into account by every protocol in such an environment. The protocol should thus be viewed as taking the preferences of the different participants and aggregating them into a social choice: the outcome of the run of the protocol.

Conceptually, one can look at two different types of motivations: those that use economics to solve computer science issues and those that use computer science to solve economic issues:

- **Economics for CS:** Consider your favorite algorithmic challenge in a computer network environment: routing of messages, scheduling of tasks, allocation of memory, etc. When running in an environment with multiple owners of resources or requests, this algorithm must take into account the different preferences of the different owners. The algorithm should function well assuming strategic selfish behavior of each participant. Thus we desire a Mechanism Design approach for a multitude of algorithmic challenges – leading to a field that has been termed *Algorithmic Mechanism Design*.
- **CS for economics:** Consider your favorite economic interaction: some type of market, an auction, a supply chain, etc. As the Internet becomes ubiquitous, this interaction will often be implemented over some computerized platform. Such an implementation enables unprecedented sophistication and complexity, handled by hyperrationally designed software. Designing these is often termed *Electronic Market Design*.

Thus, both Algorithmic Mechanism Design and Electronic Market Design can be based upon the field of Mechanism Design applied in complex algorithmic settings.

This chapter provides an introduction to classical Mechanism Design, intended for computer scientists. While the presentation is not very different from the standard economic approach, it is somewhat biased toward a worst-case (non-Bayesian) point of view common in computer science.

Section 9.2 starts with the general formulation of the social choice problem, points out the basic difficulties formulated by Arrow's famous impossibility results, and deduces the impossibility of a general strategic treatment, i.e. of Mechanism Design in the general setting. Section 9.3 then considers the important special case where "money" exists, and describes a very general positive result, the incentive-compatible Vickrey–Clarke–Grove mechanism. Section 9.4 puts everything in a wider formal context of implementation in dominant strategies. Section 9.5 provides several characterizations of dominant strategy mechanisms. All the sections up to this point have considered dominant strategies, but the prevailing economic point of view is a Bayesian one that assumes a priori known distributions over private information. Section 9.6 introduces

this setting and the notion of Bayesian-Nash equilibrium that fits it. All the treatment in this chapter is in the very basic "private value" model, and Section 9.7 shortly points out several extensions to the model. Finally, Section 9.8 provides bibliographic notes and references.

9.2 Social Choice

This section starts with the general social choice problem and continues with the strategic approach to it. The main message conveyed is that there are unavoidable underlying difficulties. We phrase things in the commonly used terms of political elections, but the reader should keep in mind that the issues are abstract and apply to general social choice.

9.2.1 Condorcet's Paradox

Consider an election with two candidates, where each voter has a preference for one of them. If society needs to jointly choose one of the candidates, intuitively it is clear that taking a *majority vote* would be a good idea. But what happens if there are three candidates? In 1785, The Marquis de Condorcet pointed out that the natural application of majority is problematic: consider three candidates – a, b, and c – and three voters with the following preferences:

(i) $a \succ_1 b \succ_1 c$
(ii) $b \succ_2 c \succ_2 a$
(iii) $c \succ_3 a \succ_3 b$

(The notation $a \succ_i b$ means that voter i prefers candidate a to candidate b.) Now, notice that a majority of voters (1 and 3) prefer candidate a to candidate b. Similarly, a majority (1 and 2) prefers b to c, and, finally, a majority (2 and 3) prefers c to a. The joint majority choice is thus $a \succ b \succ c \succ a$ which is not consistent. In particular for any candidate that is jointly chosen, there will be a majority of voters who would want to change the chosen outcome.

This immediately tells us that in general a social choice cannot be taken simply by the natural system of taking a majority vote. Whenever there are more than two alternatives, we must design some more complex "voting method" to undertake a social choice.

9.2.2 Voting Methods

A large number of different *voting methods* – ways of determining the outcome of such multicandidate elections – have been suggested. Two of the simpler ones are *plurality* (the candidate that was placed first by the largest number of voters wins) and Borda count (each candidate among the n candidates gets $n - i$ points for every voter who ranked him in place i, and the candidate with most points wins). Each of the suggested voting methods has some "nice" properties but also some problematic ones.

One of the main difficulties encountered by voting methods is that they may encourage *strategic voting*. Suppose that a certain voter's preferences are $a \succ_i b \succ_i c$, but he knows that candidate a will not win (as other voters hate him). Such a voter may be

motivated to strategically vote for b instead of a, so that b is chosen which he prefers to c. Such strategic voting is problematic as it is not transparent, depends closely on the votes of the other voters, and the interaction of many strategic voters is complex. The main result of this section is the Gibbard–Satterthwaite theorem that states that this strategic vulnerability is unavoidable. We will prove the theorem as a corollary of Arrow's impossibility theorem that highlights the general impossibility of designing voting methods with certain natural good desired properties.

Formally, we will consider a set of alternatives A (the candidates) and a set of n voters I. Let us denote by L the set of linear orders on A (L is isomorphic to the set of permutations on A). Thus for every $\prec \in L$, $\prec$ is a total order on A (antisymmetric and transitive). The preferences of each voter i are formally given by $\succ_i \in L$, where $a \succ_i b$ means that i prefers alternative a to alternative b.

Definition 9.1

- A function $F : L^n \to L$ is called a *social welfare function*.
- A function $f : L^n \to A$ is called a *social choice function*.

Thus a social welfare function aggregates the preferences of all voters into a common preference, i.e., into a total social order on the candidates, while a social choice function aggregates the preferences of all voters into a social choice of a single candidate. Arrow's theorem states that social welfare functions with "nice" properties must be trivial in a certain sense.

9.2.3 Arrow's Theorem

Here are some natural properties desired from a social welfare function.

Definition 9.2

- A social welfare function F satisfies *unanimity* if for every $\prec \in L$, $F(\prec, \ldots, \prec) = \prec$. That is, if all voters have identical preferences then the social preference is the same.
- Voter i is a *dictator* in social welfare function F if for all $\prec_1 \ldots \prec_n \in L$, $F(\prec_1, \ldots, \prec_n) = \prec_i$. The social preference in a dictatorship is simply that of the dictator, ignoring all other voters. F is not a *dictatorship* if no i is a dictator in it.
- A social welfare function satisfies *independence of irrelevant alternatives* if the social preference between any two alternatives a and b depends only on the voters' preferences between a and b. Formally, for every $a, b \in A$ and every $\prec_1, \ldots, \prec_n, \prec'_1, \ldots, \prec'_n \in L$, if we denote $\prec = F(\prec_1, \ldots, \prec_n)$ and $\prec' = F(\prec'_1, \ldots, \prec'_n)$ then $a \prec_i b \Leftrightarrow a \prec'_i b$ for all i implies that $a \prec b \Leftrightarrow a \prec' b$.

The first two conditions are quite simple to understand, and we would certainly want any good voting method to satisfy the unanimity condition and not to be a dictatorship. The third condition is trickier. Intuitively, indeed, independence of irrelevant alternatives seems quite natural: why should my preferences about c have anything to do with

the social ranking of a and b? More careful inspection will reveal that this condition in some sense captures some consistency property of the voting system. As we will see, lack of such consistency enables strategic manipulation.

Theorem 9.3 (Arrow) *Every social welfare function over a set of more than 2 candidates ($|A| \geq 3$) that satisfies unanimity and independence of irrelevant alternatives is a dictatorship.*

Over the years a large number of proofs have been found for Arrow's theorem. Here is a short one.

PROOF For the rest of the proof, fix F that satisfies unanimity and independence of irrelevant alternatives. We start with a claim showing that the same social ranking rule is taken within any pair of alternatives.

Claim (pairwise neutrality) Let $\succ_1, \ldots, \succ_n$ and $\succ'_1, \ldots, \succ'_n$ be two player profiles such that for every player i, $a \succ_i b \Leftrightarrow c \succ'_i d$. Then $a \succ b \Leftrightarrow c \succ' d$, where $\succ = F(\succ_1, \ldots, \succ_n)$ and $\succ' = F(\succ'_1, \ldots, \succ'_n)$.

By renaming, we can assume without loss of generality that $a \succ b$ and that $c \neq b$. Now we merge each $\succ_i$ and $\succ'_i$ into a single preference $\succ_i$ by putting c just above a (unless $c = a$) and d just below b (unless $d = b$) and preserving the internal order within each of the pairs (a, b) and (c, d). Now using unanimity, we have that $c \succ a$ and $b \succ d$, and by transitivity $c \succ d$. This concludes the proof of the claim.

We now continue with the proof of the theorem. Take any $a \neq b \in A$, and for every $0 \leq i \leq n$ define a preference profile π^i in which exactly the first i players rank a above b, i.e., in π^i, $a \succ_j b \Leftrightarrow j \leq i$ (the exact ranking of the other alternatives does not matter). By unanimity, in $F(\pi^0)$, we have $b \succ a$, while in $F(\pi^n)$ we have $a \succ b$. By looking at $\pi^0, \pi^1, \ldots, \pi^n$, at some point the ranking between a and b flips, so for some i^* we have that in $F(\pi^{i^*-1})$, $b \succ a$, while in $F(\pi^{i^*})$, $a \succ b$. We conclude the proof by showing that i^* is a dictator.

Claim Take any $c \neq d \in A$. If $c \succ_{i^*} d$ then $c \succ d$ where $\succ = F(\succ_1, \ldots, \succ_n)$.

Take some alternative e which is different from c and d. For $i < i^*$ move e to the top in $\succ_i$, for $i > i^*$ move e to the bottom in $\succ_i$, and for i^* move e so that $c \succ_{i^*} e \succ_{i^*} d$ – using independence of irrelevant alternatives we have not changed the social ranking between c and d. Now notice that players' preferences for the ordered pair (c, e) are identical to their preferences for (a, b) in π^{i^*}, but the preferences for (e, d) are identical to the preferences for (a, b) in π^{i^*-1} and thus using the pairwise neutrality claim, socially $c \succ e$ and $e \succ d$, and thus by transitivity $c \succ d$. □

9.2.4 The Gibbard–Satterthwaite Theorem

It turns out that Arrow's theorem has devastating strategic implications. We will study this issue in the context of social choice functions (rather than social welfare functions as we have considered until now). Let us start by defining strategic manipulations.

Definition 9.4 A social choice function f can be *strategically manipulated* by voter i if for some $\prec_1, \ldots, \prec_n \in L$ and some $\prec_i' \in L$ we have that $a \prec_i a'$ where $a = f(\prec_1, \ldots, \prec_i, \ldots, \prec_n)$ and $a' = f(\prec_1, \ldots, \prec_i', \ldots, \prec_n)$. That is, voter i that prefers a' to a can ensure that a' gets socially chosen rather than a by strategically misrepresenting his preferences to be $\prec_i'$ rather than $\prec_i$. f is called *incentive compatible* if it cannot be manipulated.

The following is a more combinatorial point of view of the same notion.

Definition 9.5 A social choice function f is monotone if $f(\prec_1, \ldots, \prec_i, \ldots, \prec_n) = a \neq a' = f(\prec_1, \ldots, \prec_i', \ldots, \prec_n)$ implies that $a' \prec_i a$ and $a \prec_i' a'$. That is, if the social choice changed from a to a' when a single voter i changed his vote from $\prec_i$ to $\prec_i'$ then it must be because he switched his preference between a and a'.

Proposition 9.6 *A social choice function is incentive compatible if and only if it is monotone.*

PROOF Take $\prec_1, \ldots, \prec_{i-1}, \prec_{i+1}, \ldots, \prec_n$ out of the quantification. Now, logically, "NOT monotone between $\prec_i$ and $\prec_i'$" is equivalent to "A voter with preference $\prec$ can strategically manipulate f by declaring $\prec'$" OR "A voter with preference $\prec'$ can strategically manipulate f by declaring $\prec$". □

The obvious example of an incentive compatible social choice function over two alternatives is taking the majority vote between them. The main point of this section is, however, that when the number of alternatives is larger than 2, only trivial social choice functions are incentive compatible.

