Econ. Theory 4, 417–435 (1994)

Economic Theory

Research articles

Top dominance and the possibility of strategy-proof stable solutions to matching problems*

José Alcalde and Salvador Barberà

Departament d'Economia i d'Historia Econòmica. Universitat Autònoma de Barcelona, E-08290 Bellaterra, SPAIN

Received: February 10, 1993; revised version June 2, 1993

Summary. This paper explores the possibility of designing strategy-proof mechanisms yielding satisfactory solutions to the marriage and to the college admissions problem. Our first result is negative. We prove that no strategy-proof mechanism can always choose marriages that are individually rational and Pareto efficient. This strengthens a result by Roth (1982) showing that strategy-proof mechanisms cannot always select stable marriages. The result also applies, a fortiori, to college admissions. Since finding difficulties with strategy-proofness is quite an expected result, we then address a second question which is classical within the incentives literature. Are there restrictions on the preferences of agents under which strategy-proof and stable mechanisms do exist? We identify a nontrivial restriction on the domain of preferences, to be called top dominance, under which there exist strategy-proof and stable mechanisms for both types of matching problems. The mechanisms turn out to be exactly those that derive from the most classical algorithms in the literature; namely, the women's optimal, the men's optimal and the student's optimal. Finally, top dominance is shown to be essentially necessary, as well as sufficient, for the existence of strategy-proof stable matching mechanisms.

1. Introduction

This paper explores the possibility of designing strategy-proof mechanisms yielding satisfactory allocations in two-sided matching markets.

Markets are two-sided if agents belong, from the outset, to one of two disjoint sets.[1] Matching markets are those where the exchange is bilateral and in fixed

* This work is partially supported by grant PB 89-0294, from the Dirección General de Investigación Ciencia y Tecnología of the Spanish Ministerio de Educación y Ciencia. Salvador Barberà is also grateful to the Instituto de Estudios Fiscales. This research was initiated while both authors were visting GREMAQ, Université des Sciencies Sociales, Toulouse, whose hospitality is gratefully acknowledged. The paper extends results that were circulated as GREMAQ W.P. 91.22.232. We are grateful to Matthew Jackson and Marilda Sotomayor for their comments.

[1] All informal definitions in this introduction are made precise later on in the text. Our purpose here is just to give a general feeling for the subject. For a masterful introduction to matching models we refer the reader to the book "Two-sided Matching: A Study in Game-Theoretic Modelling and Analysis", by Roth and Sotomayor (1991).

quantities. Some very important allocation problems are solved in markets which share exactly these two features. Some leading examples are the allocation of jobs, with firms and workers as the two sides of the market, or the admissions of new students to colleges. Allocations in these markets are matchings, assigning each agent in one side of the market the agent(s) in the other side with whom he or she will carry out the fixed quantity trade. Admittedly, not all markets are appropriately described in these terms. The role of agents as suppliers or demanders may not be clearly predetermined, multilateral trade may be important and the quantities to exchange may not be fixed. But even then, the models we analyze can often serve as useful first approximations.

Most of our results are presented for the simplest case, where each agent on one side of the market can be matched to at most one agent on the other side. This is colloquially known as the marriage market. The results also extend with some added assumptions to the case where agents on one of the sides can be matched with more than one in the other; this is the many-to-one case, also known as the college admissions model. Brief indications on that case are provided in section 6.

Given two sets of agents, representing the two sides of a potential market, different specific markets can arise depending on these agents' preferences. We concentrate on matching mechanisms; that is, rules to select one matching for each market that arises as the agents' preferences vary across some admissible class. More specifically, we look for mechanisms satisfying two desirable properties: stability and strategy-proofness.

A matching mechanism is stable if it picks one stable matching for each one of the markets on its domain. A matching is stable if it is individually rational and no pair of agents, one for each side, would rather be matched to each other than to their present match. Stability is clearly desirable as a necessary condition for the durability of the allocations resulting from a mechanism.

A mechanism is strategy-proof if it provides incentives for agents to reveal their true preferences while participating in the allocation process. Strategy-proofness is a guarantee that the results (and possible virtues) of the mechanism will not be biased through strategic misrepresentation of their preferences by some individual participants in the market.

Our first, preliminary result is negative. We prove that no strategy-proof mechanism can always choose marriages that are individually rational and Pareto efficient. The same impossibility applies, a fortiori, to the college admissions problem. This strengthens a result by Roth (1982), showing that strategy-proof mechanisms cannot always select stable marriages.

Since finding difficulties with strategy-proofness is quite an expected result, we then turn to the major question addressed by this paper. Are there restrictions on the preferences of agents under which strategy-proof and stable mechanisms do exist? Our approach is standard in the literature on incentives. For instance, the Clarke-Groves "solution" to the free rider problem consists in (1) identifying a class of preferences—additively separable preferences—under which strategy-proof and satisfactory mechanisms can be defined; (2) defining a family of procedures based on the pivotal mechanism having these properties on that preference domain; (3) proving that such procedures are in fact the only ones to satisfy both requirements

on the given domain, and (4) showing that the original, restricted class of admissible preferences is indeed the maximal class admitting any such solution.

