CHAPTER 10

Mechanism Design without Money

James Schummer and Rakesh V. Vohra

Abstract

Despite impossibility results on general domains, there are some classes of situations in which there exist interesting dominant-strategy mechanisms. While some of these situations (and the resulting mechanisms) involve the transfer of money, we examine some that do not. Specifically, we analyze problems where agents have single-peaked preferences over a one-dimensional "public" policy space; and problems where agents must match with each other.

10.1 Introduction

The Gibbard–Satterthwaite Theorem (Theorem 9.8) is a Procrustean bed[1] that is escaped only by relaxing its assumptions. In conjunction with the Revelation Principle (Proposition 9.25), it states that on the *general domain* of preferences, only dictatorial rules can be implemented in dominant strategies (if the range contains at least three alternatives). In this chapter we escape Procrustes by examining dominant strategy implementation on *restricted* domains of preferences.[2]

In most applications it is clearly unreasonable to assume that agents' preferences are completely unrestricted, as was assumed in the voting context of Section 9.2.4. For instance, in situations involving the allocation of goods, including money, one can safely assume that each agent prefers to receive more money (or other goods). As can be seen in the following chapters, the ability for agents to make monetary transfers allows for a rich class of *strategy-proof* rules.

Nevertheless there are many important environments where money cannot be used as a medium of compensation. This constraint can arise from ethical and/or institutional

[1] Procrustes was a giant that lived by one of the roads that led to Attica. He boasted of a bed whose length exactly matched the size of its occupant. What he neglected to mention was that this remarkable feature was obtained by either stretching or butchering his guest to fit the bed.

[2] Other avenues of escape not discussed here include randomization, making preferences common knowledge, and using weaker notions of implementation.

considerations: many political decisions must be made without monetary transfers; organ donations can be arranged by "trade" involving multiple needy patients and their relatives, yet monetary compensation is illegal. In this chapter we focus on a few examples of just this kind.

Before proceeding with the examples, we formalize the idea that dominant-strategy implementation is a weaker concept on restricted domains of preferences. In general, a decision problem can be described by these parameters: a set of agents $N = \{1, 2, \ldots, n\}$, a set of alternatives A, and for each agent $i \in N$ a set of potential preference relations $\mathcal{R}_i$ over the alternatives in A.[3] The Gibbard–Satterthwaite Theorem (Theorem 9.8) applies, for example, when each $\mathcal{R}_i$ is the entire set of linear orders on A.

An allocation rule is a function $f : \times \mathcal{R}_i \rightarrow A$, mapping preferences of the agents into alternatives. It is *strategy-proof* if its use makes it a weakly dominant strategy for agents to truthfully report their preferences. (See Section 9.4). We observe the following principle.

Consider two decision problems $(N, A, \mathcal{R}_1, \ldots, \mathcal{R}_n)$ and $(N, A, \mathcal{R}'_1, \ldots, \mathcal{R}'_n)$, where $\mathcal{R}'_i \subseteq \mathcal{R}_i$ for each $i \in N$. Suppose $f : \times \mathcal{R}_i \rightarrow A$ is a *strategy-proof* rule for the former problem. Then the restriction of the function f to $(\times \mathcal{R}'_i)$, namely $f|_{\times \mathcal{R}'_i}$, defines a *strategy-proof* rule for the latter problem.

The proof of this is straightforward: on a smaller domain of preferences, *strategy-proofness* is easier to satisfy because it imposes strictly fewer constraints. This simple observation justifies the search for reasonable (or at least nondictatorial) rules for decision problems involving "smaller" domains of preferences than those that yield the Gibbard–Satterthwaite Theorem.

In Section 10.2 we analyze a problem involving a natural domain restriction when agents vote over one-dimensional policies. It is one of the canonical "public good" settings ($\mathcal{R}_i = \mathcal{R}_j$ for all $i, j \in N$) in which interesting, *strategy-proof* rules can be obtained. The analysis here is illustrative of the approach used to characterize such rules in other environments. In Sections 10.3 and 10.4 we analyze matching problems. As opposed to the previous setting, these problems have the feature that each agent cares only about his own private consumption; that is, each $\mathcal{R}_i$ contains only preference relations that are sensitive only to certain dimensions of the alternative space A; hence $\mathcal{R}_i \neq \mathcal{R}_j$ whenever $i \neq j$. These are examples of what are called "private good" problems. Two kinds of matching problems are analyzed, demonstrating the limits of what can be implemented in dominant strategies in such environments.

10.2 Single-Peaked Preferences over Policies

A simple but elegant class of domains involves *single-peaked preferences* over one-dimensional policy spaces. This domain can be used to model political policies, economic decisions, location problems, or any allocation problem where a single point

[3] A preference relation is a weak order on A.

must be chosen in an interval. The key assumption we make is that agents' preferences are assumed to have a single most-preferred point in the interval, and that preferences are "decreasing" as one moves away from that peak.

Formally, the allocation space (or policy space) is the unit interval $A = [0, 1]$. An *outcome* in this model is a single point $x \in A$. Each agent $i \in N$ has a preference ordering $\succeq_i$ (i.e., a weak order) over the outcomes in $[0, 1]$. The preference relation $\succeq_i$ is *single-peaked* if there exists a point $p_i \in A$ (the *peak* of $\succeq_i$) such that for all $x \in A \setminus \{p_i\}$ and all $\lambda \in [0, 1)$, $(\lambda x + (1 - \lambda)p_i) \succ_i x$.[4] Let $\mathcal{R}$ denote the class of single-peaked preferences.

We denote the peaks of preference relations $\succeq_i$, $\succeq_i'$, $\succeq_j$, etc., respectively by p_i, p_i', p_j, etc. Denote a *profile* (n-tuple) of preferences as $\succeq \in \mathcal{R}^n$.

One can imagine this model as representing a political decision such as an income tax rate, another political issue with conservative/liberal extremes, the location of a public facility on a road, or even something as simple as a group of people deciding on the temperature setting for a shared office. In these and many other examples, the agents have an ideal preferred policy in mind, and would prefer that a decision be made as close as possible to this "peak."

A *rule* f: $\mathcal{R}^n \to A$ assigns an outcome $f(\succeq)$ to any preference profile $\succeq$. As before, a rule is *strategy-proof* if it is a dominant strategy for each agent to report his preferences truthfully when the rule is being used to choose a point.

In contrast to the impossibility result of Gibbard (1973) and Satterthwaite (1975), that obtain on the universal domain of preferences, we shall see that this class of problems admits a rich family of *strategy-proof* rules whose ranges include more than two alternatives. In fact, the family of such rules remains rich even when one restricts attention (as we do in this chapter) to rules that satisfy the following condition.

We say that a rule f is *onto* if for all $x \in A$ there exists $\succeq \in \mathcal{R}^n$ such that $f(\succeq) = x$. An *onto* rule cannot preclude an outcome from being chosen *ex ante*. It is not without loss of generality to impose this condition. For instance, fix two points $x, y \in [0, 1]$ and consider a rule that chooses whichever of the two points is preferred to the other by a majority of agents (and where x is chosen in case of a tie). Such a rule is *strategy-proof*, but not *onto*. Similar *strategy-proof* rules can even break ties between x and y by using preference information about other points x', y', ..., in $[0, 1]$, even though x', etc., are not in the range of the rule.

The *onto* condition is even weaker than what is called *unanimity*, which requires that whenever all agents' preferences have the same peak ($p_i = p_j$ for all i, j), the rule must choose that location as the outcome. In turn, *unanimity* is weaker than *Pareto-optimality*: for all $\succeq \in \mathcal{R}^n$, there exists no point $x \in [0, 1]$ such that $x \succ_i f(\succeq)$ for all $i \in N$.

As it turns out, these three requirements are all equivalent among *strategy-proof* rules.

Lemma 10.1 *Suppose f is* strategy-proof. *Then f is* onto *if and only if it is unanimous if and only if it is Pareto-optimal.*

[4] The binary relation $\succ_i$ is the strict (asymmetric) part of $\succeq_i$. Under a single-peaked preference relation, preference is strictly decreasing as one moves away from p_i.

