Chapter 4

Mechanism Design

Honesty is the best policy - when there is money in it.

Mark Twain

In order for a preference aggregator to choose a good outcome, she needs to be provided with the agents' (relevant) preferences. Usually, the only way of learning these preferences is by having the agents report them. Unfortunately, in settings where the agents are self-interested, they will report these preferences truthfully if and only if it is in their best interest to do so. Thus, the preference aggregator has the difficult task of not only choosing good outcomes for the given preferences, but also choosing outcomes in such a way that agents will not have any incentive to misreport their preferences. This is the topic of *mechanism design*, and the resulting outcome selection functions are called *mechanisms*.

This chapter gives an introduction to some basic concepts and results in mechanism design. In Section 4.1, we review basic concepts in mechanism design (although discussions of the game-theoretic justifications for this particular framework, in particular the *revelation principle*, will be postponed to Chapter 7). In Section 4.2, we review the famous and widely-studied *Vickrey-Clarke-Groves* mechanisms and their properties. In Section 4.3, we briefly review some other positive results (mechanisms that achieve particular properties), while in Section 4.4, we briefly review some key impossibility results (combinations of properties that no mechanism can achieve).

4.1 Basic concepts

If all of the agents' preferences were public knowledge, there would be no need for mechanism design—all that would need to be done is solve the outcome optimization problem. Techniques from mechanism design are useful and necessary only in settings in which agents' have *private information* about their preferences. Formally, we say that each agent i has a privately known *type* θ_i that corresponds to that agent's private information, and we denote by Θ_i the space of all of agent i's possible types. In general, it is possible to have private information that has implications for how other agents value outcomes—for example, one agent may privately know that the painting that is being auctioned is a forgery, which would be relevant to other agents that may not know this [Ito *et al.*, 2002, 2003, 2004]. In this dissertation, as is most commonly done in the mechanism design

literature, we will only consider private information about the agent's own preferences (which is the most common type of private information). We model these preferences by saying that each agent i has a utility function $u_i : \Theta_i \times O \to \mathbb{R}$, where $u_i(\theta_i, o)$ gives the agent's utility for outcome o when the agent has type θ_i. The utility function u_i is common knowledge, but it is still impossible for other agents to precisely assess agent i's utility for a given outcome o without knowing agent i's type. For example, in an auction for a single item, an agent's type θ_i could be simply that agent's valuation for the item. Then, the agent's utility for an outcome in which he receives the item will be θ_i (not counting any payments to be made by the agent), and the utility is 0 otherwise. Hence, the utility function is common knowledge, but one still needs to know the agent's type to assess the agent's utility for (some) outcomes.

A *direct-revelation mechanism* asks each agent to report its private information, and chooses an outcome based on this (and potentially some random bits). It will generally be convenient not to consider payments imposed by the mechanism as part of the outcome, so that the mechanism also needs to specify payments to be made by/to agents. Formally:

Definition 16

- *A* deterministic direct-revelation mechanism without payments *consists of an outcome selection function* $o : \Theta_1 \times \ldots \times \Theta_n \to O$.
- *A* randomized direct-revelation mechanism without payments *consists of a distribution selection function* $p : \Theta_1 \times \ldots \times \Theta_n \to \Delta(O)$, *where* $\Delta(O)$ *is the set of probability distributions over* O.
- *A* deterministic direct-revelation mechanism with payments *consists of an outcome selection function* $o : \Theta_1 \times \ldots \times \Theta_n \to O$ *and for each agent* i, *a payment selection function* $\pi_i : \Theta_1 \times \ldots \times \Theta_n \to \mathbb{R}$, *where* $\pi_i(\theta_1, \ldots, \theta_n)$ *gives the payment made by agent* i *when the reported types are* $\theta_1, \ldots, \theta_n$.
- *A* randomized direct-revelation mechanism with payments *consists of a distribution selection function* $p : \Theta_1 \times \ldots \times \Theta_n \to \Delta(O)$, *and for each agent* i, *a payment selection function* $\pi_i : \Theta_1 \times \ldots \times \Theta_n \to \mathbb{R}$.

In some settings, it makes sense to think of an agent's type θ_i as being drawn from a (commonly known) prior distribution over Θ_i. In this case, while each agent still only knows its own type, each agent can use the commonly known prior to make probabilistic assessments of what the others will report.