A similar "solution" is given by Barberà, Sonneschein and Zhou (1991) to the problem of choosing sets of objects: voting by committees is shown to be the class of strategy-proof mechanisms respecting voter's sovereignty when individual preferences over sets are separable, and this restricted class of separable preferences is also proven to be maximal.

Other examples in the same vein are provided by Moulin's (1980) characterization of strategy-proof mechanisms that choose points on the real line, when agent's preferences are single-peaked; Sprumont's (1991) characterization of efficient, anonymous and strategy-proof mechanisms that choose on the simplex when each of the agent's preferences are single-peaked on one dimension of the simplex; and Barberà and Jackson's (1991) characterization of non-dictatorial strategy-proof mechanisms that choose on the n-dimensional plane, when individual preferences are satiated and convex.

In all of these cases, domain restrictions are imposed, and it is within these restrictions that one tries to identify strategy-proof mechanisms. In all these cases, the restrictions imposed on preferences are quite stringent, yet plausible enough under some interpretation as to deserve investigation. The very fact that an appropriate solution to an interesting problem can be found under some restrictions should make such restriction attractive. For example, single peakedness owes much of its popularity to the fact that it avoids the rather pervasive problem of majority cycles. But a restriction becomes especially important if we can show that only those domains satisfying it admit mechanisms that meet our demands. We then say that this restriction is necessary, even if this is not a necessity in the usual mathematical sense.

Our top dominance condition, to be described in what follows, meets these requirements. As we shall see, the only nontrivial domains admitting strategy-proof stable matching mechanisms are those satisfying top dominance.

Before presenting the condition, we discuss a number of examples: in each case we exhibit a class of preferences and describe market circumstances under which the preferences of agents on one of the sides of the market might be expected to belong to the class. Then we notice that all these classes share a common feature. Take any pair of preferences P, P' within the class, and any two individually rational alternatives x and y such that xPy while $yP'x$; then no other alternative z can be better than x under P and also better than y under P'. This is top dominance.

Our first main result establishes that when the preferences of men are restricted to satisfy top dominance, then the matching rule that always chooses the (uniquely defined) women's optimal matching is stable and strategy-proof. Symmetrically, when the preferences of women satisfy top dominance, then selecting the men's optimal matching defines a stable strategy-proof rule. In the college admissions problem, when colleges have responsive preferences on sets of students and their underlying preferences on individual students satisfy top dominance, then the student's optimal rule is also stable and strategy-proof.

The facts above only require to restrict the preferences of one side of the market. Our second result shows that, when the agents' preferences of one side of the market

are unrestricted, if a strategy-proof stable mechanism exists at all, it must always choose the stable allocation that is optimal for the unrestricted side of the market. As a corollary, when one side is restricted by top dominance and the other side is not restricted, then the women's (or the men's, or the students') optimal rule turns out to be the *unique* strategy-proof and stable procedure on this domain.

Top dominance is thus a sufficient condition, when imposed on one side of the market, for strategy-proof stable matching rules to exist. Moreover, the rules that emerge are those that the literature had already identified as most interesting from other points of view. Our third main result proves that top dominance is not only sufficient, but also essentially necessary in the sense described above, for a non-trivial restricted domain to admit strategy-proof stable matching procedures. This completes our full characterization result, parallel to the Clarke-Groves solution to the free-rider problem in economies with public goods.

The paper is organized as follows. Sections 2 to 5 refer to the marriage problem. After notation and definitions in Section 2, Section 3 contains our preliminary impossibility result. Top dominance is presented in Section 4. The main results are in Section 5. Section 6 briefly describes the college admissions problem and shows how to extend the previous analysis to that model. Section 7 concludes.

2. One-to-one matchings and matching rules

$M = \{m_1, m_2, \ldots, m_n\}$ is the set of *men*.

$W = \{w_1, w_2, \ldots, w_m\}$ is the set of *women*.

These sets are kept fixed throughout the paper. All ensuing definitions are relative to these sets. We assume that n and m are both greater than one.

Each m_i is endowed with a complete, transitive preference relation $P(m_i)$ on $W \cup \{m_i\}$[2]. For $w_j, w_h \in W$, $w_j P(m_i) w_h$ means that man m_i prefers woman w_j to woman w_h; $w_j P(m_i) m_i$ means that m_i would rather marry woman w_j than stay single, and $m_i P(m_i) w_h$ means that m_i would prefer to stay single rather than being married to w_h.

Similarly, each woman w_j is endowed with a complete transitive preference relation $P(w_j)$ on $M \cup \{w_j\}$. We denote by $\mathcal{P}(m_i)$ the set of all possible preferences for man m_i, and by $\mathcal{Q}(w_j)$ the set of all possible preferences for woman w_j.