PROOF It is clear that Pareto-optimality implies the other two conditions. Suppose f is *strategy-proof* and *onto*. Fix $x \in [0, 1]$ and let $\succeq \in \mathcal{R}^n$ be such that $f(\succeq) = x$. Consider any "unanimous" profile $\succeq' \in \mathcal{R}^n$ such that $p'_i = x$ for each $i \in N$. By *strategy-proofness*, $f(\succeq'_1, \succeq_2, \ldots, \succeq_n) = x$, otherwise agent 1 could manipulate f. Repeating this argument, $f(\succeq'_1, \succeq'_2, \succeq_3, \ldots, \succeq_n) = x, \ldots, f(\succeq') = x$. That is, f is unanimous.

Finally, to derive a contradiction, suppose that f is not Pareto-optimal at some profile $\succeq \in \mathcal{R}^n$. This implies that either (i) $f(\succeq) < p_i$ for all $i \in N$ or (ii) $f(\succeq) > p_i$ for all $i \in N$. Without loss of generality, assume (i) holds. Furthermore, assume that the agents are labeled so that $p_1 \leq p_2 \leq \cdots \leq p_n$.

If $p_1 = p_n$ then unanimity is violated, completing the proof. Otherwise, let $j \in N$ be such that $p_1 = p_j < p_{j+1}$; that is, $j < n$ agents have the minimum peak. For all $i > j$, let $\succeq'_i$ be a preference relation such that both $p'_i = p_1$ and $f(\succeq) \succeq'_i p_i$.

Let $x_n = f(\succeq_1, \ldots, \succeq_{n-1}, \succeq'_n)$. By *strategy-proofness*, $x_n \in [f(\succeq), p_n]$, otherwise agent n (with preference $\succeq'_n$) could manipulate f by reporting preference $\succeq_n$. Similarly, $x_n \notin (f(\succeq), p_n]$, otherwise agent n (with preference $\succeq_n$) could manipulate f by reporting preference $\succeq'_n$. Therefore $x_n = f(\succeq)$.

Repeating this argument as each $i > j$ replaces $\succeq_i$ with $\succeq'_i$, we have

$$f(\succeq_1, \ldots, \succeq_j, \succeq'_{j+1}, \ldots, \succeq'_n) = f(\succeq)$$

which contradicts unanimity. Since a *strategy-proof*, *onto* rule must be unanimous, this is a contradiction. □

10.2.1 Rules

The central *strategy-proof* rule on this domain is the simple median-voter rule. Suppose that the number of agents n is odd. Then the rule that picks the median of the agents' peaks (p_i's) is a *strategy-proof* rule.

It is straightforward to see why this rule is *strategy-proof*: If an agent's peak p_i lies *below* the median peak, then he can change the median only by reporting a preference relation whose peak lies *above* the true median. The effect of this misreport is for the rule to choose a point even further away from p_i, making the agent worse off. A symmetric argument handles the case in which the peak is above the median. Finally, an agent cannot profitably misreport his preferences if his peak is the median one to begin with.

More generally, for any number of agents n and any positive integer $k \leq n$, the rule that picks the kth highest peak is *strategy-proof* for precisely the same reasons as above. An agent can only move the kth peak further from his own. The median happens to be the case where $k = (n + 1)/2$.

The *strategy-proofness* of such rules stands in contrast to the incentives properties of rules that choose *average*-type statistics. Consider the rule that chooses the average of the n agents' peaks. Any agent with peak $p_i \in (0, 1)$ that is not equal to the average can manipulate the rule by reporting preferences with a more extreme peak (closer to 0 or 1) than his true peak.

This would also hold for any *weighted* average of the agents' peaks, with one exception. If a rule allocated all of the weight to one agent, then the resulting rule simply picks that agent's peak always. Such a *dictatorial* rule is *strategy-proof* and *onto*.

In addition to favorable incentives properties, rules based on order statistics have the property that they require little information to be computed. Technically a rule requires agents to report an entire preference ordering over [0, 1]. The rules we have discussed so far, however, only require agents to report their most preferred point, i.e., a single number. In fact, under the *onto* assumption, this informational property is a consequence of the *strategy-proofness* requirement; that is, *all strategy-proof* and *onto* rules have the property that they can be computed solely from information about the agents' peaks.

To begin showing this, we first observe that the class of "kth-statistic rules" can be further generalized as follows. Consider a fixed set of points $y_1, y_2, \ldots, y_{n-1} \in A$. Consider the rule that, for any profile of preferences $\succeq$, chooses the median of the $2n-1$ points consisting of the n agents' peaks and the $n-1$ points of y. This kind of rule differs from the previous ones in that, for some choices of y and some profiles of preferences, the rule may choose a point that is not the peak of *any* agent's preferences. Yet, for the same reasons as above, such a rule is *strategy-proof.*

It turns out that such rules compose the entire class of *strategy-proof* and *onto* rules that treat agents symmetrically. To formalize this latter requirement, we call a rule *anonymous* if for any $\succeq \in \mathcal{R}^n$ and any permutation $\succeq'$ of $\succeq$, $f(\succeq') = f(\succeq)$. This requirement captures the idea that the agents' names play no role in the behavior of a rule. Dictatorial rules mentioned above are examples of rules that are *strategy-proof* and *onto*, but not *anonymous*.

Theorem 10.2 *A rule f is* strategy-proof*,* onto*, and* anonymous *if and only if there exist $y_1, y_2, \ldots, y_{n-1} \in [0, 1]$ such that for all $\succeq \in \mathcal{R}^n$,*

$$f(\succeq) = \text{med}\{p_1, p_2, \ldots, p_n, y_1, y_2, \ldots, y_{n-1}\}. \tag{10.1}$$

PROOF We leave it as an exercise to verify that such a rule satisfies the three axioms in the Theorem. To prove the converse, suppose f is *strategy-proof*, *onto*, and *anonymous*.

We make extensive use of the two (extreme) preference relations that have peaks at 0 and 1 respectively. Since preferences relations are ordinal, there is only one preference relation with a peak at 0 and only one with a peak at 1. Denote these two preference relations by $\succeq_i^0$ and $\succeq_i^1$ respectively.

(Construct the y_m's.) For any $1 \leq m \leq n-1$, let y_m denote the outcome of f when m agents have preference relation $\succeq_i^1$ and the remainder have $\succeq_i^0$:

$$y_m = f\big(\succeq_1^0, \ldots, \succeq_{n-m}^0, \succeq_{n-m+1}^1, \ldots, \succeq_n^1\big).$$

Recall that by *anonymity* the order of the arguments of f is irrelevant; if precisely m agents have preference relation $\succeq_i^1$ and the rest have $\succeq_i^0$ then the outcome is y_m. In addition, we leave it to the reader to verify that *stragegy proofness* implies monotonicity of the y_m's: $y_m \leq y_{m+1}$ for each $1 \leq m \leq n-2$. We prove the theorem by showing that f satisfies Eq. (10.1) with respect to this list of y_m's.

Consider a profile of preferences $\succeq \in \mathcal{R}^n$ with peaks $p_1, \ldots, p_n$. Without loss of generality (by *anonymity*) assume that $p_i \leq p_{i+1}$ for each $i \leq n-1$. We wish to show $f(\succeq) = x^* \equiv \operatorname{med}\{p_1, \ldots, p_n, y_1, \ldots, y_{n-1}\}$.

(Case 1: the median is some y_m.) Suppose $x^* = y_m$ for some m. By monotonicity of the peaks and y_m's, since x^* is the median of $2n-1$ points this implies $p_{n-m} \leq x^* = y_m \leq p_{n-m+1}$. By assumption,

$$x^* = y_m = f\left(\succeq_1^0, \ldots, \succeq_{n-m}^0, \succeq_{n-m+1}^1, \ldots, \succeq_n^1 \right). \tag{10.2}$$

Let $x_1 = f(\succeq_1, \succeq_2^0, \ldots, \succeq_{n-m}^0, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)$. *Strategy-proofness* implies $x_1 \geq x^*$, otherwise agent 1 with preference $\succeq_1^0$ could manipulate f. Similarly, since $p_1 \leq y_m$, we cannot have $x_1 > x^*$, otherwise agent 1 with preference $\succeq_1$ could manipulate f. Hence $x_1 = x^*$. Repeating this argument for all $i \leq n-m$, $x^* = f(\succeq_1, \ldots, \succeq_{n-m}, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)$. The symmetric argument for all $i > n-m$ implies

$$f(\succeq_1, \ldots, \succeq_n) = x^*. \tag{10.3}$$

(Case 2: the median is an agent's peak.) The remaining case is that $y_m < x^* < y_{m+1}$ for some m. (The cases where $x^* < y_1$ and $x^* > y_{n-1}$ are similar, denoting $y_0 = 0$ and $y_n = 1$.) In this case, since the agents' peaks are in increasing order, we have $x^* = p_{n-m}$.