So, what makes for a good mechanism? Typically, there is an *objective function* that the designer wants to maximize. One common objective is social welfare (the sum of the agents' utilities with respect to their *true*, not reported, types), but there are many others—for example, the designer may wish to maximize revenue (the sum of the agents' payments). However, there are certain constraints on what the designer can do. For example, it would not be reasonable for the designer to specify that a losing bidder in an auction should pay the designer a large sum: if so, the bidder would simply not participate in the auction. We next present constraints, called *participation* or *individual rationality (IR)* constraints, that prevent this. Before we do so, we note that we will assume *quasilinear preferences* when payments are involved.

Definition 17 *An agent i has* quasilinear preferences *if the agent's utility function can be written as $u_i(\theta_i, o) - \pi_i$.*

We are now ready to present the IR constraints.

Definition 18 Individual rationality (IR) *is defined as follows.*

- *A deterministic mechanism is* ex post IR *if for any agent i, and any type vector $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, we have $u_i(\theta_i, o(\theta_1, \ldots, \theta_n)) - \pi_i(\theta_1, \ldots, \theta_n) \geq 0$.*

  A randomized mechanism is ex post IR *if for any agent i, and any type vector $(\theta_1, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_n$, we have $E_{o|\theta_1,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

- *A deterministic mechanism is* ex interim IR *if for any agent i, and any type $\theta_i \in \Theta_i$, we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1, .., \theta_n)) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

  A randomized mechanism is ex interim IR *if for any agent i, and any type $\theta_i \in \Theta_i$, we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, .., \theta_n)] \geq 0$.*

The terms involving payments are left out if payments are not possible.

Thus, participating in an *ex post* individually rational mechanism never makes an agent worse off; participating in an *ex interim* individually rational mechanism may make an agent worse off in the end, but not in expectation (assuming that the agent's belief over the other agents' reported types matches the common prior).

Still, as long as these are the only constraints, all that the designer needs to do is solve the outcome optimization problem (perhaps charging the agents' their entire utility as payment, in case revenue maximization is the objective). But we have not yet considered the agents' *incentives*. Agents will only report their preferences truthfully if they have an incentive to do so. We will impose *incentive compatibility (IC)* constraints that ensure that this is indeed the case. Again, there is an *ex post* and an *ex interim* variant; in this context, these variants are usually called *dominant-strategies incentive compatible* and *Bayes-Nash equilibrium (BNE) incentive compatible*, respectively. Given the (potential) difference between true and reported types, we will use the standard notation $\hat{\theta}_i$ to refer to agent i's reported type.

Definition 19 *A mechanism is* dominant-strategies incentive compatible *(or* strategy-proof*) if telling the truth is always optimal, even when the types reported by the other agents are already known. Formally, for any agent i, any type vector $(\theta_1, \ldots, \theta_i, \ldots, \theta_n) \in \Theta_1 \times \ldots \times \Theta_i \times \ldots \times \Theta_n$, and any alternative type report $\hat{\theta}_i \in \Theta_i$, in the case of deterministic mechanisms we require $u_i(\theta_i, o(\theta_1, \ldots, \theta_i, \ldots, \theta_n)) - \pi_i(\theta_1, \ldots, \theta_i, \ldots, \theta_n) \geq u_i(\theta_i, o(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)) -$ $\pi_i(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)$. In the case of randomized mechanisms we have $E_{o|\theta_1,..,\theta_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, \ldots, \theta_i, \ldots, \theta_n)] \geq E_{o|\theta_1,..,\hat{\theta}_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1, \ldots, \hat{\theta}_i, \ldots, \theta_n)]$.*
The terms involving payments are left out if payments are not possible.