Elements of $\prod_{i=1}^{n} \mathcal{P}(m_i) \times \prod_{j=1}^{m} \mathcal{Q}(w_j)$ are called *preference profiles*. We denote preference profiles by $(P(m_1), \ldots, P(m_n); P(w_1), \ldots, P(w_m))$, or by $\underline{P}$. $\mathcal{P}$ denotes the set of all possible preference profiles. Given a preference profile $\underline{P}$, we denote by $\underline{P}/P'(m_i)$ (resp. $(\underline{P}/P'(w_j))$ the profile obtained from $\underline{P}$ by changing the preferences of man m_i

[2] For simplicity, we assume that no agent is indifferent between any two potential mates, or between any possible mate and the no-marriage option. Our conclusions would not change if we allowed for indifferences.

(resp. woman w_j) from $P(m_i)$ to $P'(m_i)$ (resp. $P(w_j), P'(w_j)$), and keeping all other preferences unchanged.

A (one-to-one) *matching* is a function $\mu: M \cup W \to M \cup W$, such that

(1) $\quad (\mu(m_i) \notin W \Rightarrow \mu(m_i) = m_i), \quad \text{and} \quad (\mu(w_j) \notin M \Rightarrow \mu(w_j) = w_j), \quad \text{and}$

(2) $\quad \mu(m_i) = w_j \Leftrightarrow \mu(w_j) = m_i$

Condition (1) requires that each woman (resp. man) should either be assigned to a man (resp. woman) or to herself (resp. himself). The latter represents the possibility that some agents stay unmatched, or single. Condition (2) requires that if a man is assigned to a woman, then this woman should be assigned to that man.

Let M be the set of all possible matchings of women W with men M.

A matching μ is *individually rational* at preference profile $\underline{P}$ iff each agent which is assigned a partner considers it at least as good as staying single. Formally,

$$\mu(m_i) \in W \Rightarrow \mu(m_i) P(m_i) m_i \quad \text{for all } m_i \in M, \quad \text{and}$$

$$\mu(w_j) \in M \Rightarrow \mu(w_j) P(w_j) w_j \quad \text{for all } w_j \in W.$$

A matching μ is *efficient* at preference profile $\underline{P}$ iff there does not exist another matching $\mu' \neq \mu$ such that

$$[\mu'(m_i) \neq \mu(m_i)] \Rightarrow \mu'(m_i) P(m_i) \mu(m_i), \quad \text{and} \quad [\mu'(w_j) \neq \mu(w_j)] \Rightarrow \mu'(w_j) P(w_j) \mu(w_j).$$

A matching μ is *blocked by a pair* $(m_i, w_j) \in M \times W$ at profile $\underline{P}$ iff $w_j P(m_i) \mu(m_i)$ and $m_i P(w_j) \mu(w_j)$.

A matching μ is *stable* at profile $\underline{P}$ iff it is individually rational and it is not blocked by any pair in $M \times W$.

Let $\mathscr{T} \subseteq \mathscr{P}$. A matching rule on $\mathscr{T}$ is a function φ, $\varphi: \mathscr{T} \to M$.

Thus, a matching rule on $\mathscr{T}$ assigns one matching to each preference profile in $\mathscr{T} \subseteq \mathscr{P}$. We interpret $\mathscr{T}$ as the set of admissible preference profiles. In all that follows, we assume that $\mathscr{T}$ is a cartesian product

$$\mathscr{T}(m_1) \times \cdots \times \mathscr{T}(m_n) \times \mathscr{T}(w_1) \times \cdots \times \mathscr{T}(w_m).$$

A matching rule φ on T is *manipulable* by man m_i (resp. woman w_j) at profile $\underline{P} \in \mathscr{T}$ via $P'(m_i) \in \mathscr{T}(m_i)$ (resp. $P'(w_j) \in \mathscr{T}(w_j)$ iff $\varphi(\underline{P}/P'(m_i)) P(m_i) \varphi(\underline{P})$ (resp. $\varphi(\underline{P}/P'(w_j)) P(w_j) \varphi(\underline{P})$).

A matching rule φ is *strategy-proof* on $\mathscr{T}$ iff it cannot be manipulated at any profile in $\mathscr{T}$ by any man m_i or any woman w_j, via preferences in $\mathscr{T}(m_i)$, $\mathscr{T}(w_j)$.

A *matching rule is stable* (resp. *individually rational*, resp. *efficient*) iff its image at any profile $\underline{P} \in \mathscr{P}$ is stable (resp. individually rational, resp. efficient) at $\underline{P}$.

3. An impossibility result

The possibility of designing strategy-proof and stable rules for some special domains of preferences will be the main focus of this paper. This would be unnecessary if such possibility was wide open even when no domain restrictions held. Our first proposition, which extends a result by Roth (1982), proves that under the type of

standard preferences assumed for men and women in matching models[3] it is not even possible to achieve efficiency and individual rationality, which are milder requirements than stability, along with strategy-proofness.

Proposition 1
There exists no strategy-proof, efficient and individually rational matching rule on $\mathscr{P}$.

Proof
We prove the result for a society with two men and two women. The proof will be identical for larger societies: just endow two men and two women with the same preferences that we consider here, and let the preferences of all others be such that the rest of the assignment problem becomes trivial[4].

