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 $F : L^n \rightarrow L$ is called a *social welfare function*.
- A function $f : L^n \rightarrow A$ 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* $v_i : A \rightarrow \Re$, 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 $v_i : A \rightarrow \Re$, 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 $f : V_1 \times \cdots \times V_n \rightarrow A$ and a vector of payment functions $p_1, \ldots, p_n$, where $p_i : V_1 \times \cdots \times V_n \rightarrow \Re$ 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

- $f(v_1, \ldots, v_n) \in \text{argmax}_{a \in A} \sum_i v_i(a)$; that is, f maximizes the social welfare, and
- for some functions $h_1, \ldots, h_n$, where $h_i : V_{-i} \rightarrow \Re$ (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 $A = \{\text{no–trade, trade}\}$ 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