If

$$f\left(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \succeq_{n-m}, \succeq_{n-m+1}^1, \ldots, \succeq_n^1 \right) = x^* = p_{n-m} \tag{10.4}$$

then, analogous to the way Eq. (10.2) implied Eq. (10.3), repeated applications of *strategy-proofness* (to the $n-1$ agents other than $i = n-m$) would imply $f(\succeq_1, \ldots, \succeq_n) = x^*$, and the proof would be finished. The remainder of the proof is devoted to showing that indeed Eq. (10.4) must hold.

Suppose to the contrary that

$$f\left(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \succeq_{n-m}, \succeq_{n-m+1}^1, \ldots, \succeq_n^1 \right) = x' < x^*. \tag{10.5}$$

(The case $x' > x^*$ can be proven symmetrically.) If agent $(n-m)$ were to report preference $\succeq_{n-m}^0$ instead, f would choose outcome y_m; hence *strategy-proofness* implies $y_m \leq x' < x^*$. See Figure 10.1.

Denote the outcomes that agent $(n-m)$ can obtain by varying his preferences, fixing the others, as[5]

$$O = \left\{ x : \exists \tilde{\succeq}_{n-m} \text{ s.t. } x = f\left(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \tilde{\succeq}_{n-m}, \succeq_{n-m+1}^1, \ldots, \succeq_n^1 \right) \right\}.$$

By definition, $x' \in O$; Case 1 implies $y_m, y_{m+1} \in O$. *Strategy proofness* implies that $x' = \max\{x \in O : x \leq x^*\}$, otherwise by reporting some other preference, agent $(n-m)$ could obtain some $x \in (x', x^*)$, violating *strategy proofness*.

[5] The literature on *strategy proofness* refers to this as an option set.

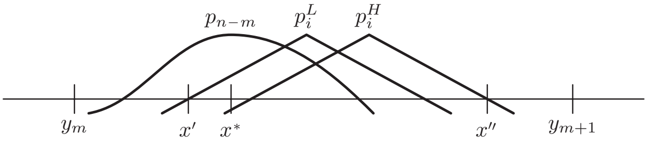

Figure 10.1. Proof of Theorem 10.2. If a *strategy-proof, onto* rule does not pick x^* when it is the median of peaks and y_m's, then a contradiction is reached using preferences with peaks at p_i^L and p_i^H.

Letting $x'' \equiv \inf\{x \in O : x \geq x^*\}$, *strategy proofness* implies $x'' \in O$.[6] To see this, let $\succeq''_{n-m}$ be a preference relation with peak $p''_{n-m} = x''$ and such that $(x'' + \epsilon) \succ''_{n-m} x'$ for some small $\epsilon > 0$. Then *strategy proofness* implies $f(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \succeq''_{n-m}, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)\} = \hat{x} \in [x'', x'' + \epsilon]$. But if $\hat{x} \neq x''$, then there would exist a misreport resulting in an outcome arbitrarily closer to x'', making agent $(n - m)$ (with preference $\succeq''_{n-m}$) better off. Hence $\hat{x} = x'' = \min\{x \in O : x \geq x^*\}$. With Eq. (10.5), we have $x'' > x^*$.

We have shown that $O \cap (x', x'') = \emptyset$. Let p_i^L be a symmetric preference relation with peak at $p^L = (x' + x'')/2 - \varepsilon$, where $\varepsilon > 0$ is sufficiently small; see Figure 10.1. Similarly let p_i^H be a symmetric preference relation with peak at $(x' + x'')/2 + \varepsilon$. Then *strategy-proofness* implies

$$f(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \succeq_{n-m}^H, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)\} = x''.$$

By repeated application of *strategy-proofness* (along the lines used in proving Eq. (10.3)), this implies

$$f(\succeq_1^L, \ldots, \succeq_{n-m-1}^L, \succeq_{n-m}^H, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)\} = x''.$$

Lemma 10.1 (Pareto-optimality) implies

$$f(\succeq_1^L, \ldots, \succeq_{n-m-1}^L, \succeq_{n-m}^L, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)\} \geq p_i^L.$$

Therefore, *strategy-proofness* implies

$$f(\succeq_1^L, \ldots, \succeq_{n-m-1}^L, \succeq_{n-m}^L, \succeq_{n-m+1}^1, \ldots, \succeq_n^1)\} = x'' \tag{10.6}$$

otherwise agent $n - m$ could manipulate at one of the two profiles (since ε is small).

On the other hand, *strategy-proofness* implies

$$f(\succeq_1^0, \ldots, \succeq_{n-m-1}^0, \succeq_{n-m}^L, \succeq_{n-m+1}^1, \ldots, \succeq_n^1) = x'$$

by the definition of $\succeq_i^L$. *Strategy-proofness* implies that if agent $(n - m - 1)$ instead reports preference $\succeq^L$, a point must be chosen that is in the interval $[x', x'' - 2\varepsilon]$, otherwise, he could report $\succeq^0$ to gain. By repeated application of this argument, this continues to hold as each agent $1 \leq i \leq n - m - 1$ changes his report from $\succeq_i^0$ to $\succeq_i^L$, so

$$f(\succeq_1^L, \ldots, \succeq_{n-m-1}^L, \succeq_{n-m}^L, \succeq_{n-m+1}^1, \ldots, \succeq_n^1) \in [x', x'' - 2\varepsilon].$$

[6] More generally, *strategy-proofness* alone implies O is closed. For brevity we prove only $x'' \in O$.

This contradicts Eq. (10.6). Hence Eq. (10.5) cannot hold, so $x' \geq x^*$; the symmetric argument implies $x' = x^*$, resulting in Eq. (10.4). Thus f chooses the median of these $2n - 1$ points for profile $\succeq$. □

The parameters (y_m's) in Theorem 10.2 can be thought of as the rule's degree of compromise when agents have extremist preferences. If m agents prefer the highest possible outcome (1), while $n - m$ prefer the lowest (0), then which point should be chosen? A true median rule would pick whichever extreme (0 or 1) contains the most peaks. On the other hand, the other rules described in the Theorem may choose intermediate points (y_m) as a compromise. The degree of compromise (which y_m) can depend on the degree to which the agents' opinions are divided (the size of m).

The *anonymity* requirement is a natural one in situations where agents are to be treated as equals. If one does not require this, however, the class of *strategy-proof* rules becomes even larger. We have already mentioned *dictatorial* rules, which always chooses a predetermined agent's peak. There are less extreme violations of anonymity: The full class of *strategy-proof*, *onto* rules, which we now define, allows agents to be treated with varying degrees of asymmetry.

Definition 10.3 A rule f is a *generalized median voter scheme* (g.m.v.s.) if there exist 2^n points in [0, 1], $\{\alpha_S\}_{S \subseteq N}$, such that

(i) $S \subseteq T \subseteq N$ implies $\alpha_S \leq \alpha_T$,

(ii) $\alpha_\emptyset = 0, \alpha_N = 1$, and

(iii) for all $\succeq \in \mathcal{R}^n$, $f(\succeq) = \max_{S \subset N} \min\{\alpha_S, p_i : i \in S\}$.

An example is given below. It is worth making two observations regarding Definition 10.3. First, the monotonicity condition (i) is actually redundant. If parameters $\{\alpha_S\}_{S \subseteq N}$ fail this condition, they still define some *strategy-proof* rule via condition (iii). However, the resulting rule could also be defined by an alternate set of parameters $\{\alpha'_S\}_{S \subseteq N}$ that do satisfy condition (i). Second, condition (ii) is present merely to guarantee the rule to be *onto*. Parameters that fail this condition still define a *strategy-proof* rule whose range is $[\alpha_\emptyset, \alpha_N]$.[7]

Consider the rule described by the parameters (α_S's) in Figure 10.2, for the 3-agent case. The reader should first verify the following. If each agent in some set $S \subseteq N$ were to have a preference peak at 1, while each remaining agent (in $N \setminus S$) were to have a preference peak at 0, then the rule would choose α_S as the outcome. In this sense, the α_S parameters reflect a (nonanonymous) degree of compromise at extreme preference profiles, analogous to the y_m parameters of Theorem 10.2.