Thus, let $M = \{m_1, m_2\}$, $W = \{w_1, w_2\}$.

We assume that f is an efficient and individually rational matching rule. We shall prove that it must then be manipulable.

Consider first the following problems[5]

$$\text{(I)} \quad \left\{ \begin{array}{ll} P^1(m_1) = w_1 w_2 & P^1(w_1) = m_2 m_1 \\ P^1(m_2) = w_2 w_1 & P^1(w_2) = m_1 m_2 \end{array} \right\}$$

$$\text{(II)} \quad \left\{ \begin{array}{ll} P^2(m_1) = w_1 w_2 & P^2(w_1) = m_2 \\ P^2(m_2) = w_2 w_1 & P^2(w_2) = m_1 m_2 \end{array} \right\}$$

$$\text{(III)} \quad \left\{ \begin{array}{ll} P^3(m_1) = w_1 w_2 & P^3(w_1) = m_2 \\ P^3(m_2) = w_2 w_1 & P^3(w_2) = m_1 \end{array} \right\}$$

$$\text{(IV)} \quad \left\{ \begin{array}{ll} P^4(m_1) = w_1 & P^4(w_1) = m_2 m_1 \\ P^4(m_2) = w_2 w_1 & P^4(w_2) = m_1 m_2 \end{array} \right\}$$

$$\text{(V)} \quad \left\{ \begin{array}{ll} P^5(m_1) = w_1 & P^5(w_1) = m_2 m_1 \\ P^5(m_2) = w_2 & P^5(w_2) = m_1 m_2 \end{array} \right\}$$

The set of individually rational, efficient matchings for these preference profiles are the following[6]:

$$\text{(I)} \quad \text{i.r.e.} \quad \mu_1^1 : \left\{ \begin{array}{ll} m_1 & m_2 \\ w_1 & w_2 \end{array} \right\} \qquad \text{and} \quad \mu_2^1 : \left\{ \begin{array}{ll} m_1 & m_2 \\ w_2 & w_1 \end{array} \right\}$$

[3] Notice that the standard assumptions already impose domain restrictions: the preferences of each agent are defined over a different set of agents than those of any other; equivalently, if we look at their induced preferences on matchings (a common set of objects to be ranked), then feasible preferences for each agent are different than for any other, because of selfishness. Thus, our result cannot be obtained as a consequence of the celebrated impossibility theorem of Gibbard (1973) and Satterthwaite (1975).

[4] For example, by assuming that everybody else would prefer to stay single. Or by assuming for some set of t men and t women that each man prefers one of the women and that this woman prefers him too.

[5] We use the conventional notation. An ordered list of mates is assumed to indicate the agent's preferences from better to worse, among those mates that are preferred to remaining single. Potential mates that are worse than remaining single are not listed.

[6] Here again, we use standard notation. If a man and a woman appear on the same vertical, they are matched to each other. An agent with no mate on its vertical is unmatched.

(II) i.r.e. $\mu_1^2: \begin{Bmatrix} m_1 & m_2 & - \\ - & w_2 & w_1 \end{Bmatrix}$ and $\mu_2^2: \begin{Bmatrix} m_1 & m_2 \\ w_2 & w_1 \end{Bmatrix}$

(III) i.r.e. $\mu^3: \begin{Bmatrix} m_1 & m_2 \\ w_2 & w_1 \end{Bmatrix}$

(IV) i.r.e. $\mu_1^4: \begin{Bmatrix} m_1 & m_2 \\ w_1 & w_2 \end{Bmatrix}$ and $\mu_2^4: \begin{Bmatrix} m_1 & m_2 & - \\ - & w_1 & w_2 \end{Bmatrix}$

(V) i.r.e. $\mu^5: \begin{Bmatrix} m_1 & m_2 \\ w_1 & w_2 \end{Bmatrix}$

Suppose $f(\underset{\sim}{P}^1) = \mu_1^1$. Then if $f(\underset{\sim}{P}^2) = \mu_2^2$, w_1 can manipulate f at $\underset{\sim}{P}^1$ via $P^2(w_1)$. If $f(\underset{\sim}{P}^2) = \mu_1^2$, then w_2 can manipulate at $\underset{\sim}{P}^2$ via $P^3(w_2)$.

A parallel argument works if $f(\underset{\sim}{P}^1) = \mu_2^1$. Then, if $f(\underset{\sim}{P}^4) = \mu_1^4$, m_1 can manipulate f at $\underset{\sim}{P}^1$ via $P^4(m_1)$. If $f(\underset{\sim}{P}^4) = \mu_2^4$, then m_2 can manipulate f at $\underset{\sim}{P}^4$ via $P^5(m_2)$.

This exhausts all possible choices of individually rational, efficient matching for f to take values at these profiles, and shows that f will necessarily be manipulable.

Remark 1

Our result extends Roth's proof that all stable matching rules are manipulable by relaxing stability to the weaker requirement that matchings should be individually rational and efficient. Notice that our arguments have considered certain matchings that satisfy our conditions but are not stable. For example, μ_1^2 and μ_2^4 are not stable at $\underset{\sim}{P}^2$ and $\underset{\sim}{P}^4$, respectively.