Definition 20 *A mechanism is* Bayes-Nash equilibrium (BNE) incentive compatible *if telling the truth is always optimal to an agent when that agent does not yet know anything about the other*

agents' types, and the other agents are telling the truth. Formally, for any agent i, any type $\theta_i \in \Theta_i$, and any alternative type report $\hat{\theta}_i \in \Theta_i$, in the case of deterministic mechanisms we have $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1,\ldots,\theta_i,\ldots,\theta_n)) - \pi_i(\theta_1,\ldots,\theta_i,\ldots,\theta_n)] \geq$ $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i}[u_i(\theta_i, o(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)) - \pi_i(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)]$*. In the case of randomized mechanisms we have* $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\theta_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1,\ldots,\theta_i,\ldots,\theta_n)] \geq$ $E_{(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_n)|\theta_i} E_{o|\theta_1,..,\hat{\theta}_i,..,\theta_n}[u_i(\theta_i, o) - \pi_i(\theta_1,\ldots,\hat{\theta}_i,\ldots,\theta_n)]$.

The terms involving payments are left out if payments are not possible.

One may wonder whether it is possible to obtain better outcomes by using a direct-revelation mechanism that is not truthful—perhaps the cost of the resulting strategic misreporting is not as great as the cost of having to honor the incentive compatibility constraints. Or, perhaps we could do even better using a mechanism that is not a direct-revelation mechanism—that is, a mechanism under which agents have other actions to take besides merely reporting their preferences. A famous result called the *revelation principle* [Gibbard, 1973; Green and Laffont, 1977; Myerson, 1979, 1981] shows that, when agents are perfectly strategic (unboundedly rational), the answer to both of these questions is "no": there is no loss in restricting attention to truthful, direct-revelation mechanisms.[1] For now, we do not yet have a definition of strategic behavior in non-truthful or indirect mechanisms, so we will postpone detailed discussion of the revelation principle to Chapter 7 (and we will question the assumption of unbounded rationality in Chapters 8 and 9). However, the intuition behind the revelation principle is simple: suppose we envelop a non-truthful mechanism with an *interface layer*, to which agents input their preferences. Then, the interface layer interacts with the original mechanism on behalf of each agent, *playing strategically in the agent's best interest* based on the reported preferences. (Compare, for example, proxy agents on eBay [eBay UK, 2004].) The resulting mechanism is truthful: an agent has no incentive to misreport to the interface layer, because the layer will play the agent's part in the original mechanism in the agent's best interest. Moreover, the final outcome of the new, truthful mechanism will be the same, because the layer will play strategically optimally—just as the agent would have.

In the next section, we will define the famous Vickrey-Clarke-Groves mechanisms.

4.2 Vickrey-Clarke-Groves mechanisms

The most straightforward direct-revelation mechanism for selling a single item is the *first-price sealed-bid* auction, in which each bidder submits a bid for the item in (say) a sealed envelope, and the highest bidder wins and pays the value that he bid. This is certainly not an incentive-compatible mechanism: in fact, bidding one's true valuation guarantees a utility of 0 (even if the bid wins, the bidder will pay his entire valuation). Rather, to obtain positive utility, a bidder needs to reduce (or *shave*) his bid, ideally to the point where it is only slightly higher than the next highest bid. Another direct-revelation mechanism is the *Vickrey* [Vickrey, 1961] or *second-price sealed-bid* auction, in which the highest bidder still wins, but pays the value of the *second* highest bid. The Vickrey auction is strategy-proof. To see why, imagine a bidder that knows the other bids. This bidder has only two

[1] The result requires that we can use randomized truthful mechanisms. Moreover, if there are multiple strategic equilibria in the original non-truthful mechanism, then we can choose any one of them to be preserved in the truthful mechanism, but not *all* the equilibria are necessarily preserved.

choices: bid higher than the highest other bid, to win and pay the value of that other bid; or bid lower, and do not win the item. The bidder will prefer to do the former if his valuation is higher than the highest other bid, and the latter otherwise. But in fact, bidding truthfully accomplishes exactly this! Hence, bidding truthfully guarantees one the same utility that an omniscient bidder would receive, and therefore the mechanism is strategy-proof.