Without the anonymity condition, some agents – more generally some *coalitions* of agents – are more powerful than others. To see this, consider the profile of preferences represented in Figure 10.2 with peaks p_1, p_2, p_3. Following condition (iii) of Definition 10.3, calculate $\min\{\alpha_S, p_i : i \in S\}$ for each $S \subseteq N$. Beginning with the three

[7] To avoid potential confusion, we point out that, in some of the literature, the term *generalized median voter scheme* also refers to such rules.

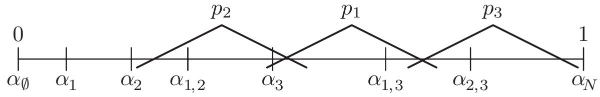

Figure 10.2. An example of a generalized median voter scheme for $n = 3$.

singleton coalitions of the form $S = \{i\}$, these values are α_1, α_2, and α_3, because each p_i is above that agent's corresponding $\alpha_{\{i\}}$. (For peak p_3', the third value would have been p_3' instead.) Since the g.m.v.s. eventually chooses the maximum of these kinds of values (after we also check larger coalitions), agent 3 can be said to have more power than the other two agents, *as a singleton*. A large α_3 corresponds to more instances in which agent 3's peak is a candidate outcome for this rule. A small α_1 corresponds to more instances in which agent 1 has no impact on the outcome (i.e., whenever $p_1 > \alpha_{\{1\}}$).

On the other hand, we also need to calculate these minimum-values for larger coalitions. For the pairs of agents $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$, these values are $\alpha_{\{1,2\}}$, p_1, and p_2 respectively. Coalition $\{1, 2\}$ is the weakest two-agent coalition in the sense that they have the lowest α_S. After checking $S = \emptyset$ (which yields 0) and $S = N$ (yielding a repetition of the value p_2), we calculate the rule's outcome to be the maximum of the 2^n values $\{0, \alpha_1, \alpha_2, \alpha_3, \alpha_{\{1,2\}}, p_1, p_2, p_2\}$ we have obtained, which is $\alpha_{\{3\}}$.

We close by stating the main result of this section. We omit its proof, which has much in common with the proof of Theorem 10.2.

Theorem 10.4 *A rule f is* strategy-proof *and* onto *if and only if it is a generalized median voter scheme.*

10.2.2 Application to Public Good Cost Sharing

Consider a group of n agents who have access to a machine that can convert their labor into some public good. Specifically, suppose that the machine requires the simultaneous labor of all n agents in order to work. The agents are free to jointly decide how many hours of labor, ℓ, to work. Implicit is the requirement that each agent work for ℓ hours, however, since the machine requires all n agents' labor simultaneously. After ℓ hours of labor, the machine outputs $y = Y(\ell)$ units of some public good, where the production function Y is assumed to be an increasing and strictly concave function, with $Y(0) = 0$.

Different agents may have different preferences over how much labor they should provide, in exchange for the public good. Let us suppose that we know nothing about their preferences, other than the fact that they are represented by some utility function $u_i(\ell, y)$ which is strictly increasing in y, strictly decreasing in ℓ, and is quasi-concave.[8] See Figure 10.3.

In this environment, a rule takes as input the reported utility functions of the agents, subject only to the assumptions we have made. It then gives as output a single labor requirement $\ell = f(u_1, \ldots, u_n)$. Each agent is then required to provide ℓ units of labor,

[8] The function $u()$ is quasi-concave if, at each (ℓ, y), the upper contour set $\{(\ell', y') : u(\ell', y') \geq u(\ell, y)\}$ is convex.

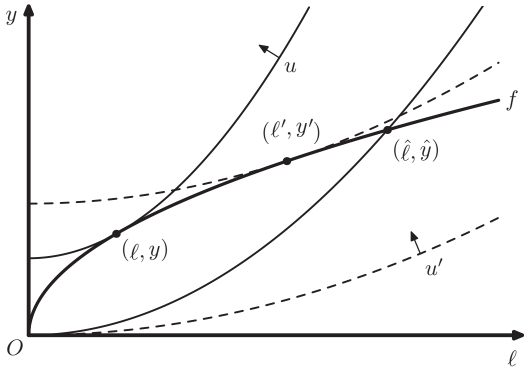

Figure 10.3. An agent with utility function u most prefers the outcome (y, ℓ); one with u' prefers (y', ℓ').

and they enjoy $Y(\ell)$ units of output as a reward. What rules are *strategy-proof* and *onto*?

By assumption, outcomes may only be attained along the graph of Y. Because of the assumptions on Y and on preferences, it is clear that agents have single-peaked preferences over this consumption space. It follows that any *strategy-proof*, *onto* rule for this environment is a generalized median voter schemes operating along the graph of Y.

Proving this is not difficult, but involves some technical details that we omit. First the outcome space is not bounded as we assumed before, although it would certainly be reasonable to bound it by assumption. Second, the preference domain here should be verified to yield all the single-peaked preferences necessary to characterize generalized median voter schemes; e.g., we used *symmetric* single-peaked preferences to construct the proof of Theorem 10.2. Third, one should demonstrate that a *strategy-proof* rule in this environment is invariant to utility information away from the graph of Y. We leave it to the interested reader to verify our claim despite these technicalities.

In this kind of problem, it may be reasonable to add additional requirements to a rule. One that we address is the requirement that an agent should be better off as part of this decision-making group than if he were simply to walk away. Formally, if this public good technology did not exist, each agent would provide no labor ($\ell = 0$), and would enjoy none of the public good ($y = 0$). We say a rule is *individually rational* if for all $U = (u_1, \ldots, u_n)$ and $1 \geq i \geq n$, we have $u_i(f(U), Y(f(U))) \geq u_i(0, 0)$.

What *strategy-proof* and *onto* rules satisfy *individual rationality*? In terms of our earlier model, where agents have single-peaked preferences on [0, 1], that question translates as follows: What g.m.v.s. has the property that, for any preference profile, each agent (weakly) prefers the chosen outcome to the outcome $x = 0$?

The answer is that there is a unique such rule. As an exercise, we leave it to the reader to show that the rule that chooses the minimum peak is the unique *strategy-proof*, *onto* rule that satisfies this individual rationality condition. In terms of this public good model, this corresponds to asking each agent their most preferred labor level ℓ, and choosing the minimum.

10.3 House Allocation Problem

The House allocation problem is a model for understanding the allocation of indivisible goods. It involves a set N of n agents, each owning a unique house and a strict preference ordering over all n houses. The objective is to reallocate the houses among the agents in an appropriate way. A modern version of the same would replace houses by kidneys.

While any possible (strict) preference ordering over the homes is permitted, the set of preferences over allocations is restricted. In particular, an agent is indifferent between all allocations that give her the same house. Therefore the Gibbard–Satterthwaite Theorem does not apply in this setting.

One could select an allocation of homes in a variety of ways, perhaps so as to optimize some function of the preferences and then investigate if the resulting allocation rule is *strategy-proof*. However, this ignores an important feature not present in earlier examples. In this environment, agents control the resources to be allocated. Therefore an allocation can be subverted by a subset of agents who might choose to break away and trade among themselves. For this reason it is natural to focus on allocations that are invulnerable to agents opting out.

Number each house by the number of the agent who owns that house. An allocation is an n vector a whose ith component, a_i, is the number of the house assigned to agent i. If a is the initial allocation then $a_i = i$. For an allocation to be feasible, we require that $a_i \neq a_j$ for all $i \neq j$. The preference ordering of an agent i will be denoted $\succ_i$ and $x \succ_i y$ will mean that agent i ranks house x above house y. Denote by A the set of all feasible allocations. For every $S \subseteq N$ let $A(S) = \{z \in A : z_i \in S \ \forall i \in S\}$ denote the set of allocations that can be achieved by the agents in S trading among themselves alone. Given an allocation $a \in A$, a set S of agents is called a **blocking coalition** (for a) if there exists a $z \in A(S)$ such that for all $i \in S$ either $z_i \succ_i a_i$ or $z_i = a_i$ and for at least one $j \in S$ we have that $z_j \succ_j a_j$. A blocking coalition can, by trading among themselves, receive homes that each strictly prefers (or is equivalent) to the home she receives under a, with at least one agent being strictly better off. The set of allocations that is not blocked by any subset of agents is called the **core**.

The reader will be introduced to the notion of the core in Chapter 15 (Section 15.2) where it will be defined for a cooperative game in which utility is transferable via money (a TU game). The house allocation problem we consider is an example of a cooperative game with nontransferable utility (an NTU game). The definition of the core offered here is the natural modification of the notion of TU core to the present setting.