4. Top dominance: a new restriction on preferences

Our previous result has shown that strategy-proof rules will be extremely unsatisfactory if they operate on the set of all profiles that can be formed with selfish preferences.

We now want to investigate whether there exist nontrivial domain restrictions under which satisfactory matching rules can be found. We shall exhibit a class of preference domains and stable matching rules which are strategy-proof on these domains. More specifically, we'll show that the rules which obtain by application of the classical M-optimal or W-optimal algorithms are indeed strategy-proof in domains within our class[7].

Before we discuss the classes of preferences admitting a solution to our problem, let us comment on a rather natural preference restriction: that all men and all women prefer any possible mate to remaining single.

Our previous result is no longer true under that restriction, because the condition of individual rationality looses all its bite. Here is an efficient, individually rational and strategy-proof matching rule when all men and all women like all their potential

[7] We remind the reader that the "marriage" interpretation is a metaphor. There will be uses of the model where the restriction is natural, and others where it is not.

mates: rank individuals on one side of the market, and let them pick their favourite mate among those not already chosen by people of higher rank.

What about stable and strategy-proof rules when everybody prefers all potential mates to remaining single? For the case with two men and two women, the M-optimal and the W-optimal rule are examples of strategy-proof, stable matching rules under that domain restriction. Unfortunately, the result does not generalize. Roth's negative result still applies when there are at least three men and three women. In fact Roth's (1982) proof only uses preferences within this class, and builds upon the three men and three women case.

The above discussion warns us that domains admitting stable and strategy-proof rules will be quite restrictive. We now proceed to identify a family of such domains. Preferences of either men or women will be restricted, those of the other side of the market may remain free. Let us begin with some examples that will fit into our formal conditions. For ease of interpretation, we refer to the case where one side of the market consists of a set of potential employers, and the other side is a set of workers[8].

Example 1

Assume that workers can be classified by level of qualification (nobody being equal than anyone else), and that differences in these levels have cardinal meaning. Assume that the preferences of employers are formed as follows: they define an ideal level of qualification and rank any two workers according to the inverse of their distance to the ideal degree of qualification.

Figure 1 represents two preference orderings that conform to this description. Sets of preferences of this type do respect the top dominance condition. Notice that

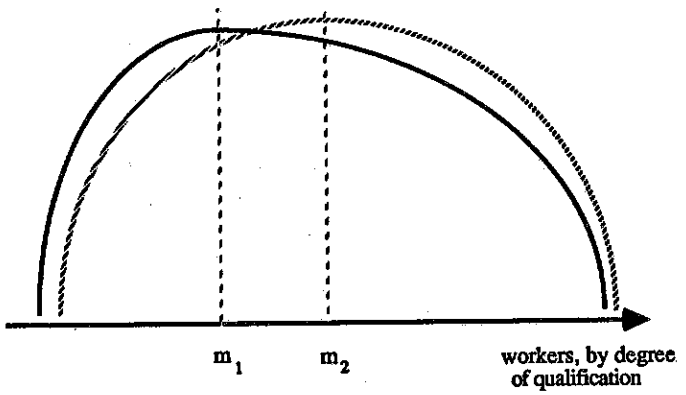

Figure 1.

[8] Implicit in this interpretation are two assumptions: (1) salaries are exogenously fixed and (2) each employer has only one opening available. The latter is relaxed in the college admissions problem considered in Section 6.

in fact they constitute a set of single-peaked preferences. However, the restriction is stronger here: once the distances in qualification among workers are fixed, there is only one admissible preference for each possible ideal.

Example 2

Assume again that workers are ranked by degree of qualification (nobody being equal to any one else). Assume that the preferences of employers are formed as follows: (1) they define an ideal level of qualification and prefer an employee at this level to any other; (2) they prefer anybody who is overqualified to anybody underqualified; (3) within each of these two groups, they prefer workers who are closer to the ideal.

Figure 2 represents two preference orderings that conform to this description. Sets of preferences within this type will also satisfy the top dominance condition.

Notice that this time the restriction is purely ordinal. In view of the graphical representation, these preferences may seem less attractive than those in Example 1, because of some apparent "discontinuity" in their treatment of alternatives. We do not attach much importance to this apparent feature, because the model in hand is essentially discrete and ordinal.

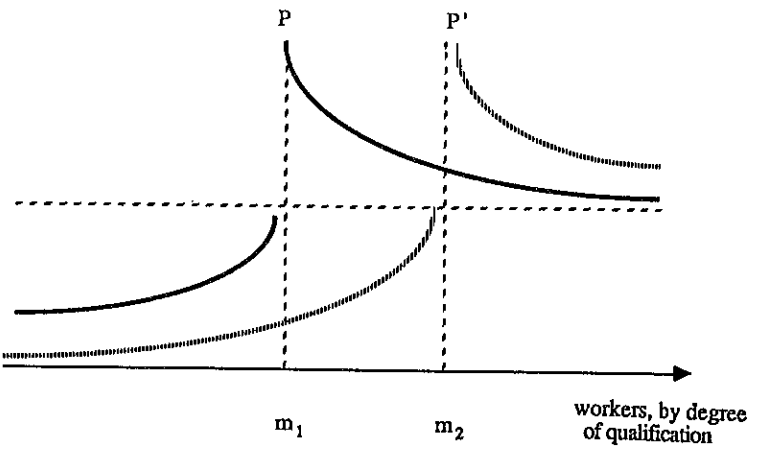

Figure 2.