Definition 9.7 Voter i is a *dictator* in social choice function f if for all $\prec_1, \ldots, \prec_n \in L$, $\forall b \neq a$, $a \succ_i b \Rightarrow f(\prec_1, \ldots, \prec_n) = a$. f is called a *dictatorship* if some i is a dictator in it.

Theorem 9.8 (Gibbard–Satterthwaite) *Let f be an incentive compatible social choice function onto A, where $|A| \geq 3$, then f is a dictatorship.*

Note the requirement that f is onto, as otherwise the bound on the size of A has no bite. To derive the theorem as a corollary of Arrow's theorem, we will construct a social welfare function F from the social choice function f. The idea is that in order to decide whether $a \prec b$, we will "move" a and b to the top of all voters' preferences, and then see whether f chooses a or b. Formally,

Definition 9.9

- Notation: Let $S \subset A$ and $\prec \in L$. Denote by $\prec^S$ the order obtained by moving all alternatives in S to the top in $\prec$. Formally, for $a, b \in S$, $a \prec^S b \Leftrightarrow a \prec b$; for $a, b \notin S$, also $a \prec^S b \Leftrightarrow a \prec b$; but for $a \notin S$ and $b \in S$, $a \prec^S b$.

- The social welfare function F that extends the social choice function f is defined by $F(\prec_1, \ldots, \prec_n) = \prec$, where $a \prec b$ iff $f(\prec_1^{\{a,b\}}, \ldots, \prec_n^{\{a,b\}}) = b$.

We first have to show that F is indeed a social welfare function, i.e., that it is antisymmetric and transitive.

Lemma 9.10 *If f is an incentive compatible social choice function onto A then the extension F is a social welfare function.*

To conclude the proof of the theorem as a corollary of Arrow's, it then suffices to show:

Lemma 9.11 *If f is an incentive compatible social choice function onto A, which is not a dictatorship then the extension F satisfies unanimity and independence of irrelevant alternatives and is not a dictatorship.*

PROOF OF LEMMAS 9.10 AND 9.11 We start with a general claim which holds under the conditions on f:

Claim: For any $\prec_1, \ldots, \prec_n$ and any S, $f(\prec_1^S, \ldots, \prec_n^S) \in S$.

Take some $a \in S$ and since f is onto, for some $\prec_1', \ldots, \prec_n'$, $f(\prec_1', \ldots, \prec_n') = a$. Now, sequentially, for $i = 1, \ldots, n$, change $\prec_i'$ to $\prec_i^S$. We claim that at no point during this sequence of changes will f output any outcome $b \notin S$. At every stage this is simply due to monotonicity since $b \prec_i^S a'$ for $a' \in S$ being the previous outcome. This concludes the proof of the claim.

We can now prove all properties needed for the two lemmas:

- Antisymmetry is implied by the claim since $f(\prec_1^{\{a,b\}}, \ldots, \prec_n^{\{a,b\}}) \in \{a, b\}$.
- Transitivity: assume for contradiction that $a \prec b \prec c \prec a$ (where $\prec = F(\prec_1, \ldots, \prec_n)$). Take $S = \{a, b, c\}$ and using the claim assume without loss of generality that $f(\prec_1^S, \ldots, \prec_n^S) = a$. Sequentially changing $\prec_i^S$ to $\prec_i^{\{a,b\}}$ for each i, monotonicity of f implies that also $f(\prec_1^{\{a,b\}}, \ldots, \prec_n^{\{a,b\}}) = a$, and thus $a \succ b$.
- Unanimity: If for all i, $b \prec_i a$, then $(\prec_i^{\{a,b\}})^{\{a\}} = \prec_i^{\{a,b\}}$ and thus by the claim $f(\prec_1^{\{a,b\}}, \ldots, \prec_n^{\{a,b\}}) = a$.
- Independence of irrelevant alternatives: If for all $i, b \prec_i a \Leftrightarrow b \prec_i' a$, then $f(\prec_1^{\{a,b\}}, \ldots, \prec_n^{\{a,b\}}) = f(\prec_1'^{\{a,b\}}, \ldots, \prec_n'^{\{a,b\}})$ since when we, sequentially for all i, flip $\prec_i^{\{a,b\}}$ into $\prec_i'^{\{a,b\}}$, the outcome does not change because of monotonicity and the claim.
- Nondictatorship: obvious. □

The Gibbard–Satterthwaite theorem seems to quash any hope of designing incentive compatible social choice functions. The whole field of Mechanism Design attempts escaping from this impossibility result using various modifications in the model. The next section describes how the addition of "money" offers an escape route. Chapter 10 offers other escape routes that do not rely on money.

9.3 Mechanisms with Money

In the previous section, we modeled a voter's preference as an order on the alternatives. $a \succ_i b$ implies that i prefers a to b, but we did not model "by how much" is a preferred to b. "Money" is a yardstick that allows measuring this. Moreover, money can be transferred between players. The existence of money with these properties is an assumption, but a fairly reasonable one in many circumstances, and will allow us to do things that we could not do otherwise.

Formally, in this section we redefine our setting. We will still have a set of alternatives A and a set of n players I (which we will no longer call voters). The preference of a player i is now given by a *valuation function* $v_i : A \to \Re$, where $v_i(a)$ denotes the "value" that i assigns to alternative a being chosen. This value is in terms of some currency; i.e., we assume that if a is chosen and then player i is additionally given some quantity m of money, then i's *utility* is $u_i = v_i(a) + m$, this utility being the abstraction of what the player desires and aims to maximize. Utilities of this form are called *quasilinear preferences*, denoting the separable and linear dependence on money.

9.3.1 Vickrey's Second Price Auction

Before we proceed to the general setting, in this subsection we study a basic example: a simple auction. Consider a single item that is auctioned for sale among n players. Each player i has a scalar value w_i that he is "willing to pay" for this item. More specifically, if he wins the item, but has to pay some price p for it, then his utility is $w_i - p$, while if someone else wins the item then i's utility is 0. Putting this scenario into the terms of our general setting, the set of alternatives here is the set of possible winners, $A = \{i\text{–wins} | i \in I\}$, and the valuation of each bidder i is $v_i(i\text{–wins}) = w_i$ and $v_i(j\text{–wins}) = 0$ for all $j \neq i$. A natural social choice would be to allocate the item to the player who values it highest: choose i–wins, where $i = \text{argmax}_j w_j$. However, the challenge is that we do not know the values w_i but rather each player knows his own value, and we want to make sure that our mechanism decides on the allocation – the social choice – in a way that *cannot be strategically manipulated*. Our degree of freedom is the definition of the payment by the winner.

Let us first consider the two most natural choices of payment and see why they do not work as intended:

- **No payment:** In this version we give the item for free to the player with highest w_i. Clearly, this method is easily manipulated: every player will benefit by exaggerating his w_i, reporting a much larger $w_i' \gg w_i$ that can cause him to win the item, even though his real w_i is not the highest.
- **Pay your bid:** An attempt of correction will be to have the winner pay the declared bid. However, this system is also open to manipulation: a player with value w_i who wins and pays w_i gets a total utility of 0. Thus it is clear that he should attempt declaring a somewhat lower value $w_i' < w_i$ that still wins. In this case he can still win the item getting a value of w_i (his real value) but paying only the smaller w_i' (his declared value), obtaining a net positive utility $u_i = w_i - w_i' > 0$. What value w_i' should i bid then?

Well, if i knows the value of the second highest bid, then he should declare just above it. But what if he does not know?

Here is the solution.

Definition 9.12 Vickrey's second price auction: Let the winner be the player i with the highest declared value of w_i, and let i pay the second highest declared bid $p^* = \max_{j \neq i} w_j$.

Now it turns out that manipulation never can increase any players' utility. Formally,

Proposition 9.13 (Vickrey) *For every $w_1, \ldots, w_n$ and every w_i', Let u_i be i's utility if he bids w_i and u_i' his utility if he bids w_i'. Then, $u_i \geq u_i'$.*

PROOF Assume that by saying w_i he wins, and that the second highest (reported) value is p^*, then $u_i = w_i - p^* \geq 0$. Now, for an attempted manipulation $w_i' > p^*$, i would still win if he bids w_i' and would still pay p^*, thus $u_i' = u_i$. On the other hand, for $w_i' \leq p^*$, i would lose so $u_i' = 0 \leq u_i$.

If i loses by bidding w_i, then $u_i = 0$. Let j be the winner in this case, and thus $w_j \geq w_i$. For $w_i' < w_j$, i would still lose and so $u_i' = 0 = u_i$. For $w_i' \geq w_j$, i would win, but would pay w_j, thus his utility would be $u_i' = w_i - w_j \leq 0 = u_i$. □

This very simple and elegant idea achieves something that is quite remarkable: it reliably computes a function (argmax) of n numbers (the w_i's) that are each held secretly by a different self-interested player! Taking a philosophical point of view, this may be seen as the mechanics for the implementation of Adam Smith's *invisible hand*: despite private information and pure selfish behavior, social welfare is achieved. All the field of Mechanism Design is just a generalization of this possibility.

9.3.2 Incentive Compatible Mechanisms

In a world with money, our mechanisms will not only choose a social alternative but will also determine monetary payments to be made by the different players. The complete social choice is then composed of the alternative chosen as well as of the transfer of money. Nevertheless, we will refer to each of these parts separately, calling the alternative chosen the social choice, not including in this term the monetary payments.

Formally, a mechanism needs to socially choose some alternative from A, as well as to decide on payments. The preference of each player i is modeled by a valuation function $v_i : A \to \Re$, where $v_i \in V_i$. Throughout the rest of this chapter, $V_i \subseteq \Re^A$ is a commonly known set of possible valuation functions for player i.

Starting at this point and for the rest of this chapter, it will be convenient to use the following standard notation.

Notation Let $v = (v_1, \ldots, v_n)$ be an n-dimensional vector. We will denote the $(n-1)$-dimensional vector in which the i'th coordinate is removed by $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$. Thus we have three equivalent notations: $v = (v_1, \ldots, v_n) = (v_i, v_{-i})$. Similarly, for $V = V_1 \times \cdots \times V_n$, we will denote $V_{-i} = V_1 \times \cdots \times V_{i-1} \times V_{i+1} \times \cdots \times V_n$. Similarly we will use t_{-i}, x_{-i}, X_{-i}, etc.

Definition 9.14 A (direct revelation) *mechanism* is a social choice function $f : V_1 \times \cdots \times V_n \to A$ and a vector of payment functions $p_1, \ldots, p_n$, where $p_i : V_1 \times \cdots \times V_n \to \Re$ is the amount that player i pays.

The qualification "direct revelation" will become clear in Section 9.4, where we will generalize the notion of a mechanism further. We are now ready for the key definition in this area, *incentive compatibility* also called *strategy-proofness* or *truthfulness*.

Definition 9.15 A mechanism $(f, p_1, \ldots, p_n)$ is called incentive compatible if for every player i, every $v_1 \in V_1, \ldots, v_n \in V_n$ and every $v_i' \in V_i$, if we denote $a = f(v_i, v_{-i})$ and $a' = f(v_i', v_{-i})$, then $v_i(a) - p_i(v_i, v_{-i}) \geq v_i(a') - p_i(v_i', v_{-i})$.

Intuitively this means that player i whose valuation is v_i would prefer "telling the truth" v_i to the mechanism rather than any possible "lie" v_i', since this gives him higher (in the weak sense) utility.