The theorem below shows the core to be nonempty. The proof is by construction using the top trading cycle algorithm (TTCA).

Definition 10.5 (Top Trading Cycle Algorithm) Construct a directed graph using one vertex for each agent. If house j is agent i's kth ranked choice, insert a directed edge from i to j and color the edge with color k. An edge of the form (i, i) will be called a loop. First, identify all directed cycles and loops consisting only of edges colored 1. The strict preference ordering implies that the set of such cycles and loops is node disjoint. Let N_1 be the set of vertices (agents) incident to these cycles. Each cycle implies a sequence of swaps. For example, suppose $i_1 \to i_2 \to i_3 \to \cdots \to i_r$ is one such cycle. Give house i_1 to agent i_r, house i_r to agent i_{r-1}, and so on. After all such swaps are performed, delete all edges colored 1. Repeat with the edges colored 2 and call the corresponding set of vertices incident to these edges N_2, and so on. The TTCA yields the resulting matching.

This algorithm is used to prove the following result.

Theorem 10.6 *The core of the house allocation problem consists of exactly one matching.*

PROOF We prove that if a matching is in the core, it must be the one returned by the TTCA.

Under the TTCA, each agent in N_1 receives his favorite house, i.e., the house ranked first in his preference ordering. Therefore, N_1 would form a blocking coalition to any allocation that does not assign to all of those agents the houses they would receive under the TTCA. That is, any core allocation must assign N_1 to houses just as the TTCA assigns them.

Given this fact, the same argument applies to N_2: Under the TTCA, each agent in N_2 receives his favorite house *not including* those houses originally endowed by agents in N_1. Therefore, if an allocation is in the core and the agents in N_1 are assigned each other's houses, then agents in N_2 must receive the same houses they receive under the TTCA.

Continuing the argument for each N_k proves that if an allocation is in the core, then it is the one determined by the TTCA. This proves that there is at most one core allocation.

To prove that the TTCA allocation is in the core, it remains to be shown that there is no other blocking coalition $S \subseteq N$. This is left to the reader. □

To apply the TTCA, one must know the preferences of agents over homes. Do they have an incentive to truthfully report these? To give a strongly positive answer to this question, we first associate the TTCA with its corresponding direct revelation mechanism. Define the **Top Trading Cycle (TTC) Mechanism** to be the function (mechanism) that, for each profile of preferences, returns the allocation computed by the TTCA.

Theorem 10.7 *The TTC mechanism is* strategy-proof.

PROOF Let π be a profile of preference orderings and a the allocation returned by TTCA when applied to π. Suppose that agent $j \in N_k$ for some k misreports her preference ordering. Denote by π' the new profile of preference orderings. Let a' the allocation returned by TTCA when applied to π'. If the TTCA is not *strategy-proof* $a'_i >^i a_i$. Observe that $a_i = a'_i$ for all $i \in \bigcup_{r=1}^{k-1} N_r$. Therefore, $a'_i \in N \setminus \{\bigcup_{r=1}^{k-1} N_r\}$. However, the TTCA chooses a_i to be agent i's top ranked choice from $N \setminus \{\bigcup_{r=1}^{k-1} N_r\}$ contradicting the fact that $a'_i >^i a_i$. □

If we relax the requirement that preferences be strict, what we had previously called a blocking set is now called a **weakly** blocking set. What we had previously called the core is now called the *strict* core. With indifference, a **blocking** set S is one where *all* agents in S are *strictly* better off by trading among themselves. Note the requirement that all agents be strictly better off. The *core* is the set of allocations not blocked by any set S.

When preferences are strict, every minimal weakly blocking set is a blocking set. To see this, fix a weakly blocking set S. An agent in S who is not made strictly better off by trade among agents in S must have been assigned their own home. Remove them from S. Repeat. The remaining agents must all be allocated houses that make them strictly better off. Hence, when preferences are strict the core and strict core coincide. With indifference permitted, the strict core can be different from the core. In fact, there are examples where the strict core is empty and others where it is not unique. Deciding emptiness of the strict core is polynomial in $|N|$.

Another possible extension of the model is to endow the agents with more than one good. For example, a home and a car. Clearly, if preferences over pairs of goods are sufficiently rich, the core can be empty. It turns out that even under very severe restrictions the core can still be empty. For example, when preferences are separable, i.e., one's ranking over homes does not depend on which car one has.

10.4 Stable Matchings

The stable matching problem was introduced as a model of how to assign students to colleges. Since its introduction, it has been the object of intensive study by both computer scientists and economists. In computer science it used as vehicle for illustrating basic ideas in the analysis of algorithms. In economics it is used as a stylized model of labor markets. It has a direct real-world counterpart in the procedure for matching medical students to residencies in the United States.

The simplest version of the problem involves a set M of men and a set W of women. Each $m \in M$ has a strict preference ordering over the elements of W and each $w \in W$ has a strict preference ordering over the men. As before the preference ordering of agent i will be denoted $\succ_i$ and $x \succ_i y$ will mean that agent i ranks x above y. A **matching** is an assignment of men to women such that each man is assigned to at most one woman and vice versa. We can accommodate the possibility of an agent choosing to remain single as well. This is done by including for each man (woman) a dummy woman (man) in the set W (M) that corresponds to being single (or matched with oneself). With this construction we can always assume that $|M| = |W|$.

As in the house allocation problem a group of agents can subvert a prescribed matching by opting out. In a manner analogous to the house allocation problem, we can define a blocking set. A matching is called **unstable** if there are two men m, m' and two women w, w' such that

(i) m is matched to w,
(ii) m' is matched to w', and
(iii) $w' \succ_m w$ and $m \succ_{w'} m'$

The pair (m, w') is called a **blocking pair**. A matching that has no blocking pairs is called **stable**.

Example 10.8 The preference orderings for the men and women are shown in the table below

$\succ_{m_1}$	$\succ_{m_2}$	$\succ_{m_3}$	$\succ_{w_1}$	$\succ_{w_2}$	$\succ_{w_3}$
w_2	w_1	w_1	m_1	m_3	m_1
w_1	w_3	w_2	m_3	m_1	m_3
w_3	w_2	w_3	m_2	m_2	m_2

Consider the matching $\{(m_1, w_1), (m_2, w_2), (m_3, w_3)\}$. This is an unstable matching since (m_1, w_2) is a blocking pair. The matching $\{(m_1, w_1), (m_3, w_2), (m_2, w_3)\}$, however, is stable.

Given the preferences of the men and women, is it always possible to find a stable matching? Remarkably, yes, using what is now called the deferred acceptance algorithm. We describe the male-proposal version of the algorithm.

Definition 10.9 (Deferred Acceptance Algorithm, male-proposals) First, each man proposes to his top-ranked choice. Next, each woman who has received at least two proposals keeps (tentatively) her top-ranked proposal and rejects the rest. Then, each man who has been rejected proposes to his top-ranked choice among the women who have not rejected him. Again each woman who has at least two proposals (including ones from previous rounds) keeps her top-ranked proposal and rejects the rest. The process repeats until no man has a woman to propose to or each woman has at most one proposal. At this point the algorithm terminates and each man is assigned to a woman who has not rejected his proposal. Notice that no man is assigned to more than one woman. Since each woman is allowed to keep only one proposal at any stage, no woman is assigned to more than one man. Therefore the algorithm terminates in a matching.

We illustrate how the (male-proposal) algorithm operates using Example 10.8 above. In the first round, m_1 proposes to w_2, m_2 to w_1, and m_3 to w_1. At the end of this round w_1 is the only woman to have received two proposals. One from m_3 and the other from m_2. Since she ranks m_3 above m_2, she keeps m_3 and rejects m_2. Since m_3 is the only man to have been rejected, he is the only one to propose again in the second round. This time he proposes to w_3. Now each woman has only one proposal and the algorithm terminates with the matching $\{(m_1, w_2), (m_2, w_3), (m_3, w_2)\}$. It is easy to verify that the matching is stable and that it is different from the one presented earlier.

Theorem 10.10 *The male propose algorithm terminates in a stable matching.*

PROOF Suppose not. Then there exists a blocking pair (m_1, w_1) with m_1 matched to w_2, say, and w_1 matched to m_2. Since (m_1, w_1) is blocking and $w_1 \succ_{m_1} w_2$, in the proposal algorithm, m_1 would have proposed to w_1 before w_2. Since m_1 was not matched with w_1 by the algorithm, it must be because w_1 received a proposal from a man that she ranked higher than m_1. Since the algorithm matches her to m_2 it follows that $m_2 \succ_{w_1} m_1$. This contradicts the fact that (m_1, w_1) is a blocking pair. □

One could just as well have described an algorithm where the women propose and the outcome would also be a stable matching. Applied to the example above, this would produce a stable matching different from the one generated when the men propose. Thus, not only is a stable matching guaranteed to exist but there can be more than 1. If there can be more than one stable matching, is there a reason to prefer one to another? Yes. To explain why, some notation.