Example 3

Suppose each of n workers can be defined by n different numbers. These numbers are ordered around a circle, and the description of each worker is obtained by choosing one of these numbers as the first component, followed by the others in clockwise order. We could interpret that each of these workers, say $(m_1, m_2, \ldots, m_n)$, is fully described by the levels m_i at which he or she holds characteristic i.

Assume now that each employer forms his or her preferences over these workers according to the level of just one of these characteristics, the only one that matters to him or her.

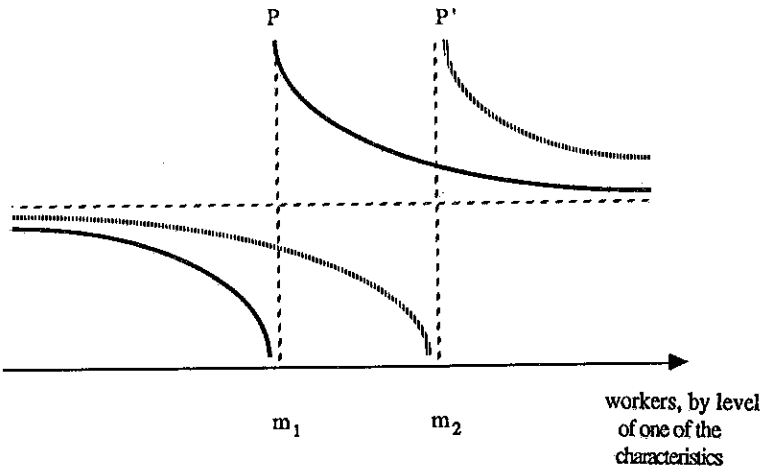

Figure 3.

We can then describe the preferences of employers as follows. For an appropriate choice of position, we can arrange workers around a circle; once we know an employer's preferred worker m^*, he or she will rank the rest of them in clockwise order, starting from m^*. Another equivalent description of the class of employers' preferences fitting the above description is given by Figure 3. Again, classes of this type will satisfy the top dominance condition.

Example 4

Suppose that all workers are perceived by all employers in the same basic order. Yet each employer is entitled to have an absolute favourite, while the rest of workers will be ranked according to the basic order.

Figure 4 represents two possible preferences in a class that fits the above description. Again, a class of preferences generated in this way will satisfy the top dominance condition.

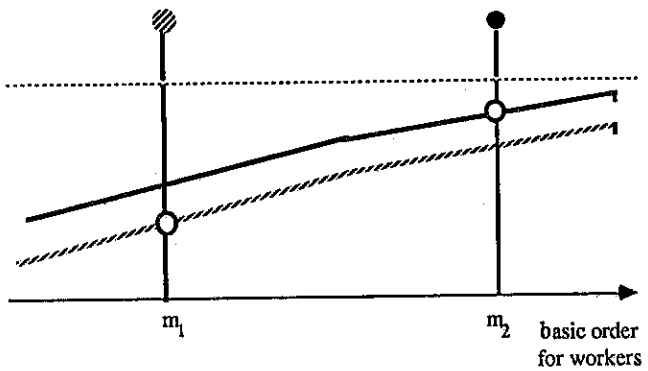

Figure 4.

Example 5
Suppose each employer is known to like one and only one of the workers. He or she would rather employ that particular worker, but if that was impossible he or she would prefer not to employ anybody. Any worker can be chosen for this role by any employer.

Once again, this class of preferences will satisfy top dominance.

We now proceed with the formal definition of our domain restriction.

Definition
Let $\mathscr{P}$ be a set of preferences (on $M \cup \{w\}$ or in $W \cup \{m\}$). $\mathscr{P}$ satisfies *the top dominance condition* iff for any pair of preferences P and P' in $\mathscr{P}$, and any x, y such that:

$$\left.\begin{array}{l} x \text{ is individually rational for } P \\ y \text{ is individually rational for } P' \\ xPy \text{ and } yP'x \end{array}\right\} \quad \text{then, there is no } z \text{ for which } zPx \text{ and } zP'y.$$

The reader may check that the classes of preferences in examples 1 to 5 satisfy top dominance.

Remark 2
It is clear from the definition that two different preferences in the same class cannot have the same maximal element. The top choice determines the rest of the preference: hence our name for the condition.