9.3.3 Vickrey–Clarke–Groves Mechanisms

While in the general setting without money, as we have seen, nothing nontrivial is incentive compatible, the main result in this setting is positive and provides an incentive compatible mechanism for the most natural social choice function: optimizing the social welfare. The social welfare of an alternative $a \in A$ is the sum of the valuations of all players for this alternative, $\sum_i v_i(a)$.

Definition 9.16 A mechanism $(f, p_1, \ldots, p_n)$ is called a Vickrey–Clarke–Groves (VCG) mechanism if

- $f(v_1, \ldots, v_n) \in \operatorname{argmax}_{a \in A} \sum_i v_i(a)$; that is, f maximizes the social welfare, and
- for some functions $h_1, \ldots, h_n$, where $h_i : V_{-i} \to \Re$ (i.e., h_i does not depend on v_i), we have that for all $v_1 \in V_1, \ldots, v_n \in V_n$: $p_i(v_1, \ldots, v_n) = h_i(v_{-i}) - \sum_{j \neq i} v_j(f(v_1, \ldots, v_n))$.

The main idea lies in the term $-\sum_{j \neq i} v_j(f(v_1, \ldots, v_n))$, which means that each player is paid an amount equal to the sum of the values of all other players. When this term is added to his own value $v_i(f(v_1, \ldots, v_n))$, the sum becomes exactly the total social welfare of $f(v_1, \ldots, v_n)$. Thus this mechanism aligns all players' incentives with the social goal of maximizing social welfare, which is exactly archived by telling the truth. The other term in the payment $h_i(v_i)$ has no strategic implications for player i since it does not depend, in any way, on what he says, and thus from player i's point of view it is just a constant. Of course, the choice of h_i does change significantly how

much money is paid and in which direction, but we will postpone this discussion. What we have just intuitively explained is as follows.

Theorem 9.17 (Vickrey–Clarke–Groves) *Every VCG mechanism is incentive compatible.*

Let us prove it formally.

PROOF Fix i, v_{-i}, v_i, and v_i'. We need to show that for player i with valuation v_i, the utility when declaring v_i is not less than the utility when declaring v_i'. Denote $a = f(v_i, v_{-i})$ and $a' = f(v_i', v_{-i})$. The utility of i, when declaring v_i, is $v_i(a) + \sum_{j \neq i} v_j(a) - h_i(v_{-i})$, but when declaring v_i' is $v_i(a') + \sum_{j \neq i} v_j(a') - h_i(v_{-i})$. But since $a = f(v_i, v_{-i})$ maximizes social welfare over all alternatives, $v_i(a) + \sum_{j \neq i} v_j(a) \geq v_i(a') + \sum_{j \neq i} v_j(a')$ and thus the same inequality holds when subtracting the same term $h_i(v_{-i})$ from both sides. □

9.3.4 Clarke Pivot Rule

Let us now return to the question of choosing the "right" h_i's. One possibility is certainly choosing $h_i = 0$. This has the advantage of simplicity but usually does not make sense since the mechanism pays here a great amount of money to the players. Intuitively we would prefer that players pay money to the mechanism, but not more than the gain that they get. Here are two conditions that seem to make sense, at least in a setting where all valuations are nonnegative.

Definition 9.18

- A mechanism is (ex-post) *individually rational* if players always get nonnegative utility. Formally if for every $v_1, \ldots, v_n$ we have that $v_i(f(v_1, \ldots, v_n)) - p_i(v_1, \ldots, v_n) \geq 0$.
- A mechanism has no positive transfers if no player is ever paid money. Formally if for every $v_1, \ldots, v_n$ and every i, $p_i(v_1, \ldots, v_n) \geq 0$.

The following choice of h_i's provides the following two properties.

Definition 9.19 (Clarke pivot rule) The choice $h_i(v_{-i}) = \max_{b \in A} \sum_{j \neq i} v_i(b)$ is called the Clarke pivot payment. Under this rule the payment of player i is $p_i(v_1, \ldots, v_n) = \max_b \sum_{j \neq i} v_i(b) - \sum_{j \neq i} v_i(a)$, where $a = f(v_1, \ldots, v_n)$.

Intuitively, i pays an amount equal to the total damage that he causes the other players – the difference between the social welfare of the others with and without i's participation. In other words, the payments make each player internalize the externalities that he causes.

Lemma 9.20 *A VCG mechanism with Clarke pivot payments makes no positive transfers. If $v_i(a) \geq 0$ for every $v_i \in V_i$ and $a \in A$ then it is also individually rational.*

PROOF Let $a = f(v_1, \ldots, v_n)$ be the alternative maximizing $\sum_j v_j(a)$ and b be the alternative maximizing $\sum_{j \neq i} v_j(b)$. To show individual rationality, the utility of player i is $v_i(a) + \sum_{j \neq i} v_j(a) - \sum_{j \neq i} v_j(b) \geq \sum_j v_j(a) - \sum_j v_j(b) \geq 0$, where the first inequality is since $v_i(b) \geq 0$ and the second is since a was chosen as to maximize $\sum_j v_j(a)$. To show no positive transfers, note that $p_i(v_1, \ldots, v_n) = \sum_{j \neq i} v_i(b) - \sum_{j \neq i} v_i(a) \geq 0$, since b was chosen as to maximize $\sum_{j \neq i} v_j(b)$. □

As stated, the Clarke pivot rule does not fit many situations where valuations are negative; i.e., when alternatives have costs to the players. Indeed, with the Clarke pivot rule, players always pay money to the mechanism, while the natural interpretation in case of costs would be the opposite. The spirit of the Clarke pivot rule in such cases can be captured by a modified rule that chooses b as to maximize the social welfare "when i does not participate" where the exact meaning of this turns out to be quite natural in most applications.

9.3.5 Examples

9.3.5.1 Auction of a Single Item

The Vickrey auction that we started our discussion with is a special case of a VCG mechanism with the Clarke pivot rule. Here $A = \{i\text{–wins} | i \in I\}$. Each player has value 0 if he does not get the item, and may have any positive value if he does win the item, thus $V_i = \{v_i | v_i(i\text{–wins}) \geq 0 \text{ and } \forall j \neq i,\ v_i(j\text{–wins}) = 0\}$. Notice that finding the player with highest value is exactly equivalent to maximizing $\sum_i v_i(i)$ since only a single player gets nonzero value. VCG payments using the Clarke pivot rule give exactly Vickrey's second price auction.

9.3.5.2 Reverse Auction

In a reverse auction (procurement auction) the bidder wants to *procure* an item from the bidder with lowest cost. In this case the valuation spaces are given by $V_i = \{v_i | v_i(i\text{–wins}) \leq 0 \text{ and } \forall j \neq i\ v_i(j\text{–wins}) = 0\}$, and indeed procuring the item from the lowest cost bidder is equivalent to maximizing the social welfare. The natural VCG payment rule would be for the mechanism to pay to the lowest bidder an amount equal to the second lowest bid, and pay nothing to the others. This may be viewed as capturing the spirit of the pivot rule since the second lowest bid is what would happen "without i."

9.3.5.3 Bilateral Trade

In the bilateral trade problem a seller holds an item and values it at some $0 \leq v_s \leq 1$ and a potential buyer values it at some $0 \leq v_b \leq 1$. (The constants 0 and 1 are arbitrary and may be replaced with any commonly known constants $0 \leq v_l \leq v_h$.) The possible outcomes are $A = \{\text{no–trade}, \text{trade}\}$ and social efficiency implies that trade is chosen if $v_b > v_s$ and *no-trade* if $v_s > v_b$. Using VCG payments and decreeing that no payments be made in case of *no-trade*, implies that in case of trade the buyer pays v_s and the seller is paid v_b. Notice that since in this case $v_b > v_s$,

the mechanism subsidizes the trade. As we will see below in Section 9.5.5, this is unavoidable.

9.3.5.4 Multiunit Auctions

In a multiunit auction, k identical units of some good are sold in an auction (where $k < n$). In the simple case each bidder is interested in only a single unit. In this case $A = \{S\text{–wins}|S \subset I,\ |S| = k\}$, and a bidder's valuation v_i gives some fixed value v^* if i gets an item, i.e. $v_i(S) = v^*$ if $i \in S$ and $v_i(S) = 0$ otherwise. Maximizing social welfare means allocating the items to the k highest bidders, and in the VCG mechanism with the pivot rule, each of them should pay the $k + 1$'st highest offered price. (Losers pay 0.)

In a more general case, bidders may be interested in more than a single unit and have a different value for each number of units obtained. The next level of sophistication comes when the items in the auction are heterogeneous, and valuations can give a different value to each combination of items. This is called a combinatorial auction and is studied at length in Chapter 11.

9.3.5.5 Public Project

The government is considering undertaking a public project (e.g., building a bridge). The project has a commonly known cost C, and is valued by each citizen i at (a privately known) value v_i. (We usually think that $v_i \geq 0$, but the case of allowing $v_i < 0$, i.e., citizens who are hurt by the project is also covered.) Social efficiency means that the government will undertake this project iff $\sum_i v_i > C$. (This is not technically a subcase of our definition of maximizing the social welfare, since our definition did not assume any costs or values for the designer, but becomes so by adding an extra player "government" whose valuation space is the singleton valuation, giving cost C to undertaking the project and 0 otherwise.) The VCG mechanism with the Clarke pivot rule means that a player i with $v_i \geq 0$ will pay a nonzero amount only if he is pivotal: $\sum_{j \neq i} v_j \leq C$ but $\sum_j v_j > C$ in which case he will pay $p_i = C - \sum_{j \neq i} v_j$. (A player with $v_i < 0$ will make a nonzero payment only if $\sum_{j \neq i} v_j > C$ but $\sum_j v_j \leq C$ in which case he will pay $p_i = \sum_{j \neq i} v_j - C$.) One may verify that $\sum_i p_i < C$ (unless $\sum_i v_i = C$), and thus the payments collected do not cover the project's costs. As we will see in Section 9.5.5, this is unavoidable.

9.3.5.6 Buying a Path in a Network

Consider a communication network, modeled as a directed graph $G = (V, E)$, where each link $e \in E$ is owned by a different player, and has a cost $c_e \geq 0$ if his link is used for carrying some message. Suppose that we wish to procure a communication path between two specified vertices $s, t \in V$; i.e., the set of alternatives is the set of all possible $s - t$ paths in G, and player e has value 0 if the path chosen does not contain e and value $-c_e$ if the path chosen does contain e. Maximizing social welfare means finding the shortest path p (in terms of $\sum_{e \in p} c_e$). A VCG mechanism that makes no payments to edges that are not in p, will pay to each $e_0 \in p$ the quantity $\sum_{e \in p'} c_e - \sum_{e \in p - \{e_0\}} c_e$, where p is the shortest $s - t$ path in G and p' is the shortest

$s - t$ path in G that does not contain the edge e (for simplicity, assume that G is 2-edge connected so such a p' always exists). This corresponds to the spirit of the pivot rule since "without e" the mechanism can simply not use paths that contain e.

9.4 Implementation in Dominant Strategies

In this section our aim is to put the issue of incentive compatibility in a wider context. The mechanisms considered so far extract information from the different players by motivating them to "tell the truth." More generally, one may think of other, indirect, methods of extracting sufficient information from the participants. Perhaps one may devise some complex protocol that achieves the required social choice when players act strategically. This section will formalize these more general mechanisms, and the associated notions describing what happens when "players act strategically."

Deviating from the common treatment in economics, in this section we will describe a model that does not involve any distributional assumptions. Many of the classical results of Mechanism Design are captured in this framework, including most of the existing applications in computational settings. In Section 9.6 we will add this ingredient of distributional assumptions reaching the general "Bayesian" models.

9.4.1 Games with Strict Incomplete Information

How do we model strategic behavior of the players when they are missing some of the information that specifies the game? Specifically in our setting a player does not know the private information of the other players, information that determines their preferences. The standard setting in Game Theory supposes on the other hand that the "rules" of the game, including the utilities of all players, are public knowledge.