Denote a matching by μ. the woman assigned to man m in the matching μ is denoted $\mu(m)$. Similarly, $\mu(w)$ is the man assigned to woman w. A matching μ is **male-optimal** if there is no stable matching ν such that $\nu(m) \succ_m \mu(m)$ or $\nu(m) = \mu(m)$ for all m with $\nu(j) \succ_j \mu(j)$ for at least one $j \in M$. Similarly define **female-optimal**.

Theorem 10.11 *The stable matching produced by the (male-proposal) Deferred Acceptance Algorithm is male-optimal.*

PROOF Let μ be the matching returned by the male-propose algorithm. Suppose μ is not male optimal. Then, there is a stable matching ν such that $\nu(m) \succ_m \mu(m)$ or $\nu(m) = \mu(m)$ for all m with $\nu(j) \succ_j \mu(j)$ for at least one $j \in M$. Therefore, in the application of the proposal algorithm, there must be an iteration where some man j proposes to $\nu(j)$ before $\mu(j)$ since $\nu(j) \succ_j \mu(j)$ and is rejected by woman $\nu(j)$. Consider the first such iteration. Since woman $\nu(j)$ rejects j she must have received a proposal from a man i she prefers to man j. Since this is the first iteration at which a male is rejected by his partner under ν it follows that man i ranks woman $\nu(j)$ higher than $\nu(i)$. Summarizing, $i \succ_{\nu(j)} j$ and $\nu(j) \succ_i \nu(i)$ implying that ν is not stable, a contradiction. □

Clearly one can replace the word "male" by the word "female" in the statement of the theorem above. It is natural to ask if there is a stable matching that would be optimal with respect to both men and women. Alas, no. The example above has two stable matchings: one male optimal and the other female optimal. At least one female is strictly better off under the female optimal matching than the male optimal one and no female is worse off. A similar relationship holds when comparing the two stable matchings from the point of view of the men.

A stable matching is immune to a pair of agents opting out of the matching. We could be more demanding and ask that no subset of agents should have an incentive to opt out of the matching. Formally, a matching μ' **dominates** a matching μ if there is a set $S \subset M \cup W$ such that for all $m, w \in S$, both (i) $\mu'(m), \mu'(w) \in S$ and (ii) $\mu'(m) \succ_m \mu(m)$ and $\mu'(w) \succ_w \mu(w)$. Stability is a special case of this dominance condition when we restrict attention to sets S consisting of a single couple. The set of undominated matchings is called the **core** of the matching game. The next result is straightforward.

Theorem 10.12 *The core of the matching game is the set of all stable matchings.*

Thus far we have assumed that the preference orderings of the agents is known to the planner. Now suppose that they are private information to the agent. As before we can associate a direct revelation mechanism with an algorithm for finding a stable matching.

Theorem 10.13 *The direct mechanism associated with the male propose algorithm is* strategy-proof *for the males.*

PROOF Suppose not. Then there is a profile of preferences $\pi = (\succ_{m_1}, \succ_{m_2}, \dots, \succ_{m_n})$ for the men, such that man m_1, say, can misreport his preferences and obtain a better match. To express this formally, let μ be the stable matching obtained by applying the male proposal algorithm to the profile π. Suppose that m_1 reports the preference ordering $\succ_*$ instead. Let ν be the stable matching that results when the male-proposal algorithm is applied to the profile $\pi^1 = (\succ_*, \succ_{m_2}, \dots, \succ_{m_n})$. For a contradiction, suppose $\nu(m_1) \succ_{m_1} \mu(m_1)$. For notational convenience we will write $a \succeq_m b$ to mean that $a \succ_m b$ or $a = b$.

First we show that m_1 can achieve the same effect by choosing an ordering $\succsim$ where woman $\nu(m_1)$ is ranked first. Let $\pi^2 = (\succsim, \succ_{m_2}, \dots, \succ_{m_n})$. Knowing that ν is stable with respect to the profile π^1 we show that it is stable with respect to the profile π^2. Suppose not. Then under the profile π^2 there must be a pair (m, w) that blocks ν. Since ν assigns to m_1 its top choice with respect to π^2, m_1 cannot be part of this blocking pair. Now the preferences of all agents other than m_1 are the same in π^1 and π^2. Therefore, if (m, w) blocks ν with respect to the profile π^2, it must block ν with respect to the profile π^1, contradicting the fact that ν is a stable matching under π^1.

Let λ be the male propose stable matching for the profile π^2. Since ν is a stable matching with respect to the profile π^2. As λ is male optimal with respect to the profile π^2, it follows that $\lambda(m_1) = \nu(m_1)$.

Thus we can assume that $\nu(m_1)$ is the top-ranked woman in the ordering $\succ_*$. Next we show that the set $B = \{m_j : \mu(m_j) \succ_{m_j} \nu(m_j)\}$ is empty. This means that all men, not just m_1, are no worse off under ν compared to μ. Since ν is stable with respect to the original profile, π this contradicts the male optimality of μ and completes the proof.

Suppose $B \neq \emptyset$. Therefore, when the male proposal algorithm is applied to the profile π^1, each $m_j \in B$ is rejected by their match under μ, i.e., $\mu(m_j)$. Consider the first iteration of the proposal algorithm where some m_j is rejected by $\mu(m_j)$. This means that woman $\mu(m_j)$ has a proposal from man m_k that she ranks higher, i.e., $m_k \succ_{\mu(m_j)} m_j$. Since m_k was not matched to $\mu(m_j)$ under μ it must be that $\mu(m_k) \succ_{m_k} \mu(m_j)$. Hence $m_k \in B$, otherwise

$$\mu(m_j) \succeq m_k \nu(m_k) \succeq_{m_k} \mu(m_k) \succ_{m_k} \mu(m_j),$$

which is a contradiction.

Since $m_k \in B$ and m_k has proposed to $\mu(m_j)$ at the time man m_j proposes, it means that m_k must have been rejected by $\mu(m_k)$ prior to m_j being rejected, contradicting our choice of m_j. □

The mechanism associated with the male propose algorithm is not *strategy-proof* for the females. To see why, it is enough to consider example. The male propose algorithm returns the matching $\{(m_1, w_2), (m_2, w_3), (m_3, w_1)\}$. In the course of the algorithm the only woman who receives at least two proposals is w_1. She received proposals from m_2 and m_3. She rejects m_2 who goes on to propose to w_3 and the algorithm terminates.

Notice that w_1 is matched with her second choice. Suppose now that she had rejected m_3 instead. Then m_3 would have gone on to proposes to w_2. Woman w_2 now has a choice between m_1 and m_3. She would keep m_3 and reject m_1, who would go on to propose to w_1. Woman w_1 would keep m_1 over m_2 and in the final matching be paired with a her first-rank choice.

It is interesting to draw an analogy between the existence of stable matchings and that of Walrasian equilibrium. We know (Chapter 6) that Walrasian equilibria exist. Furthermore, they are the solutions of a fixed point problem. In the cases when they can be computed efficiently it is because the set of Walrasian equilibria can be described by a set of convex inequalities. The same can be said of stable matchings. The set of stable matchings is fixed points of a nondecreasing function defined on a lattice. In addition, one can describe the set of stable matchings as the solutions to a set of linear inequalities.

10.4.1 A Lattice Formulation

We describe a proof of the existence of stable matchings using Tarski's fixed point theorem. It will be useful to relax the notion of a matching. Call an assignment of women to men such that each man is assigned to at most one woman (but a woman may be assigned to more than one man) a **male semimatching**. The analogous object for women will be called a **female semimatching**. For example, assigning each man his first choice would be a male semimatching. Assigning each woman her third choice would be an example of a female semimatching.

A pair of male and female semimatchings will be called a **semimatching** which we will denote by μ, ν, etc. An example of a semi-matching would consist of each man being assigned his first choice and each woman being assigned her last choice.