5. Stable and strategy-proof rules: possibility and characterization results

We can now state and prove results about the possibility of strategy-proof and stable rules when one side of the market satisfies top dominance. First, we show the condition to be sufficient for the existence of strategy-proof rules on marriage markets. Second, we prove that, if the preferences for the other side of the market are completely unrestricted, then there is only one strategy-proof rule, and it is the one choosing the (unique) optimal stable outcome for this unrestricted side. All this is nice and well, but one can still feel uneasy about top dominance: isn't it too restrictive a condition?. Our third result shows that top dominance is essentially necessary (in a sense to be made precise) as well as sufficient, for a strategy-proof stable solution to the marriage problem to exist.

Theorem 2
Let $\mathscr{T}$ be a family of preferences for sets M and W of men and women. If the set of admissible preferences, $\mathscr{T}(m_i)$, for every man, (resp. $\mathscr{T}(w_j)$ for every woman), satisfy the top dominance condition, then there exist stable and strategy-proof matching rules on $\mathscr{T}$.

Proof
We prove the theorem under the assumption that the women's preferences all satisfy top dominance. The proof consists in showing that the matching rule f resulting from applying the men's optimal algorithm will be stable and strategy-proof under

these circumstances. Should the men's preferences be the ones to satisfy the restrictions, we could then apply an identical reasoning to the women's optimal algorithm.

It is well known that declaring their true preferences is a dominant strategy for men under the men's optimal algorithm (see, for example, Theorem 4.7 in Roth and Sotomayor (1991)). Thus, it suffices to show that women holding our restricted preferences will not have any incentive to misrepresent their preferences at any profile.

In order to prove it, let's take a closer look at the algorithm. For any preference profile $\underline{P}$, let $M_j^t(\underline{P})$ be the set of all men that have proposed to woman w_j along the application of the men's algorithm at profile $\underline{P}$, up to the t-th iteration, and let $f_j^t(\underline{P})$ be w_j's mate at this t-th iteration. Notice that at the t-th iteration, woman w_j is tentatively matched to the man she considers best among those who have already proposed her, or else is single, if she has received no proposition yet. That is, $f_j^t(\underline{P}) = C(P(w_j), M_j^t(\underline{P}) \cup \{w_j\})$. Because of that, it will always be the case that $f_j(\underline{P}) P(w_j) f_j^t(\underline{P})$, or else $f(\underline{P}) = f_j^t(\underline{P})$. When the latter holds for some $\bar{t}$, it also holds for any $t > \bar{t}$.

Suppose that w_j could manipulate at profile $\underline{P}$ by declaring $P'(w_j)$. Then, declaring $P'(w_j)$ rather than $P(w_j)$ must make a difference at some point, during the application of the men's algorithm. Specifically, there must exist a k such that

$$M_j^t(\underline{P}) = M_j^t(\underline{P}/P'(w_j)) \quad \text{for } t = 1, 2, \ldots, k, \text{ and yet}$$

$$x \equiv f_j^k(\underline{P}) = C(P(w_j), M_j^k(\underline{P}) \cup \{w_j\}) \neq C(P'(w_j), M_j^k(\underline{P}) \cup \{w_j\}) = f_j^k(\underline{P}/P'(w_j)) \equiv y.$$

Now, if f is manipulable, we should have that

$$z \equiv f_j(\underline{P}/P'(w_j)) P(w_j) f_j(\underline{P}) P(w_j) f_j^k(\underline{P}) P(w_j) f_j^k(\underline{P}, P'(w_j)).$$

But we also have

$$z \equiv f_j(\underline{P}/P'(w_j)) P'(w_j) f_j^k(\underline{P}, P'(w_j)) P'(w_j) f_j^k(\underline{P}).$$

This is a contradiction, since $P(w_j)$ and $P'(w_j)$ belong to the same class satisfying top dominance, $xP(w_j)y$ and $yP'(w_j)x$, and yet we claim that both $zP(w_j)x$ and $zP'(w_j)y$. Therefore, w_j cannot manipulate f at any $\underline{P}$ by means of any $P'(w_j)$, as we wanted to show.

Our existence proof is based on the fact that one side is restricted to satisfy top dominance, and places no specific requirements on the other. Then we show that the rule which attributes its best stable outcome to the side on which no specific restrictions are placed is indeed stable and strategy-proof. The next results go one step further: if the agents on the unrestricted side can indeed have any of the possible preferences, then the rule we have located is in fact the only stable and strategy-proof rule one can use.

Theorem 3

Let $\mathcal{T}$ be a family of preferences for sets M and W of men and women. If the preferences of every man (resp. of every woman) are unrestricted and there is a stable and strategy-proof matching rule on $\mathcal{T}$, then this rule must always choose the men's (resp. the women's) optimal stable matching.

Proof

In order to prove Theorem 3, we assume that the men's preferences are unrestricted. We'll show that any stable rule, ρ, that does not always choose the matching resulting from the men's optimal algorithm will be manipulable.