It turns out that the Vickrey mechanism is a special case of a general mechanism called the *Clarke* mechanism (or Clarke tax) [Clarke, 1971], which can be applied to combinatorial auctions and exchanges, as well as other preference aggregation settings. The Clarke mechanism works as follows. First, choose the optimal outcome based on the bidders' reported preferences; call this outcome o^*. Then, to determine agent i's payment, remove agent i from the preference aggregation problem, and solve this problem again to obtain o^*_{-i}. Agent i will be required to pay $\sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i}) - \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$. Informally, agent i's payment is exactly the amount by which the other agents are worse off due to agent i's presence—the *externality* that i imposes on the other agents. The Clarke mechanism is strategy-proof, for the following reason. Agent i seeks to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*) - \sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i})$. Since o^*_{-i} does not depend on agent i's report, agent i cannot affect the term $\sum_{j \neq i} u_j(\hat{\theta_j}, o^*_{-i})$, so equivalently, agent i seeks to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$. Agent i can only affect this expression by influencing the choice of o^*, and the mechanism will select o^* to maximize $\sum_{j=1}^{n} u_j(\hat{\theta_j}, o^*)$. But then, if the agent reports truthfully, that is, $\hat{\theta_i} = \theta_i$, then the mechanism will choose o^* precisely to maximize $u_i(\theta_i, o^*) + \sum_{j \neq i} u_j(\hat{\theta_j}, o^*)$, thereby maximizing agent i's utility.

The Clarke mechanism is also *ex post* individually rational, *if* 1) the presence of an agent never makes it impossible to choose some outcome that could have been chosen without that agent, and 2) no agent ever has a negative utility for an outcome that would be selected if that agent were not present. Note that if either 1) or 2) does not hold, then the Clarke mechanism may require a payment from an agent that receives a utility of 0 for the chosen outcome, and is therefore not individually rational. Both 1) and 2) will hold in the remainder of this dissertation.

Additionally, the Clarke mechanism is *weak budget balanced*, that is, the sum of the payments from the agents is always nonnegative, *if* the following condition holds: when an agent is removed from the system, the new optimal (welfare-maximizing) outcome is at least as good for the remaining agents as the optimal outcome before the first agent was removed. That is, if an agent leaves, that does not make the other agents worse off in terms of the chosen outcome (not considering payments). This condition does not hold in, for example, task allocation settings: if an agent leaves, the tasks allocated to that agent must be re-allocated to the other agents, who will therefore be worse off. Indeed, in task allocation settings, the agents must be compensated for taking on tasks, so we do not expect weak budget balance. Green and Laffont [1977] show that it is not possible to obtain *strong* budget balance—the sum of the payoffs always being zero—in addition to choosing optimal outcomes and having dominant-strategies incentive compatibility.

Finally, the Clarke mechanism is just one mechanism among the class of *Groves* mechanisms [Groves, 1973]. To introduce this class of mechanisms, we note that in the Clarke mechanism, agent

i's type report $\hat{\theta}_i$ does not affect the terms $\sum_{j \neq i} u_j(\hat{\theta}_j, o^*_{-i})$ in agent i's payment; short of colluding with the other agents, there is nothing that agent i can do about paying these terms. Hence, if we removed these terms from the payment function, the mechanism would still be strategy-proof. Moreover, *any* term that we add to agent i's payment *that does not depend on* $\hat{\theta}_i$ will not compromise strategy-proofness. The class of Groves mechanisms consists precisely of all mechanisms that can be obtained in this manner. Additionally, Groves mechanisms are in fact the *only* mechanisms that are efficient (*i.e.* the mechanism chooses the optimal outcome) and dominant-strategies incentive-compatible, given that there is no restriction on what the agents' types can be [Green and Laffont, 1977] or even only given that agents' type spaces are smoothly connected [Holmström, 1979]. It should be noted that the Clarke mechanism is often referred to as "the" VCG mechanism, and we will follow this convention.

In the next section, I survey some other positive results in mechanism design (without presenting them in full detail).

4.3 Other possibility results

Interestingly, in some settings, there are Groves mechanisms that require a smaller total payment from the agents than the Clarke mechanism, while maintaining individual rationality and never incurring a deficit. The idea here is to *redistribute* some of the Clarke surplus back to the agents. To maintain incentive compatibility, how much is redistributed to an agent cannot depend on that agent's type. Nevertheless, if the other agents' reported types are such that a certain amount of Clarke surplus will be obtained *regardless* of the given agent's report, then we can redistribute a share of that guaranteed surplus (most naturally, $1/n$) to the agent. For example, in a single-item auction, each agent receives $1/n$ of the second-highest bid among the other bids [Cavallo, 2006].