We will use a model of games with *independent private values* and *strict incomplete information*. Let us explain the terms: "independent private values" means that the utility of a player depends fully on his private information and not on any information of others as it is independent from his own information. *Strict incomplete information* is a (not completely standard) term that means that we will have no probabilistic information in the model. An alternative term sometimes used is "pre-Bayesian." From a CS perspective, it means that we will use a worst case analysis over unknown information. So here is the model.

Definition 9.21 A game with (independent private values and) strict incomplete information for a set of n players is given by the following ingredients:

(i) For every player i, a set of *actions* X_i.

(ii) For every player i, a set of *types* T_i. A value $t_i \in T_i$ is the private information that i has.

(iii) For every player i, a *utility function* $u_i : T_i \times X_1 \times \cdots \times X_n \to \Re$, where $u_i(t_i, x_1, \ldots, x_n)$ is the utility achieved by player i, if his type (private information) is t_i, and the profile of actions taken by all players is $x_1, \ldots, x_n$.

The main idea that we wish to capture with this definition is that each player i must choose his action x_i when knowing t_i but not the other t_j's. Note that the t_j's do not

affect his utility, but they do affect how the other players behave. Thus the interplay between the different x_i's is more delicate than in "regular" games. The total behavior of player i in such a setting is captured by a function that specifies which action x_i is taken for every possible type t_i – this is termed a strategy. It is these strategies that we want to be in equilibrium.

Definition 9.22

- A strategy of a player i is a function $s_i : T_i \to X_i$.
- A profile of strategies $s_1, \ldots, s_n$ is an ex-post-Nash equilibrium if for every $t_1, \ldots, t_n$ we have that the actions $s_1(t_1), \ldots, s_n(t_n)$ are in Nash equilibrium in the full information game defined by the t_i's. Formally: For all i, all $t_1, \ldots, t_n$, and all x_i' we have that $u_i(t_i, s_i(t_i), s_{-i}(t_{-i})) \geq u_i(t_i, x_i', s_{-i}(t_{-i}))$.
- A strategy s_i is a (weakly) dominant strategy if for every t_i we have that the action $s_i(t_i)$ is a dominant strategy in the full information game defined by t_i. Formally: for all t_i, all x_{-i} and all x_i' we have that $u_i(t_i, s_i(t_i), x_{-i}) \geq u_i(t_i, x_i', x_{-i})$. A profile $s_1, \ldots, s_n$ is called a dominant strategy equilibrium if each s_i is a dominant strategy.

Thus the notion of ex-post Nash requires that $s_i(t_i)$ is a best response to $s_i(t_{-i})$ for every possible value of t_{-i}, i.e., without knowing anything about t_{-i} but rather only knowing the forms of the other players' strategies s_{-i} as functions. The notion of dominant strategy requires that $s_i(t_i)$ is a best response to any x_{-i} possible, i.e., without knowing anything about t_{-i} or about s_{-i}. Both of these definitions seem too good to be true: how likely is it that a player has a single action that is a best response to all x_{-i} or even to all $s_{-i}(t_{-i})$? Indeed in usual cases one does not expect games with strict incomplete information to have any of these equilibria. However, in the context of Mechanism Design – where we get to design the game – we can sometimes make sure that they do exist.

While at first sight the notion of dominant strategy equilibrium seems much stronger than ex-post Nash, this is only due to actions that are never used.

Proposition 9.23 *Let $s_1, \ldots, s_n$ be an ex-post-Nash equilibrium of a game $(X_1, \ldots, X_n; T_1, \ldots, T_n; u_1, \ldots, u_n)$. Define $X_i' = \{s_i(t_i) | t_i \in T_i\}$ (i.e. X_i' is the actual range of s_i in X_i), then $s_1, \ldots, s_n$ is a dominant strategy equilibrium in the game $(X_1', \ldots, X_n'; T_1, \ldots, T_n; u_1, \ldots, u_n)$.*

PROOF Let $x_i = s_i(t_i) \in X_i', x_i' \in X_i'$, and for every $j \neq i$ $x_j \in X_j'$. By definition of X_j', for every $j \neq i$, there exists $t_j' \in T_j$ such that $s_j(t_j) = x_j$. Since $s_1, \ldots, s_n$ is an ex-post-Nash equilibrium, $u_i(t_i, s_i(t_i), s_{-i}(t_{-i})) \geq u_i(t_i, x_i', s_{-i}(t_{-i}))$, and as $x_{-i} = s_{-i}(t_{-i})$ we get exactly $u_i(t_i, s_i(t_i), x_{-i}) \geq u_i(t_i, x_i', x_{-i})$ as required in the definition of dominant strategies. □

9.4.2 Mechanisms

We are now ready to formalize the notion of a general – nondirect revelation – mechanism. The idea is that each player has some private information $t_i \in T_i$ that captures his

preference over a set of alternatives A; i.e., $v_i(t_i, a)$ is the value that player i assigns to a when his private information is t_i. We wish to "implement" some social choice function $F : T_1 \times \cdots \times T_n \to A$ that aggregates these preferences. We design a "mechanism" for this purpose: this will be some protocol for interaction with the players, specifying what each can "say" and what is done in each case. Formally, we can specify a set of possible actions X_i for each player, an outcome function $a : X_1 \times \cdots \times X_n \to A$ that chooses an alternative in A for each profile of actions, and payment functions $p : X_1 \times \cdots \times X_n \to \Re$ that specify the payment of each player for every profile of actions. Now the players are put in a game with strict incomplete information and we may expect them to reach an equilibrium point (if such exists).

Definition 9.24

- A mechanism for n players is given by (a) players' type spaces $T_1, \ldots, T_n$, (b) players' action spaces $X_1, \ldots, X_n$, (c) an alternative set A, (d) players' valuations functions $v_i : T_i \times A :\to \Re$, (e) an outcome function $a : X_1 \times \cdots \times X_n \to A$, and (f) payment functions $p_1, \ldots, p_n$, where $p_i : X_1 \times \cdots \times X_n \to \Re$. The game with strict incomplete information induced by the mechanism is given by using the types spaces T_i, the action spaces X_i, and the utilities $u_i(t_i, x_1, \ldots, x_n) = v_i(t_i, a(x_1, \ldots, x_n)) - p_i(x_1, \ldots, x_n)$.
- The mechanism implements a social choice function $f : T_1 \times \cdots \times T_n \to A$ in dominant strategies if for some dominant strategy equilibrium $s_1, \ldots, s_n$ of the induced game, where $s_i : T_i \to X_i$, we have that for all $t_1, \ldots, t_n$, $f(t_1, \ldots, t_n) = a(s_1(t_1), \ldots, s_n(t_n))$.
- Similarly we say that the mechanism implements f in ex-post-equilibrium if for some ex-post equilibrium $s_1, \ldots, s_n$ of the induced game we have that for all $t_1, \ldots, t_n$, $f(t_1, \ldots, t_n) = a(s_1(t_1), \ldots, s_n(t_n))$.

Clearly every dominant strategy implementation is also an ex-post-Nash implementation. Note that our definition only requires that for *some* equilibrium $f(t_1, \ldots, t_n) = a(s_1(t_1), \ldots, s_n(t_n))$ and allows other equilibria to exist. A stronger requirement would be that all equilibria have this property, or stronger still, that only a unique equilibrium point exists.

9.4.3 The Revelation Principle

At first sight it seems that the more general definition of mechanisms will allow us to do more than is possible using incentive compatible direct revelation mechanisms introduced in Section 9.3. This turns out to be false: any general mechanism that implements a function in dominant strategies can be converted into an incentive compatible one.

Proposition 9.25 (Revelation principle) *If there exists an arbitrary mechanism that implements f in dominant strategies, then there exists an incentive compatible mechanism that implements f. The payments of the players in the incentive compatible mechanism are identical to those, obtained at equilibrium, of the original mechanism.*

PROOF The proof is very simple: the new mechanism will simply simulate the equilibrium strategies of the players. That is, Let $s_1, \ldots, s_n$ be a dominant strategy equilibrium of the original mechanism, we define a new direct revelation mechanism: $f(t_1, \ldots, t_n) = a(s_1(t_1), \ldots, s_n(t_n))$ and $p_i'(t_1, \ldots, t_n) = p_i(s_1(t_1), \ldots, s_n(t_n))$. Now, since each s_i is a dominant strategy for player i, then for every t_i, x_{-i}, x_i' we have that $v_i(t_i, a(s_i(t_i), x_{-i})) - p_i(s_i(t_i), x_{-i}) \geq v_i(t_i, a(x_i', x_{-i})) - p_i(x_i', x_{-i})$. Thus in particular this is true for all $x_{-i} = s_{-i}(t_{-i})$ and any $x_i' = s_i(t_i')$, which gives the definition of incentive compatibility of the mechanism $(f, p_1', \ldots, p_n')$. □

Corollary 9.26 *If there exists an arbitrary mechanism that ex-post-Nash implements f, then there exists an incentive compatible mechanism that implements f. Moreover, the payments of the players in the incentive compatible mechanism are identical to those, obtained in equilibrium, of the original mechanism.*

PROOF We take the ex-post implementation and restrict the action space of each player, as in Proposition 9.23, to those that are taken, for some input type, in the ex-post equilibrium $s_1, \ldots, s_n$. Proposition 9.23 states that now $s_1, \ldots, s_n$ is a dominant strategy equilibrium of the game with the restricted spaces, and thus the mechanism with the restricted action spaces is an implementation in dominant strategies. We can now invoke the revelation principle to get an incentive compatible mechanism. □

The revelation principle does not mean that indirect mechanisms are useless. In particular, general mechanisms may be adaptive (multiround), significantly reducing the communication (or computation) burden of the players or of the auctioneer relative to a nonadaptive direct mechanism. An example is the case of combinatorial auctions studied in Chapter 11.

9.5 Characterizations of Incentive Compatible Mechanisms

In Section 9.3 we saw how to implement the most natural social choice function: maximization of the social welfare. The question that drives this section is: What other social choice functions can we implement? In economic settings, the main reasons for attempting implementations of other social choice functions are increasing the revenue or introducing some kind of fairness. In computerized settings there are many natural optimization goals and we would like to be able to implement each of them. For example, in scheduling applications, a common optimization goal is that of the "makespan" – completion time of the last job. This is certainly a social choice function that is very different than maximizing the total social welfare – how can it be implemented? Another major motivation for social choice functions that do not maximize social welfare comes from computational considerations. In many applications the set of alternatives A is complex, and maximizing social welfare is a hard computational problem (NP-complete). In many of these cases there are computationally efficient algorithms that *approximate* the maximum social welfare. Such an algorithm in effect gives a social choice function

that approximates social welfare maximization, but is different from it. Can it be implemented?

Chapter 12 and parts of Chapter 11 address these issues specifically. This section limits itself to laying the foundations by providing basic characterizations of implementable social choice functions and their associated payments.

Because of the revelation principle, we can restrict ourselves again to look at incentive compatible mechanisms. Thus, in this section we revert to the notation used in Subsection 9.3.3: A mechanism $M = (f, p_1, \ldots, p_n)$ over domain of preferences $V_1 \times \cdots \times V_n$ ($V_i \subseteq \Re^A$) is composed of a social choice function $f : V_1 \times \cdots \times V_n \to A$ and payment functions $p_1, \ldots, p_n$, where $p_i : V_1 \times \cdots \times V_n \to \Re$ is the amount that player i pays. In the rest of the section we will provide characterizations of when such mechanisms are incentive compatible.

9.5.1 Direct Characterization

We start by stating explicitly the required properties from an incentive compatible mechanism.