The woman assigned to the man m under the semi-matching μ will be denoted $\mu(m)$. If man m is assigned to no woman under μ, then $\mu(m) = m$. Similarly for $\mu(w)$. Next we define a partial order over the set of semimatchings. Write $\mu \succeq \nu$ if

(i) $\mu(m) \succ_m \nu(m)$ or $\mu(m) = \mu(m)$ for all $m \in M$ and
(ii) $\mu(w) \prec_w \nu(w)$ or $\mu(w) = \nu(w)$ for all $w \in W$.

Therefore $\mu \succeq \nu$ if all the men are better off under μ than in ν and all the women are worse off under μ than in ν.

Next we define the meet and join operations. Given two semimatchings μ and ν define $\lambda = \mu \vee \nu$ as follows:

(i) $\lambda(m) = \mu(m)$ if $\mu(m) \succ_m \nu(m)$ otherwise $\lambda(m) = \nu(m)$,
(ii) $\lambda(w) = \mu(w)$ if $\mu(w) \prec_w \nu(w)$ otherwise $\lambda(w) = \nu(w)$.

Define $\lambda' = \mu \wedge \nu$ as follows:

(i) $\lambda'(m) = \mu(m)$ if $\mu(m) \prec_m \nu(m)$ otherwise $\lambda(m) = \nu(m)$,
(ii) $\lambda(w) = \mu(w)$ if $\mu(w) \succ_w \nu(w)$ otherwise $\lambda(w) = \nu(w)$.

With these definitions it is easy to check that the set of semimatchings forms a compact lattice.

Now define a function f on the set of semi-matchings that is nondecreasing. Given a semi-matching μ define $f(\mu)$ to be the following semi-matching:

(i) $f(\mu)(m)$ is man m's most preferred woman from the set $\{w\colon m \succ_w \mu(w), m = \mu(w)\}$. If this set is empty set $f(\mu)(m) = m$.

(ii) $f(\mu)(w)$ is woman w's most preferred man from the set $\{m\colon w \succ_m \mu(m), w = \mu(m)\}$. If this set is empty set $f(\mu)(w) = w$.

It is clear that f maps semi-matchings into semi-matchings.

Theorem 10.14 *There is a semi-matching μ such that $f(\mu) = \mu$ and that μ is a stable matching.*

PROOF We use Tarski's theorem. It suffices to check that f is nondecreasing. Suppose $\mu \succeq \nu$. Pick any $m \in M$. From the definition of $\succeq$, the women are worse off under μ than in ν. Thus

$$\{w\colon m \succ_w \nu(w)\} \subseteq \{w\colon m \succ_w \mu(w)\}$$

and so $f(\mu)(m) \succ_m f(\nu)(m)$ or $f(\mu)(m) = f(\nu)(m)$. A similar argument applies for each $w \in W$. Thus f is nondecreasing.

Since the conditions of Tarski's theorem hold, it follows that there is a semi-matching μ such that $f(\mu) = \mu$. We show that the semi-matching is a stable matching.

By the definition of a semi-matching we have for every $m \in M$, $\mu(m)$ single valued as is $\mu(w)$ for all $w \in W$. To show that μ is a matching, suppose not. Then there is a pair $m_1, m_2 \in M$, say, such that $\mu(m_1) = \mu(m_2) = w^*$. Since $f(\mu) = \mu$ it follows that w^* is m_1's top-ranked choice in $\{w\colon m_1 \succ_w \mu(w), m_1 = \mu(w)\}$ and m_2's top ranked choice in $\{w\colon m_2 \succ_w \mu(w), m_2 = \mu(w)\}$. From this we deduce that $\mu(w^*) = m_3$ where $m_1, m_2 >^{w^*} m_3$. However, $m_3 = \mu(w^*) = f(\mu^*)(w^*)$, which is woman w^*'s top-ranked choice in $\{m\colon w^* \succ_m \mu(m), \mu(m) = w^*\}$. Since m_1, m_2 are members of this set, we get a contradiction.

To show that the matching μ is stable suppose not. Then there must be a blocking pair (m^*, w^*). Let $w' = \mu(m^*)$ and $m' = \mu(w^*)$, $m' \neq m^*$ and $w^* \neq w'$. Since (m^*, w^*) is blocking, $m^* \succ_{w^*} m'$ and $w^* \succ_{m^*} w'$. Now $w' = \mu(m^*) = f(\mu)(m^*)$, which is man m^*'s top-ranked choice from $\{w : m^* \succ_w \mu(w), m^* = \mu(w)\}$. But this set contains w^*, which is ranked higher by man m^* than w', a contradiction. □

10.4.2 The LP Formulation

One can formulate the problem of finding a stable matching as the solution to a set of linear inequalities. For each man m and woman w let $x_{mw} = 1$ if man m is matched with woman w and zero otherwise. Then, every stable matching must satisfy the following.

$$
\begin{aligned}
\sum_{w \in W} x_{mw} &= 1 && \forall m \in M \\
\sum_{m \in M} x_{mw} &= 1 && \forall w \in W \\
\sum_{j \prec_m w} x_{mj} + \sum_{i \prec_w m} x_{iw} + x_{mw} &\leq 1 && \forall m \in M, w \in W \\
x_{mw} &\geq 0 && \forall m \in M, w \in W
\end{aligned}
$$

Let P be the polyhedron defined by these inequalities.

The first two constraints of P ensure that each agent is matched with exactly one other agent of the opposite sex. The third constraint ensures stability. To see why, suppose $\sum_{j \prec_m w} x_{mj} = 1$ and $\sum_{i \prec_w m} x_{iw} = 1$. Then man m is matched to a woman, j that he ranks below w. Similarly, woman w is matched to a man she ranks below m. This would make the pair (m, w) a blocking pair.

Theorem 10.15 *P is the convex hull of all stable matchings.*

10.4.3 Extensions

We have been careful to specify that preferences are strict. If we allow for indifference, Theorem 10.7 becomes false. This is because there are instances of the stable matching problem in which no male or female optimal stable matching exists. The other theorems stated above continue to hold in the presence of indifferences.

We also limited ourselves to one-to-one matchings. There are situations where one side of the market wishes to match with more than one agent. The college admissions market is the classic example. Each student can be assigned to at most one college but each college can be assigned to many students. In this more general setup colleges will have preferences over subsets of students. In the absence of any restrictions on these preferences a stable matching need not exist. One restriction on preferences for which the results above carry over with no change in statement or proof is the quota model. Each college has a strict preference ordering over the students and a quota r of students it wishes to admit. Consider two subsets, S and T, of students of size r that differ in exactly one student. The college prefers the subset containing the more preferred student.

A third extension is to relax the bipartite nature of the stable matching problem. The nonbipartite version is called the stable roommates problem. Suppose that a set of N individuals such that $|N|$ is even. A matching in this setting is a partition of N into disjoint pairs of individuals (roommates). Each individual has a strict preference ordering over the other individuals that they would like to be paired with. As before, a matching is unstable if there exists a pair who prefer each other to the person they are matched with. Such a pair is called blocking. Unlike the stable matching problem, stable roommates need not exist as the following four person example illustrates.

$\succ_1$	$\succ_2$	$\succ_3$	$\succ_4$
3	1	2	2
2	3	1	1
4	4	4	4

Each column lists the preference ordering that one agent has over the others. A matching that pairs agent 1 with agent 4 will always be blocked by the pair (1, 2). A matching that pairs 2 with 4 will be blocked by (2, 3). A matching that pairs 3 and 4 will be blocked by (3, 1).

An $O(|N|^2)$ algorithm to determine if a stable matching exists is known. One can also associate a collection of linear inequalities with the stable roommates problem such that the system is feasible if and only if a stable roommates solution exists.

10.5 Future Directions

While the models in this chapter have been studied and extended in a variety of ways, there are plenty of open questions for the creative researcher.

One direction of future research on the single-peaked preference model of Section 10.2 would be to consider choosing multiple alternatives (locations) on an interval (or more general graph) when agents' preferences are single-peaked with respect to the one location that is *closest* to his peak. As an idealized example, when downloading files on the Internet one cares only about the location (distance) of the closest "mirror" site. If a planner can elicit preferences to choose the location of k mirrors on a network, how can this be done in a *strategy-proof* way?

As for the house allocation model of Section 10.3 and the stable matching model of Section 10.4, observe that both models are static in nature. Yet, there are a variety of *dynamic* environments that resemble these models in important ways. As an example, take the problem of allocating kidneys. Until quite recently those needing a kidney transplant would have to wait in a queue (the wait list) for an available kidney that would be an appropriate "fit" or else find a donor fulfilling the appropriate medical conditions.