It turns out that if we are willing to use Bayes-Nash incentive compatibility rather than dominant-strategies incentive compatibility, then we can obtain (strong) budget balance, using the dAGVA [d'Aspremont and Gérard-Varet, 1979; Arrow, 1979] mechanism. This mechanism is similar to a Groves mechanism, except that, instead of being paid the sum of other agents' utilities according to their reported types, an agent is paid the *expected* sum of other agent's utilities given only the agent's own report. In addition, payment terms that do not depend on the agent's own report can be set in such a way as to obtain budget balance.

As noted before, maximizing social welfare is not always the objective. Another common objective is to maximize revenue. In the context of auctions, this is often referred to as the problem of designing an "optimal" auction. The Myerson auction [Myerson, 1981] is a general mechanism for maximizing the expected revenue of an auctioneer selling a single item. The Maskin-Riley auction [Maskin and Riley, 1989] generalizes this to the case of multiple units of the same item. Only very limited characterizations of revenue-maximizing combinatorial auctions (with more than one item) are known [Avery and Hendershott, 2000; Armstrong, 2000].

Another positive result exists in the context of voting: if preferences are single-peaked, then choosing the median voter's peak as the winner (as we did in Chapter 2) is a strategy-proof mechanism.

4.4 Impossibility results

In the previous sections, we saw mechanisms that achieve certain sets of desirable properties. In this section, we discuss a few negative results, that state that certain sets of desirable properties cannot be obtained by a single mechanism.

Possibly the best-known impossibility result in mechanism design is the Gibbard-Satterthwaite theorem [Gibbard, 1973; Satterthwaite, 1975]. This result shows a very strong impossibility in very general preference aggregation settings (voting settings). Specifically, it shows that when there are three or more possible outcomes (candidates), two or more agents (voters), and there is no restriction on the preferences that can be submitted (such as single-peakedness), then a (deterministic) mechanism (voting rule) cannot have the following properties simultaneously:

- For every outcome, there exist preference reports by the agents that will make this outcome win.
- The mechanism is non-dictatorial, that is, the rule does not simply always choose a single, fixed voter's most-preferred candidate.
- The mechanism is strategy-proof.

Gibbard [1977] later extended this impossibility result to encompass randomized voting rules as well: a randomized voting rule is strategy-proof only if it is a probability mixture of *unilateral* and *duple* rules. (A rule is unilateral if only one voter affects the outcome, and duple if only two candidates can win.) It is not difficult to see that this result implies the Gibbard-Satterthwaite impossibility result.

As we have seen in the previous section, this impossibility result does not apply in settings where the agents' preferences are more restricted—*e.g.* single-peaked, or quasilinear in settings where payments are possible (in which case VCG can be used). Nevertheless, impossibility results exist in these more restricted settings as well. For example, the Myerson-Satterthwaite impossibility theorem [Myerson and Satterthwaite, 1983] states that even in simple bilateral trade settings with quasilinear utility functions, where we have a single seller with a single item (and a privately held valuation for this item), and a single buyer who may procure the item (and has a privately held valuation for the item), it is impossible to have a mechanism that achieves the following properties simultaneously:

- efficiency (trade takes place if and only if the buyer's valuation for the item is greater than the seller's);
- budget-balance (money may flow between the buyer and the seller, but not from/to other places);
- Bayes-Nash incentive compatibility;
- ex-interim individual rationality.

We will show another, similar impossibility result in Chapter 5.

4.5 Summary

This chapter reviewed basic concepts and results from mechanism design. We first reviewed various types of mechanisms, as well as individual-rationality and incentive-compatibility concepts. We then reviewed Vickrey-Clarke-Groves mechanisms and their properties in detail, and we briefly reviewed some other positive results (that is, mechanisms that achieve certain sets of properties), including Cavallo's redistribution mechanism, the dAGVA mechanism, Myerson and Maskin-Riley auctions, and single-peaked preferences. Finally, we briefly reviewed the Gibbard-Satterthwaite and Myerson-Satterthwaite impossibility results.

Armed with a basic understanding of mechanism design, we are now ready to move on to deeper levels of the hierarchy. The next chapter studies difficulties for classical mechanism design in expressive preference aggregation settings.