Proposition 9.27 *A mechanism is incentive compatible if and only if it satisfies the following conditions for every i and every v_{-i}:*

(i) *The payment p_i does not depend on v_i, but only on the alternative chosen $f(v_i, v_{-i})$. That is, for every v_{-i}, there exist prices $p_a \in \Re$, for every $a \in A$, such that for all v_i with $f(v_i, v_{-i}) = a$ we have that $p(v_i, v_{-i}) = p_a$.*

(ii) *The mechanism optimizes for each player. That is, for every v_i, we have that $f(v_i, v_{-i}) \in argmax_a(v_i(a) - p_a)$, where the quantification is over all alternatives in the range of $f(\cdot, v_{-i})$.*

PROOF (if part) Denote $a = f(v_i, v_{-i})$, $a' = f(v_i', v_{-i})$, $p_a = p(v_i, v_{-i})$, and $p_{a'} = p(v_i', v_{-i})$. The utility of i, when telling the truth, is $v_i(a) - p_a$, which is not less than the utility when declaring v_i', $v_i(a') - p_{a'}$, since the mechanism optimizes for i, i.e., $a = f(v_i, v_{-i}) \in \text{argmax}_a(v_i(a) - p_a)$.

(Only-if part; first condition) If for some v_i, v_i', $f(v_i, v_{-i}) = f(v_i', v_{-i})$ but $p_i(v_i, v_{-i}) > p_i(v_i', v_{-i})$ then a player with type v_i will increase his utility by declaring v_i'.

(Only-if part; second condition) If $f(v_i, v_{-i}) \notin \text{argmax}_a(v_i(a) - p_a)$, fix $a' \in \text{argmax}_a(v_i(a) - p_a)$ in the range of $f(\cdot, v_{-i})$, and thus for some v_i', $a' = f(v_i', v_{-i})$. Now a player with type v_i will increase his utility by declaring v_i'. □

9.5.2 Weak Monotonicity

The previous characterization involves both the social choice function and the payment functions. We now provide a partial characterization that only involves the social choice function. In Section 9.5.5 we will see that the social choice function usually determines the payments essentially uniquely.

Definition 9.28 A social choice function f satisfies *Weak Monotonicity (WMON)* if for all i, all v_{-i} we have that $f(v_i, v_{-i}) = a \neq b = f(v_i', v_{-i})$ implies that $v_i(a) - v_i(b) \geq v_i'(a) - v_i'(b)$.

That is, WMON means that if the social choice changes when a single player changes his valuation, then it must be because the player increased his value of the new choice relative to his value of the old choice.

Theorem 9.29 *If a mechanism $(f, p_1, \ldots, p_n)$ is incentive compatible, then f satisfies WMON. If all domains of preferences V_i are convex sets (as subsets of an Euclidean space) then for every social choice function that satisfies WMON there exists payment functions $p_1, \ldots, p_n$ such that $(f, p_1, \ldots, p_n)$ is incentive compatible.*

The first part of the theorem is easy and we will bring it completely, the second part is quite involved, and will not be given here. It is known that WMON is not a sufficient condition for incentive compatibility in general nonconvex (more precisely, nonsimply connected) domains.

PROOF (First part) Assume first that $(f, p_1, \ldots, p_n)$ is incentive compatible, and fix i and v_{-i} in an arbitrary manner. Proposition 9.27 implies the existence of fixed prices p_a for all $a \in A$ (that do not depend on v_i) such that whenever the outcome is a then bidder i pays exactly p_a. Now assume $f(v_i, v_{-i}) = a \neq b = f(v_i', v_{-i})$. Since a player with valuation v_i does not prefer declaring v_i' we have that $v_i(a) - p_a \geq v_i(b) - p_b$. Similarly since a player with valuation v_i' does not prefer declaring v_i we have that $v_i'(a) - p_a \leq v_i'(b) - p_b$. Subtracting the second inequality from the first, we get $v_i(a) - v_i(b) \geq v_i'(a) - v_i'(b)$, as required. □

While WMON gives a pretty tight characterization of implementable social choice functions, it still leaves something to be desired as it is not intuitively clear what exactly the WMON functions are. The problem is that the WMON condition is a local condition for each player separately and for each v_{-i} separately. Is there a global characterization? This turns out to depend intimately on the domains of preferences V_i. For two extreme cases there are good global characterizations: when V_i is "unrestricted" i.e. $V_i = \Re^A$, and when V_i is severely restricted as to be essentially single dimensional. These two cases are treated in the next two subsections below. The intermediate range where the V_i's are somewhat restricted, a range in which most computationally interesting problems lie is still wide open. More on this appears in Chapter 12.

9.5.3 Weighted VCG

It turns out that when the domain of preferences is unrestricted, then the only incentive compatible mechanisms are simple variations of the VCG mechanism. These variations allow giving weights to the players, weights to the alternatives, and allow restricting the range. The resulting social choice function is an "affine maximizer":

Definition 9.30 A social choice function f is called an *affine maximizer* if for some subrange $A' \subset A$, for some player weights $w_1, \ldots, w_n \in \Re^+$ and for some outcome weights $c_a \in \Re$ for every $a \in A'$, we have that $f(v_1, \ldots, v_n) \in \text{argmax}_{a \in A'}(c_a + \sum_i w_i v_i(a))$.

It is easy to see that VCG mechanisms can be generalized to affine maximizers:

Proposition 9.31 *Let f be an affine maximizer. Define for every i, $p_i(v_1, \ldots, v_n) = h_i(v_{-i}) - \sum_{j \neq i}(w_j/w_i)v_j(a) - c_a/w_i$, where h_i is an arbitrary function that does not depend on v_i. Then, $(f, p_1, \ldots, p_n)$ is incentive compatible.*

PROOF First, we can assume wlog that $h_i = 0$. The utility of player i if alternative a is chosen is $v_i(a) + \sum_{j \neq i}(w_j/w_i)v_j(a) + c_a/w_i$. By multiplying by $w_i > 0$, this expression is maximized when $c_a + \sum_j w_j v_j(a)$ is maximized which is what happens when i reports v_i truthfully. □

Roberts' theorem states that for unrestricted domains with at least 3 possible outcomes, these are the only incentive compatible mechanisms.

Theorem 9.32 (Roberts) *If $|A| \geq 3$, f is onto A, $V_i = \Re^A$ for every i, and $(f, p_1, \ldots, p_n)$ is incentive compatible then f is an affine maximizer.*

The proof of this theorem is not trivial and is given in Chapter 12. It is easy to see that the restriction $|A| \geq 3$ is crucial (as in Arrow's theorem), since the case $|A| = 2$ falls into the category of "single parameter" domains discussed below, for which there do exist incentive compatible mechanisms beyond weighted VCG. It remains open to what extent can the restriction of $V_i = \Re^A$ be relaxed.

9.5.4 Single-Parameter Domains

The unrestricted case $V_i = \Re^A$ basically means that the valuation space has full dimensionality. The opposite case is when the space V_i is single-dimensional; i.e., there is a single real parameter that directly determines the whole vector v_i. There are several possible levels of generality in which to formalize this, and we will consider one of intermediate generality that is simple and yet suffices for most applications. In our setting each bidder has a private scalar value for "winning," with "losing" having value of 0. This is modeled by some commonly known subset of winning alternatives $W_i \subseteq A$. The main point is that all winning alternatives are equivalent to each other for player i; and similarly all losing outcomes are equivalent to each other. All the examples in Section 9.3.5 fall into this category. A simple example is an auction of one item where W_i is the single outcome where i wins. A more complex example is the setting of buying a path in a network (Subsection 9.3.5.6), where W_i is the set of all paths that contain edge i.

Definition 9.33 A single parameter domain V_i is defined by a (publicly known) $W_i \subset A$ and a range of values $[t^0, t^1]$. V_i is the set of v_i such that for some

$t^0 \leq t \leq t^1$, $v_i(a) = t$, for all $a \in W_i$ and $v_i(a) = 0$ for all $a \notin W_i$. In such settings we will abuse notation and use v_i as the scalar t.

For this setting it is quite easy to completely characterize incentive compatible mechanisms.

Definition 9.34 A social choice function f on a single parameter domain is called monotone in v_i if for every v_{-i} and every $v_i \leq v_i' \in \Re$ we have that $f(v_i, v_{-i}) \in W_i$ implies that $f(v_i', v_{-i}) \in W_i$. That is, if valuation v_i makes i win, then so will every higher valuation $v_i' \geq v_i$.

For a monotone function f, for every v_{-i} for which player i can both win and lose, there is always a *critical value* below which i loses and above which he wins. For example, in a second price auction the critical value for each player is highest declared value among the other players.

Definition 9.35 The critical value of a monotone social choice function f on a single parameter domain is $c_i(v_{-i}) = \sup_{v_i : f(v_i, v_{-i}) \notin W_i} v_i$. The critical value at v_{-i} is undefined if $\{v_i | f(v_i, v_{-i}) \notin W_i\}$ is empty.

We will call a mechanism on a single parameter domain "normalized" if the payment for losing is always 0, i.e., for every v_i, v_{-i} such that $f(v_i, v_{-i}) \notin W_i$ we have that $p_i(v_i, v_{-i}) = 0$. It is not difficult to see that every incentive compatible mechanism may be easily turned into a normalized one, so it suffices to characterize normalized mechanisms.

Theorem 9.36 *A normalized mechanism $(f, p_1, \ldots, p_n)$ on a single parameter domain is incentive compatible if and only if the following conditions hold:*

(i) *f is monotone in every v_i.*

(ii) *Every winning bid pays the critical value. (Recall that losing bids pay 0.) Formally, For every i, v_i, v_{-i} such that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = c_i(v_{-i})$. (If $c_i(v_{-i})$ is undefined we require instead that for every v_{-i}, there exists some value c_i, such that $p_i(v_i, v_{-i}) = c_i$ for all v_i such that $f(v_i, v_{-i}) \in W_i$.)*

PROOF (If part) Fix i, v_{-i}, v_i. For every declaration made by i, if he wins his utility is $v_i - c_i(v_{-i})$ and if he loses his utility is 0. Thus he prefers winning if $v_i > c_i(v_{-i})$ and losing if $v_i < c_i(v_{-i})$, which is exactly what happens when he declares the truth.

(Only-if part, first condition) If f is not monotone then for some $v_i' > v_i$ we have that $f(v_i', v_{-i})$ loses while $f(v_i, v_{-i})$ wins and pays some amount $p = p_i(v_i, v_{-i})$. Since a bidder with value v_i is not better off bidding v_i' and losing we have that $v_i - p \geq 0$. Since a bidder with value v_i' is not better off bidding v_i and winning we have that $v_i' - p \leq 0$. Contradiction.

(Only-if part, second condition) Assume that some winning v_i pays $p > c_i(v_{-i})$ then, using Proposition 9.27, all winning bids will make the same payment,

including a winning v_i' with $c_i(v_{-i}) < v_i' < p$. But such a bidder is better off losing which he can do by bidding some value $v^{\text{lose}} < c(v_{-i})$. In the other direction if v_i pays $p < c(v_{-i})$ then a losing v_i' with $c(v_{-i}) > v_i' > p$ is better of wining and paying p, which will happen if he bids v_i. □

Notice that this characterization leaves ample space for non-affine-maximization. For example we can implement social functions such as maximizing the euclidean norm $\text{argmax}_a \sum_i v_i(a)^2$ or maximizing the minimum value $\text{argmax}_a \min_i v_i(a)$. Indeed in many cases this flexibility allows the design of computationally efficient *approximation* mechanisms for problems whose exact optimization is computationally intractable – an example is given in Chapter 12.