More recently, however, exchange systems have been implemented which allow kidney patients to "swap" their incompatible (but willing) friends and relatives who are willing to donate a kidney. (Suppose that Alice needs a kidney, and her incompatible friend Bob is willing to donate; also suppose that Carmina and Dijen are in a similar situation. If Alice and Dijen are compatible, and if Carmina and Bob are compatible, then a compatible "swap" can be arranged.) Static versions of such a model have been analyzed by Roth, Sönmez, and Ünver (2004).

Those authors and others have developed a substantial literature around this important problem. If donors and recipients arrive dynamically to such a setting, how should swaps be arranged?

10.6 Notes and References

The canonical results for the single-peaked preference model are provided by Moulin (1980), who proved Theorems 10.2 and 10.4 with the additional requirement that rules take agents' peaks as their only input. Ching (1997) subsequently showed that this requirement is redundant when a rule is *strategy-proof* and *onto*.

Border and Jordan (1983) generalize these conclusions to *multidimensional* models where the outcome space is $\mathbb{R}^k$. They restrict attention to *separable* preferences, i.e., under the assumption that an agent's (relative) preferences over any one dimension are fixed, as we vary any other dimensions of the altnerative. For example with $k = 3$, if $(x_1, x_2, x_3) \succeq_i (x'_1, x_2, x_3)$ then separability would imply $(x_1, y_2, y_3) \succeq_i (x'_1, y_2, y_3)$. Border and Jordan show that a *strategy-proof*, *onto* rule for separable preferences must be decomposable into k (possibly different) one-dimensional rules. Of course, these one-dimensional rules must be generalized median voter schemes. For further reference on such generalizations, one should consult the survey of Barberà (2001).

Another direction in which these results have been generalized pertains to situations in which agents have single-peaked preferences on graphs. Schummer and Vohra (2004) obtain two types of result, depending on whether the graph contains any cycle. Finally, the book of Austen-Smith and Banks (2005). contains more details on the key results of this literature, and a proof of Theorem 10.4.

The house allocation problem was introduced by Herbert Scarf and Lloyd Shapley (1974). The TTCA is attributed by these authors to David Gale. The idea that the house allocation problem can be used as a model for kidney exchanges is discussed in Roth et al. (2004).

The stable matching problem was introduced by David Gale and Lloyd Shapley (1962). The first algorithm for finding a stable matching was developed a decade earlier in 1951 to match interns to hospitals (Stalnaker, 1953). The intrinsic appeal of the model has inspired three books. The first, by Donald Knuth (1976) uses the stable matching problem as a vehicle to illustrate some of the basic ideas in the analysis of algorithms. The book by Gusfield and Irving (1989) is devoted to algorithmic aspects of the stable matching problem and some of its relatives. On the economics side, the book by Roth and Sotomayor (1991) gives a complete game theoretic treatment of the stable matching problem as well as some of its relatives.

The lattice theoretic treatment of the stable matching problem goes back to Knuth (1976). The proof of existence based on Tarski's fixed point theorem is due to Adachi (2000). In fact, the proposal algorithm is exactly one of the algorithms for finding a fixed point when specialized to the case of stable matchings.

The linear programming formulation of the stable matching problem is due to Vande Vate (1989). The extension of it to the stable room mates problem can be found in Teo and Sethuraman (1998). Gusfield and Irving (1989) give a full algorithmic account of the stable roommates problem.

In parallel, studies have been made of matching models where monetary transfers are allowed. This has inspired models that unify both the stable matching problem as well as matching problems where monetary transfers are allowed. Descriptions can be found in Fleiner (2003) and Hatfield and Milgrom (2005).

Bibliography

H. Adachi. On a characterization of stable matchings. *Economics Letters*, 68:43–49, 2000.

D. Austen-Smith and J. Banks. *Positive Political Theory II: Strategy and Structure*. University of Michigan Press, 2005.

S. Barberà. An introduction of strategy-proof social choice functions. *Soc. Choice Welfare*, 18(4):619–653, 2001.

K. Border and J. Jordan. Straightforward elections, unanimity and phantom voters. *Rev. Econ. Stud.*, 50(1):153–170, 1983.

S. Ching. Strategy-proofness and Âmedian voters. *Intl. J. Game Theor.*, 26(4):473–490, 1997.

T. Fleiner. Some results on stable matchings and fixed points. *Math. Oper. Res.*, 28(1):103–126, 2003.

D. Gale and L.S. Shapley. College admissions and the stability of marriage. *Amer. Math. Monthly*, 69(1):9–15, 1962.

A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

D. Gusfield and R.W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, 1989.

J.W. Hatfield and P.R. Milgrom. Matching with contracts. *Amer. Econ. Rev.*, 95(4):913–935, 2005.

D. Knuth. *Marriages Stables*. Les Presses de l'Universite de Montreal, 1976.

H. Moulin. On strategy proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

A. E. Roth and M. Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis*. Cambridge University Press, 1991.

A. E. Roth, T. Sönmez, and M. U. Ünver. Kidney exchange. *Q. J. Econ.*, 119(2):457–488, 2004.

M. Satterthwaite. Strategy-proofness and arrow's conditions. *J. Econ. Theor.*, 10(2):187–217, 1975.

J. Schummer and R.V. Vohra. Strategy-proof location on a network. *J. Economic Theory*, 104(2):405–428, 2004.

L.S. Shapley and H. Scarf. On cores and indivisibility. *J. Math. Econ.*, 1(1):23–28, 1974.

J. M. Stalnaker. The matching program for intern placement: The second year of operation. *J. Med. Educ.*, 28(1):13–19, 1953.

C. P. Teo and J. Sethuraman. Geometry of fractional stable matchings and its applications. *Math. Oper. Res.*, 23(4):874–891, 1998.

J. H. VandeVate. Linear programming brings marital bliss. *Oper. Res. Lett.*, 8(3):147–153, 1989.

Exercises

10.1 To what extent is Lemma 10.1 sensitive to the richness of the preference domain? For example, does the result hold if the preference domain is even smaller, e.g., containing only *symmetric* single-peaked preferences?

10.2 Suppose that an *anonymous* rule described in Theorem 10.2 has parameters $(y_m)_{m=1}^{n-1}$. Express this rule as a generalized median voter scheme with parameters $(\alpha_S)_{S\subseteq N}$.

10.3 Suppose that a rule f is *strategy-proof* and *onto*, but not necessarily *anonymous*. Fix the preferences of agents 2 through n, $(\succeq_2,\ldots,\succeq_n)$, and denote the outcomes obtainable by agent 1 as

$$O = f(\,\cdot\,, \succeq_2, \ldots, \succeq_n) = \{x \in [0,1] : \exists \succeq_1 \in \mathcal{R} \text{ s.t. } f(\succeq_1, \succeq_2, \ldots, \succeq_n)\}.$$

Show that $O = [a, b]$ for some $a, b \in [0, 1]$ (without appealing directly to Theorem 10.4).

10.4 Prove Theorem 10.4.

10.5 *For the case of three agents, generalize Theorem 10.2 to a 3-leaved tree.* Specifically, consider a connected noncyclic graph (i.e., a tree) with exactly three leaves, ℓ_1, ℓ_2, ℓ_3. Preferences over such a graph are single-peaked if there is a peak p_i such that for any x in the graph, and any y in the (unique shortest) path from x to p_i, $y \succeq_i x$. The concepts of *strategy-proofness*, *onto*, and *anonymity* generalize in the straightforward way to this setting. Describe all the rules that satisfy these conditions for the case $n = 3$. (Hint: first show that when all agents' peaks are restricted to the interval $[\ell_1, \ell_2]$, the rule must behave like one described in Theorem 10.2.) For the nonanonymous case with $n \geq 3$, see Schummer and Vohra (2004).

10.6 Prove that the TTCA returns an outcome in the core of the house allocation game.

10.7 The TTC mechanism is immune to agents misreporting their preferences. Is it immune to agents misreporting the identity of their houses? Specifically, suppose a subset of agents trade among themselves first before participating in the TTC mechanism. Can all of them be strictly better off by doing so?

10.8 Consider an instance of the stable matching problem. Let ν be a matching (not necessarily stable) and μ the male optimal stable matching. Let $B = \{m : \nu(m) >^m \mu(m)\}$. Show that if $B \neq \emptyset$ then there is a $m' \notin B$ and woman w such that (m, w) is a blocking pair for ν.