9.5.5 Uniqueness of Prices

This section has so far focused on characterizing the implementable social choice functions. What about the payment functions? It turns out that the payment function is essentially uniquely determined by the social choice function. "Essentially" means that if we take an incentive compatible mechanisms with payments p_i and modify the payments to $p_i'(v_1, \ldots, v_n) = p_i(v_1, \ldots, v_n) + h_i(v_{-i})$ for an arbitrary function h_i that does not depend on v_i, then incentive compatibility remains. It turns out that this is the only leeway in the payment.

Theorem 9.37 *Assume that the domains of preference V_i are connected sets in the usual metric in the Euclidean space. Let $(f, p_1, \ldots, p_n)$ be an incentive compatible mechanism. The mechanism with modified payments $(f, p_1', \ldots, p_n')$ is incentive compatible if and only if for some functions $h_i : V_{-i} \to \Re$ we have that $p_i'(v_1, \ldots, v_n) = p_i(v_1, \ldots, v_n) + h_i(v_{-i})$ for all $v_1, \ldots, v_n$.*

PROOF The "if" part is clear since h_i has no strategic implications for player i, so we need only prove the only-if part. Assume that $(f, p_1', \ldots, p_n')$ is incentive compatible, and for the rest of the proof fix some i and some v_{-i}.

For every $a \in A$ denote $V^a = \{v_i \in V_i | f(v_i, v_{-i}) = a\}$. Using Proposition 9.27, the payment $p(v_i, v_{-i})$ is identical for all $v_i \in V^a$ and will be denoted by p_a. Similarly we denote $p_a' = p'(v_i, v_{-i})$ for some $v_i \in V^a$. It now suffices to show that for every $a, b \in A$, $p_a - p_b = p_a' - p_b'$.

For $a, b \in A$ we will say that a and b are *close* if for every $\epsilon > 0$ there exist $v_i^a, v_i^b \in V_i$ such that $||v_i^a - v_i^b|| = max_{c \in A} |v_i^a(c) - v_i^b(c)| \leq \epsilon$, and $f(v_i^a, v_{-i}) = a$ and $f(v_i^b, v_{-i}) = b$. We will first prove the required $p_a - p_b = p_a' - p_b'$ for close a, b. Fix $v_i^a, v_i^b \in V_i$ as in the definition of closeness. Since a bidder with type v_i^a does not gain by declaring v_i^b with payments p, we have that $v_i^a(a) - p_a \geq v_i^a(b) - p_b$, and since a bidder with v_i^b does not gain by declaring v_i^a we have that $v_i^b(a) - p_a \leq v_i^b(b) - p_b$. Putting together and rearranging we have that $v_i^a(b) - v_i^a(a) \leq p_b - p_a \leq v_i^b(b) - v_i^b(a)$. Similarly, by considering the mechanism with payments p' we have $v_i^a(b) - v_i^a(a) \leq p_b' - p_a' \leq v_i^b(b) - v_i^b(a)$. But now recall that $||v_i^a - v_i^b|| \leq \epsilon$ and thus the upper bound and the lower bound for $p_b - p_a$

and for $p'_b - p'_a$ are at most 2ϵ apart and thus $|(p_b - p_a) - (p'_b - p'_a)| \leq 2\epsilon$. Since ϵ was arbitrary $p_b - p_a = p'_b - p'_a$.

To show $p_b - p_a = p'_b - p'_a$ for general (not necessarily close) a and b, consider $B = \{b \in A | p_b - p_a = p'_b - p'_a\}$. Since $p_b - p_a = p'_b - p'_a$ and $p_c - p_b = p'_c - p'_b$ implies $p_c - p_a = p'_c - p'_a$ we have that no alternative in $A - B$ can be close to any alternative in B. Thus $V^B = \bigcup_{b \in B} V^b$ has positive distance from its complement $V^{A-B} = \bigcup_{b \notin B} V^b$ contradicting the connectedness of V. □

It is not difficult to see that the assumption that V_i is connected is essential, as for example, if the valuations are restricted to be integral, then modifying p_i by any small constants $\epsilon < 1/2$ will not modify incentive compatibility.

From this, and using the revelation principle, we can directly get many corollaries:

(i) The only incentive compatible mechanisms that maximize social welfare are those with VCG payments.

(ii) In the bilateral trade problem (Section 9.3.5.3) the only incentive compatible mechanism that maximizes social welfare and makes no payments in case of no-trade is the one shown there which subsidizes the trade. More generally, if a mechanism for bilateral trade satisfies ex-post individual rationality, then it cannot dictate positive payments from the players in case of no-trade and thus it must subsidize trade.

(iii) In the public project problem (Section 9.3.5.5) no ex-post individually rational mechanism that maximizes social welfare can recover the cost of the project. Again, the uniqueness of payments implies that if players with value 0 pay 0 (which is as much as they can pay maintaining individual rationality) then their payments in case of building the project must be identical to those obtained using the Clarke pivot rule.

In Section 9.6.3 we will see a similar theorem in the Bayesian setting, a theorem that will strengthen all of these corollaries as well to that setting.

9.5.6 Randomized Mechanisms

All of our discussion so far considered only deterministic mechanisms. It is quite natural to allow also randomized mechanisms. Such mechanisms would be allowed to produce a distribution over alternatives and a distribution over payments. Alternatively, but specifying slightly more structure, we can allow distributions over deterministic mechanisms. This will allow us to distinguish between two notions of incentive compatibility.

Definition 9.38

- A *randomized mechanism* is a distribution over deterministic mechanisms (all with the same players, types spaces V_i, and outcome space A).
- A randomized mechanism is incentive compatible in the *universal sense* if every deterministic mechanism in the support is incentive compatible.

- A randomized mechanism is incentive compatible in *expectation* if truth is a dominant strategy in the game induced by expectation. That is, if for all i, all v_i, v_{-i}, and v_i', we have that $E[v_i(a) - p_i] \geq E[v_i(a') - p_i']$, where (a, p_i), and (a', p_i') are random variables denoting the outcome and payment when i bids, respectively, v_i and v_i', and $E[\cdot]$ denotes expectation over the randomization of the mechanism.

It is clear that incentive compatibility in the universal sense implies incentive compatibility in expectation. For most purposes incentive compatibility in expectation seems to be the more natural requirement. The universal definition is important if players are not risk neutral (which we do not consider in this chapter) or if the mechanism's internal randomization is not completely hidden from the players. As we will see in Chapters 12 and 13 randomized mechanisms can often be useful and achieve more than deterministic ones.

We will now characterize randomized incentive compatible mechanisms over single parameter domains. Recall the single parameter setting and notations from Section 9.5.4. We will denote the probability that i wins by $w_i(v_i, v_{-i}) = Pr[f(v_i, v_{-i}) \in W_i]$ (probability taken over the randomization of the mechanism) and will use $p_i(v_i, v_{-i})$ to directly denote the expected payment of i. In this notation the utility of player i with valuation v_i when declaring v_i' is $v_i \cdot w(v_i', v_{-i}) - p_i(v_i', v_{-i})$. For ease of notation we will focus on normalized mechanisms in which the lowest bid $v_i^0 = t^0$ loses completely $w_i(v_i^0, v_{-i}) = 0$ and pays nothing $p_i(v_i^0, v_{-i}) = 0$.

Theorem 9.39 *A normalized randomized mechanism in a single parameter domain is incentive compatible in expectation if and only if for every i and every fixed v_{-i} we have that*

(i) *the function $w_i(v_i, v_{-i})$ is monotonically non decreasing in v_i and*

(ii) $p_i(v_i, v_{-i}) = v_i \cdot w(v_i, v_{-i}) - \int_{v_i^0}^{v_i} w(t, v_{-i})dt$.

PROOF In the proof we will simplify notation by removing the index i and the fixed argument v_{-i} everywhere. In this notation, to show incentive compatibility we need to establish that $vw(v) - p(v) \geq vw(v') - p(v')$ for every v'. Plugging in the formula for p we get $\int_{v^0}^{v} w(t)dt \geq \int_{v^0}^{v'} w(t)dt - (v' - v)w(v')$. For $v' > v$ this is equivalent to $(v' - v)w(v') \geq \int_{v}^{v'} w(t)dt$, which is true due to the monotonicity of w. For $v' < v$ we get $(v - v')w(v') \leq \int_{v'}^{v} w(t)dt$, which again is true due to the monotonicity of w.

In the other direction, combining the incentive constraint at v, $vw(v) - p(v) \geq vw(v') - p(v')$, with the incentive constraint at v', $v'w(v) - p(v) \leq v'w(v') - p(v')$, and subtracting the inequalities, we get $(v' - v)w(v) \leq (v' - v)w(v')$ which implies monotonicity of w.

To derive the formula for p, we can rearrange the two incentive constraints as

$$v \cdot (w(v') - w(v)) \leq p(v') - p(v) \leq v' \cdot (w(v') - w(v)).$$

Now by letting $v' = v + \epsilon$, dividing throughout by ϵ, and taking the limit, both sides approach the same value, $v \cdot dw/dv$, and we get $dp/dv = v \cdot dw/dv$.

Thus, taking into account the normalization condition $p(v^0) = 0$, we have that $p(v_i) = \int_{v^0}^{v_i} v \cdot w'(v)dv$, and integrating by parts completes the proof. (This seems to require the differentiability of w, but as w is monotone this holds almost everywhere, which suffices since we immediately integrate.) □

We should point out explicitly that the randomization in a randomized mechanism is completely controlled by the mechanism designer and has nothing to do with any distributional assumptions on players' valuations as will be discussed in the next section.

9.6 Bayesian–Nash Implementation

So far in this chapter we have considered only implementation in dominant strategies (and the very similar ex-post-Nash). As mentioned in Section 9.4 this is usually considered too strict a definition in economic theory. It models situations where each player has no information at all about the private information of the others – not even a prior distribution – and must operate under a "worst case" assumption. The usual working definition in economic theory takes a Bayesian approach, assumes some commonly known prior distribution, and assumes that a player that lacks some information will optimize in a Bayesian sense according to the information that he does have. The formalization of these notions, mostly by Harsanyi, was a major development in economic theory in the 1960s and 1970s, and is certainly still the dominant approach to handling lack of information in economic theory. In this section we will give these basic notions in the context of mechanism design, again limiting ourselves to settings with independent private values.

9.6.1 Bayesian–Nash Equilibrium

Definition 9.40 A game with (independent private values and) incomplete information on a set of n players is given by the following ingredients:

(i) For every player i, a set of *actions* X_i.

(ii) For every player i, a set of *types* T_i, and a prior distribution D_i on T_i. A value $t_i \in T_i$ is the private information that i has, and $D_i(t_i)$ is the a priori probability that i gets type t_i.

(iii) For every player i, a *utility function* $u_i : T_i \times X_1 \times \cdots \times X_n \to \Re$, where $u_i(t_i, x_1, \ldots, x_n)$ is the utility achieved by player i, if his type (private information) is t_i, and the profile of actions taken by all players is $x_1, \ldots, x_n$.

The main idea that we wish to capture with this definition is that each player i must choose his action x_i when knowing t_i but not the other t_j's but rather only knowing the prior distribution D_j on each other t_j. The behavior of player i in such a setting is captured by a function that specifies which action x_i is taken for every possible type t_i – this is termed a strategy. It is these strategies that we would want to be in equilibrium.

Definition 9.41 A strategy of a player i is a function $s_i : T_i \to X_i$. A profile of strategies $s_1, \ldots, s_n$ is a *Bayesian-Nash equilibrium* if for every player i and every

t_i we have that $s_i(t_i)$ is the best response that i has to $s_{-i}()$ when his type is t_i, in expectation over the types of the other players. Formally: For all i, all t_i, and all x_i': $E_{D_{-i}}[u_i(t_i, s_i(t_i), s_{-i}(t_{-i}))] \geq E_{D_{-i}}[u_i(t_i, x_i', s_{-i}(t_{-i}))]$ (where $E_{D_{-i}}[]$ denotes the expectation over the other types t_{-i} being chosen according to distribution D_{-i}).

This now allows us to define implementation in the Bayesian sense.

Definition 9.42 A Bayesian mechanism for n players is given by (a) players' type spaces $T_1, \ldots, T_n$ and prior distributions on them $D_1, \ldots, D_n$, (b) players' action spaces $X_1, \ldots, X_n$, (c) an alternative set A, (d) players' valuations functions $v_i : T_i \times A :\to \Re$, (e) an outcome function $a : X_1 \times \cdots \times X_n \to A$, and (f) payment functions $p_1, \ldots, p_n$, where $p_i : X_1 \times \cdots \times X_n \to \Re$.

The game with incomplete information induced by the mechanism is given by using the type spaces T_i with prior distributions D_i, the action spaces X_i, and the utilities $u_i(t_i, x_1, \ldots, x_n) = v_i(t_i, a(x_1, \ldots, x_n)) - p_i(x_1, \ldots, x_n)$.

The mechanism implements a social choice function $f : T_1 \times \cdots \times T_n \to A$ in the Bayesian sense if for some Bayesian–Nash equilibrium $s_1, \ldots, s_n$ of the induced game ($s_i : T_i \to X_i$) we have that for all $t_1, \ldots, t_n$, $f(t_1, \ldots, t_n) = a(s_1(t_1), \ldots, s_n(t_n))$.

In particular it should be clear that every ex-post-Nash implementation is by definition also a Bayesian implementation for any distributions D_i. In general, however, being a Bayesian implementation depends on the distributions D_i and there are many cases where a Bayesian–Nash equilibrium exists even though no dominant-strategy one does. A simple example – a first price auction – is shown in the next subsection. Just like in the case of dominant-strategy implementations, Bayesian implementations can also be turned into ones that are truthful in a Bayesian sense.

Definition 9.43 A mechanism is truthful in the Bayesian sense if (a) it is "direct revelation"; i.e., the type spaces are equal to the action spaces $T_i = X_i$, and (b) the truthful strategies $s_i(t_i) = t_i$ are a Bayesian–Nash equilibrium.

Proposition 9.44 (Revelation principle) *If there exists an arbitrary mechanism that implements f in the Bayesian sense, then there exists a truthful mechanism that implements f in the Bayesian sense. Moreover, the expected payments of the players in the truthful mechanism are identical to those, obtained in equilibrium, in the original mechanism.*

The proof is similar to the proof of the same principle in the dominant-strategy setting given in Proposition 9.25.

9.6.2 First Price Auction

As an example of Bayesian analysis we study the standard first price auction in a simple setting: a single item is auctioned between two players, Alice and Bob. Each has a private value for the item: a is Alice's value and b is Bob's value. While we

already saw that a second price auction will allocate the item to the one with higher value, here we ask what would happen if the auction rules are the usual first-price ones: the highest bidder pays his bid. Certainly Alice will not bid a since if she does even if she wins her utility will be 0. She will thus need to bid some $x < a$, but how much lower? If she knew that Bob would bid y, she would certainly bid $x = y + \epsilon$ (as long as $x \leq a$). But she does not know y or even b which y would depend on – she only knows the distribution D_{Bob} over b.

Let us now see how this situation falls in the Bayesian–Nash setting described above: The type space T_{Alice} of Alice and T_{Bob} of Bob is the nonnegative real numbers, with t_{Alice} denoted by a and t_{Bob} denoted by b. The distributions over the type space are D_{Alice} and D_{Bob}. The action spaces X_{Alice} and X_{Bob} are also the non-negative real numbers, with x_{Alice} denoted by x and x_{Bob} denoted by y. The possible outcomes are {Alice-wins, Bob-wins}, with $v_{\text{Alice}}(\text{Bob-wins}) = 0$ and $v_{\text{Alice}}(\text{Alice-wins}) = a$ (and similarly for Bob). The outcome function is that Alice-wins if $x \geq y$ and Bob-wins otherwise (we arbitrarily assume here that ties are broken in favor of Alice). Finally, the payment functions are $p_{\text{Alice}} = 0$ whenever Bob-wins and $p_{\text{Alice}} = x$ whenever Alice-wins , while $p_{\text{Bob}} = y$ whenever Bob-wins and $p_{\text{Bob}} = 0$ whenever Alice-wins. Our question translates into finding the Bayesian–Nash equilibrium of this game. Specifically we wish to find a strategy s_{Alice} for Alice, given by a function $x(a)$, and a strategy s_{Bob} for Bob, given by the function $y(b)$, that are in Bayesian equilibrium, i.e., are best-replies to each other.

In general, finding Bayesian–Nash equilibria is not an easy thing. Even for this very simple first price auction the answer is not clear for general distributions D_{Alice} and D_{Bob}. However, for the symmetric case where $D_{\text{Alice}} = D_{\text{Bob}}$, the situation is simpler and a closed form expression for the equilibrium strategies may be found. We will prove it for the special case of uniform distributions on the interval [0, 1]. Similar arguments work for arbitrary nonatomic distributions over the valuations as well as for any number of bidders.

Lemma 9.45 *In a first price auction among two players with prior distributions of the private values a, b uniform over the interval* [0, 1]*, the strategies $x(a) = a/2$ and $y(b) = b/2$ are in Bayesian–Nash equilibrium.*

Note that in particular $x < y$ if and only if $a < b$ thus the winner is also the player with highest private value. This means that the first price auction also maximizes social welfare, just like a second-price auction.

PROOF Let us consider which bid x is Alice's optimal response to Bob's strategy $y = b/2$, when Alice has value a. The utility for Alice is 0 if she loses and $a - x$ if she wins and pays x, thus her expected utility from bid x is given by $u_{\text{Alice}} = Pr[\text{Alice wins with bid } x] \cdot (a - x)$, where the probability is over the prior distribution over b. Now Alice wins if $x \geq y$, and given Bob's strategy $y = b/2$, this is exactly when $x \geq b/2$. Since b is distributed uniformly in [0, 1] we can readily calculate this probability: $2x$ for $0 \leq x \leq 1/2$, 1 for $x \geq 1/2$, and 0 for $x \leq 0$. It is easy to verify that the optimal value of x is indeed in the range $0 \leq x \leq 1/2$ (since $x = 1/2$ is clearly better than any $x > 1/2$, and since any

$x < 0$ will give utility 0). Thus, to optimize the value of x, we need to find the maximum of the function $2x(a - x)$ over the range $0 \leq x \leq 1/2$. The maximum may be found by taking the derivative with respect to x and equating it to 0, which gives $2a - 4x = 0$, whose solution is $x = a/2$ as required. □

9.6.3 Revenue Equivalence

Let us now attempt comparing the first price auction and the second price auction. The social choice function implemented is exactly the same: giving the item to the player with highest private value. How about the payments? Where does the auctioneer get higher revenue? One can readily express the revenue of the second-price auction as $min(a, b)$ and the revenue of the first-price auction as $max(a/2, b/2)$, and it is clear that each of these expressions is higher for certain values of a and b.

But which is better on the average – in expectation over the prior distributions of a and b? Simple calculations will reveal that the expected value of $min(a, b)$ when a and b are chosen uniformly in [0, 1] is exactly 1/3. Similarly the expected value of $max(a/2, b/2)$ when a and b are chosen uniformly in [0, 1] is also exactly 1/3. Thus both auctions generate equivalent revenue in expectation! This is no coincidence. It turns out that in quite general circumstances every two Bayesian–Nash implementations of the same social choice function generate the same expected revenue.

Theorem 9.46 (The Revenue Equivalence Principle) *Under certain weak assumptions (to be detailed in the proof body), for every two Bayesian–Nash implementations of the same social choice function f, we have that if for some type t_i^0 of player i, the expected (over the types of the other players) payment of player i is the same in the two mechanisms, then it is the same for every value of t_i. In particular, if for each player i there exists a type t_i^0 where the two mechanisms have the same expected payment for player i, then the two mechanisms have the same expected payments from each player and their expected revenues are the same.*

Thus, for example, all single-item auctions that allocate (in equilibrium) the item to the player with highest value and in which losers pay 0, will have identical expected revenue.

The similarity to Theorem 9.37 should be noted: in both cases it is shown that the allocation rule determines the payments, up to a normalization. In the case of dominant strategy implementation, this is true for every fixed type of the other players, while in the case of Bayesian–Nash implementation, this is true in expectation over that types of the others. The proofs of the two theorems look quite different due to technical reasons. The underlying idea is the same: take two "close" types, then the equations specifying that for neither type does a player gain by misrepresenting himself as the other type, put together, determine the difference in payments in terms of the social choice function.

PROOF Using the revelation principle, we can first limit ourselves to mechanisms that are truthful in the Bayesian–Nash sense. Let us denote by V_i the space of valuation functions $v_i(t_i, \cdot)$ over all t_i.

Assumption 1 Each V_i is convex. (Note that this holds for essentially every example we had so far. This condition can be replaced by path-connectedness, and the proof becomes just slightly messier.)

Take any type $t_i^1 \in T_i$. We will derive a formula for the expected payment for this type that depends only on the expected payment for type t_i^0 and on the social choice function f. Thus any two mechanisms that implement the same social choice function and have identical expected payments at t_i^0 will also have identical expected payments at t_i^1. For this, let us now introduce some notations:

- v^0 is the valuation $v(t_i^0, \cdot)$. v^1 is the valuation $v(t_i^1, \cdot)$. We will look at these as vectors (in $V_i \subseteq \Re^A$), and look at their convex combinations $v^\lambda = v^0 + \lambda(v_1 - v_0)$. The convexity of V_i implies that $v^\lambda \in V_i$ and thus there exists some type t_i^λ such that $v^\lambda = v(t_i^\lambda, \cdot)$.
- p^λ is the expected payment of player i at type t_i^λ: $p^\lambda = E_{t_{-i}} p_i(t_i, t_{-i})$.
- w^λ is the probability distribution of $f(t_i^\lambda, \cdot)$, i.e., for every $a \in A$ $w^\lambda(a) = Pr_{t_{-i}}[f(t_i^\lambda, t_{-i}) = a]$.

Assumption 2 w^λ is continuously differentiable in λ. (This assumption is not really needed, but allows us to simply take derivatives and integrals as convenient.)

Once we have this notation in place, the proof is easy. Note that under these notations the expected utility of player i with type t_i^λ that declares $t_i^{\lambda'}$ is given by the expression $v^\lambda \cdot w^{\lambda'} - p^{\lambda'}$. Since a player with type t_i^λ prefers reporting the truth rather than $t_i^{\lambda+\epsilon}$ we have that $v^\lambda \cdot w^\lambda - p^\lambda \geq v^\lambda \cdot w^{\lambda+\epsilon} - p^{\lambda+\epsilon}$. Similarly, a player with type $t_i^{\lambda+\epsilon}$ prefers reporting the truth rather than t_i^λ, so we have $v^{\lambda+\epsilon} \cdot w^\lambda - p^\lambda \leq v^{\lambda+\epsilon} \cdot w^{\lambda+\epsilon} - p^{\lambda+\epsilon}$. Re-arranging and putting together, we get

$$v^\lambda(w^{\lambda+\epsilon} - w^\lambda) \leq p^{\lambda+\epsilon} - p^\lambda \leq v^{\lambda+\epsilon}(w^{\lambda+\epsilon} - w^\lambda)$$

Now divide throughout by ϵ and let ϵ approach 0. $v^{\lambda+\epsilon}$ approaches v^λ, $(w^{\lambda+\epsilon} - w^\lambda)/\epsilon$ approaches the vector $dw^\lambda/d\lambda = w'(\lambda)$ and thus we get that $(p^{\lambda+\epsilon} - p^\lambda)/\epsilon)$ approaches $v^\lambda \cdot w'(\lambda)$, and thus the derivative of p^λ is defined and is continuous. Integrating, we get $p^1 = p^0 + \int_0^1 v^\lambda \cdot w'(\lambda) d\lambda$. □

Thus the revenue equivalence theorem tells us that we cannot increase revenue without changing appropriately the allocation rule (social choice function) itself. In particular, all the corollaries in Section 9.5.5 apply, in the sense of expectation, to all Bayesian–Nash implementations. However, if we are willing to modify the social choice function, then we can certainly increase revenue. Here is an example for the case of an auction with two bidders with valuations distributed uniformly in [0, 1]: Put a reservation price of 1/2, and then sell to the highest bidder for a price that is the maximum of the low bid and the reservation price, 1/2. If both bidders bid below the reservation price, then none of them wins. First, it is easy to verify that this rule is incentive compatible. Then a quick calculation will reveal that the expected revenue of this auction is 5/12 which is more than the 1/3 obtained by the regular second price or first price auctions. Chapter 13 discusses revenue maximization further.

9.7 Further Models

This chapter has concentrated on basic models. Here we shortly mention several model extensions that address issues ignored by the basic models and have received attention in economic theory.

9.7.1 Risk Aversion

All of our discussion in the Bayesian model assumed that players are risk-neutral: obtaining a utility of 2 with probability 1/2 is equivalent to obtaining a utility of 1 with probability 1. This is why we could just compute players' utilities by taking expectation. In reality, players are often risk-averse, preferring a somewhat lower utilities if they are more certain. A significant body of work in economic theory deals with formalizing and analyzing strategic behavior of such players. In our context, a particularly interesting observation is that the revenue equivalence principle fails and that with risk-averse bidders different mechanisms that implement the same social choice function may have different revenue. As an example it is known that first price auctions generate more revenue than second price auctions if the bidders are risk-averse.

9.7.2 Interdependent Values

We have considered only independent private value models: the types of the players are chosen independently of each other and each players' valuation depends only on his own private information. In a completely general setting, there would be some joint distribution over "states of the world" where such a state determines the valuations of all players. Players would not necessarily get as private information their own valuation, but rather each would get some "signal" – partial information about the state of the world – that provide some information about his own valuation and some about the valuations of others. Most of the results in this chapter cease holding for general models with interdependent values.

A case that is in the extreme opposite to the private value model is the "common value" model. In an auction of a single item under this model, we assume that the object in question has exactly the same value for all bidders. The problem is that none of them know exactly what this value is and each player's signal only provides some partial information. An example is an auction for financial instruments such as bonds. Their exact value is not completely known as it depends on future interest rates, the probability of default, etc. What is clear though is that whatever value the bonds will turn out to have, it will be the same for everyone. In such settings, an auction really serves as an information aggregation vehicle, reaching a joint estimate of the value by combining all players' signals. A common pitfall in such cases is the "winner's curse": if each bidder bids their own estimate of the object's common value, as determined from their own signal, then the winner will likely regret winning – the fact that a certain bidder won means that other signals implied a lower value, which likely means that the real value is lower than the estimate of the winner. Thus in equilibrium bidders must bid an estimate that is also conditioned on the fact that they win.

A commonly considered formalization that takes into account both a private value component and a common value component is that of affiliated signals. Roughly speaking, in such models each player gets a signal that is positively correlated (in a strong technical sense called affiliation) not only with his own value but also with the values of other players. In such settings, ascending English auctions are "better" (generate more revenue) than the non-adaptive second price auction (which is equivalent to an English auction in private value models): as the bidding progresses, each bidder gets information from the other bidders that increases his estimate of his value.

9.7.3 Complete Information Models

Our main point of view was that each player has its own private information. Some models consider a situation where all players have complete information about the game; it is only the mechanism designer who is lacking such information. A prototypical instance is that of King Solomon: two women, each claiming that the baby is hers. The women both know who the real mother is, but not King Solomon – he must design a mechanism that elicits this information from their different preferences. Several notions of implementation in such setting exists, and in general, mechanism design is much easier in this setting. In particular, many implementations without money are possible.

9.7.4 Hidden Actions

All of the theory of Mechanism Design attempts overcoming the problem that players have private information that is not known to the mechanism designer. In many settings a different stumbling block occurs: players may perform hidden actions that are not visible to the "mechanism." This complementary difficulty to the private information difficulty has been widely studied in economics and has recently started to be considered in computer science settings.

9.8 Notes

Most of the material in this chapter can be found in graduate textbooks on microeconomics such as Mas-Collel et al. (1995). The books (Krishna, 2002; Klemperer, 2004) on Auction theory contain more detail. As the Internet gained influence, during the 1990s, researchers in AI, computer networks, and economics started noticing that mechanism design can be applied in computational settings. This was put forward in a general way in Nisan and Ronen (2001) who also coined the term Algorithmic Mechanism Design.

The earliest work on voting methods including that of Condorcet and Borda goes back to the late 18th century, appropriately around the time of the French Revolution. The modern treatment of social choice theory originates with the seminal work of Arrow (1951), where Arrow's theorem also appears. Over the years many proofs for Arrow's theorem have been put forward; we bring one of those in Geanakopolos (2005). The Gibbard-Satterthwaite theorem is due to Gibbard (1973) and Satterthwaite (1975). The

computational difficulty of manipulation of voting rules was first studied in Bartholdi et al. (1989).

The positive results in Mechanism Design in the quasi-linear setting originate with the seminal work of Vickrey (1961), who, in particular, studied single-item auctions and multiunit auctions with downward sloping valuations. The public project problem was studied by Clarke (1971), who also suggested the pivot rule, and the general formulation of what is now called VCG mechanisms appears in Groves (1973). The Bilateral Trade problem was studied in Myerson and Satterthwaite (1983), and the application of buying a path in a network was put forward in Nisan and Ronen (2001).

The general framework of Mechanism Design and its basic notions have evolved in microeconomic theory mostly in the 1970s, and mostly in the general Bayesian setting that we only get to in Section 9.6. Among the influential papers in laying out the foundations are Vickrey (1961), Clarke (1971), Groves (1973), Satterthwaite (1975), Green and Laffont (1977), Dasgupta et al. (1979), and Myerson (1981).

Early papers in algorithmic Mechanism Design, such as Nisan and Ronen (2001) and Lehmann et al. (2002), pointed out the necessity and difficulty of implementing social choice functions other than welfare maximization, due to other optimization goals or due to computational hardness. Characterizations of incentive compatible mechanisms have been previously obtained in economic theory as intermediate steps on the way to theorems with clear economic motivation. The discussion here tries to put it all together independently of particular intended applications. The weak monotonicity condition is from Bikhchandani et al. (2006) and the sufficiency of this condition in convex domains is from Saks and Yu (2005). The affine-maximization characterization in complete domains is from Roberts (1979), and Lavi et al. (2003) attempts generalization to other domains. The uniqueness of pricing is the analog of the revenue equivalence theorem in the Bayesian setting which is due to Myerson (1981); Green and Laffont (1977) showed it in the dominant strategy setting for welfare maximizing social choice functions. The corollary of the impossibility of budget-balanced bilateral trade appears in Myerson and Satterthwaite (1983) in the Bayesian setting.

The Bayesian setting is currently the main vehicle of addressing lack of information in economic theory, and this development has mostly happened during the 1960s, with the main influence being the seminal work of Harsanyi (1968). As mentioned previously, most of development of the field of Mechanism Design noted above was in this setting. The revenue equivalence theorem, the form of the expected payment in single-parameter domains, as well as an analysis of revenue-maximizing auctions is from Myerson (1981).

Risk-averse bidders in (reverse) auctions are analyzed by Holt (1980). Auctions in the common value model are analyzed in Wilson (1977) and Milgrom (1981). The general model of interdependent valuations with affiliated signals was studied in Milgrom and Weber (1982). Mechanism Design in complete information models is discussed in Maskin (1985) and Moore and Repullo (1988).

Acknowledgments

I thank Shahar Dobzinski, Dana Fisman, Jason Hartline, Orna Kupferman, Ron Lavi, Ariel Procaccia, and James Schummer for comments on earlier drafts of this chapter.

Bibliography

K. Arrow. *Social Choice and Individual Values*. Yale University Press, 1951.

J. Bartholdi, III, C. Tovey, and M. Trick. Computational difficulty of manipulating an election. *Soc. Choice Welfare*, 6(3):227–241, 1989.

S. Bikhchandani, S. Chatterji, R. Lavi, A. Mu'alem, N. Nisan, and A. Sen. Weak monotonicity characterizes deterministic dominant strategy implementation. *Econometrica*, 74(4), 2006.

E.H. Clarke. Multipart pricing of public goods. *Public Choice*, 17–33, 1971.

P. Dasgupta, P. Hammond, and E. Maskin. The implementation of social choice rules: Some general results on incentive compatibility. *Rev. Econ. Stud.*, (46):185–216, 1979.

J. Geanakopolos. Three brief proofs of arrow's impossibility theorem. *Econ. Theor.*, 26(1):211–215, 2005.

A. Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, 41:587–601, 1973.

J. Green and J.J. Laffont. Characterization of satisfactory mechanism for the revelation of preferences for public goods. *Econometrica*, 427–438, 1977.

T. Groves. Incentives in teams. *Econometrica*, 617–631, 1973.

J.C. Harsanyi. Games with incomplete information played by 'bayesian' players, parts i ii and iii. *Mgmt. Sci.*, 14, 1967–68.

C. Holt. Competitive bidding for contracts under alternative auction procedures. *J. Political Econ.*, 88:433–445, 1980.

P. Klemperer. *Auctions: Theory and Practice*. Princeton University Press, 2004.

V. Krishna. *Auction Theory*. Academic Press, 2002.

R. Lavi, A. Mu'alem, and N. Nisan. Towards a characterization of truthful combinatorial auctions. In *FOCS*, 2003.

D. Lehmann, L.I. O'Callaghan, and Y. Shoham. Truth revelation in approximately efficient combinatorial auctions. *JACM* 49(5), 577–602, Sept. 2002.

A. Mas-Collel, W. Whinston, and J. Green. *Microeconomic Theory*. Oxford University Press, 1995.

E. Maskin. The theory of implementation in nash equilibrium. In *Soc. Goals and Soc. Org.: Essays in Honor of Elisha Pazner*, 1985.

P. Milgrom. Rational expectations, information acquisition, and competitive bidding. *Econometrica*, 49:921–943, 1981.

P.R. Milgrom and R.J. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122, 1982.

J. Moore and R. Repullo. Subgame perfect implementation. *Econometrica*, 56:1191–1220, 1988.

R. B. Myerson. Optimal auction design. *Math. Oper. Res.*, 6(1):58–73, 1981.

R.B. Myerson and M. Satterthwaite. Efficient mechanisms for bilateral trading. *J. Economic Theory*, (28):265–281, 1983.

N. Nisan and A. Ronen. Algorithmic mechanism design. *Games Econ. Behav.*, 35:166–196, 2001.

K. Roberts. The characterization of implementable choice rules. In *Aggregation and Revelation of Preferences*, J-J. Laffont (ed.), North Holland Publishing Company, 1979.

M. Saks and L. Yu. Weak monotonicity suffices for truthfulness. In *EC*, 2005.

M.A. Satterthwaite. Strategy-proofness and arrow's condition: Existence and correspondence theorems for voting procedures and social welfare functions. *J. Economic Theory*, 187–217, 1975.

W. Vickrey. Counterspeculation, auctions and competitive sealed tenders. *J. Finance*, 8–37, 1961.

R. Wilson. A bidding model of perfect competition. *Rev. Econ. Stud.*, 44:511–518, 1977.