Assume that $\rho(\underline{P})$ was not the men's optimal matching at $\underline{P}$ in $\mathscr{T}$. There will then be some $m_i \in M$ who is not married with his M-optimal mate, $f_i(\underline{P})$. Because of the latticial structure of the set of stable matchings, we have that

$$f_i(\underline{P})P(m_i)\rho_i(\underline{P}).$$

Consider now the following preferences for m_i

(i) $\forall w, w' \in W \quad wP(m_i)w' \Rightarrow wP'(m_i)w'$

(ii) $\forall w \in W \setminus \{f_i(\underline{P})\} \quad wP(m_i)f_i(\underline{P}) \Leftrightarrow wP'(m_i)m_i$ and

(iii) $f_i(\underline{P})P'(m_i)m_i$

In words: $P'(m_i)$ keeps the same ranking among women than $P(m_i)$. However, the only women that m_i would marry under $P'(m_i)$ are those equal or above $f_i(\underline{P})$ under $P(m_i)$.

Clearly, the men's optimal matching at $\underline{P}$ will still be stable at $\underline{P} \setminus P'(m_i)$. By theorem 2.22 of Roth and Sotomayor (1990), we know that the set of people remaining single is the same at all the stable matchings for a given profile. Thus since m_i is married to $f_i(\underline{P})$ at one stable matching for $\underline{P}/P'(m_i)$, it will be married in $\rho(\underline{P}/P'(m_i))$, and we have that $\rho(\underline{P}/P'(m_i))P'(m_i)m_i$. Then m_i can manipulate ρ at $\underline{P}$ via $P'(m_i)$, because $\rho(\underline{P}/P'(m_i))P'(m_i)m_i$ implies $\rho(\underline{P}/P'(m_i))P(m_i)\rho(\underline{P})$.

A symmetric argument can be used when the top dominance condition applies to the set M and we assume that the women's optimal matching is not always the image of a stable mechanism.

Corollary

Let $\mathscr{T}$ be a family of preferences for sets M and W of men and women. If the preferences of every man (resp. of every woman) satisfy the top dominance condition, and the preferences of every woman (resp. of every man) are unrestricted, then the matching rule obtained by application of the women's (resp. the men's) optimal algorithm is the *unique* stable and strategy-proof rule on $\mathscr{T}$.

Proof

To prove the corollary, just notice that we have not used any property of the women's preferences (if men are unrestricted) in theorem 3. Therefore, the result still holds true when the preferences of women are restricted to satisfy top dominance.

Having one side of the market satisfy top dominance and then using the rule that is optimal for the unrestricted side does avoid manipulation. We want to show more: in an essential way, top dominance is not only a sufficient but also a necessary condition for this possibility to hold.

Before we do this, we must elaborate a bit on our additional assumptions and motivation. Necessity cannot be a property of any domain restriction in the mathematical sense. For, suppose we claimed that for a stable solution to be strategy-proof,

agent i's preferences must belong to some set $\mathcal{T}(i)$. Since $\mathcal{T}(i)$ is not the set of all conceivable preferences, there must be some preference $P(i) \notin \mathcal{T}(i)$ for agent i. Now, restrict preferences for agents to be such that agent i is only allowed to have preference $P(i)$ (this can be interpreted as a way to say that i is a dummy). Then, no argument will ever prove that i's preferences being not in $\mathcal{T}(i)$ destroys strategy-proofness. Thus, belonging to $\mathcal{T}(i)$ cannot be an unqualified necessity, for any $\mathcal{T}(i)$.

What we want is to describe classes of domains that are not trivial like the one above, in that they contain rich enough sets of preferences to be admissible for each agent, and also allow everybody on the restricted side of the market the same opportunity to express some preference. [For example, if an agent on one side is restricted to be single-peaked, then all others on this same side are also assumed to have single-peaked preferences]. Then, we'll prove that the only domains within this class admitting strategy-proof stable mechanisms when one side of the market is unrestricted are those where the restricted side satisfies top dominance. This is our necessity result.

Definition

Let $\mathcal{T}$ be a set of preferences for the agents on a marriage market. We say that $\mathcal{T}$ is a *rich* domain for one side of the market, say the men, iff

(i) For each man, m_i, and each woman, w_j, the set $\mathcal{T}(m_i)$ of m_i's admissible preferences contains a preference $P(m_i)$ where w_j is most preferred; and

(ii) For each man m_i, and any two women w_j, w_k, if for some preference $P(m_i)$ in $\mathcal{T}(m_i)$, we have that $w_j P(m_i) w_k P(m_i) m_i$, then there is another admissible preference $P'(m_i)$ in $\mathcal{T}(m_i)$ for which $w_k P'(m_i) w_j P'(m_i) m_i$.

Let $\mathcal{T}$ be a set of preferences for the agents on a marriage market. We say that $\mathcal{T}$ is an *anonymous* domain for one side of the market, say the men if, for any two men m_i and m_h, if $P(m_i) \in \mathcal{T}(m_i)$, then there is $P(m_h)$ in $\mathcal{T}(m_h)$ such that:

(a) $\forall w, w' \in W \quad w P(m_i) w' \Rightarrow w P(m_h) w'$; and

(b) $\forall w \in W \quad w P(m_i) m_i \Leftrightarrow w P(m_h) m_h$

The reader can check that these conditions are satisfied by any set of preferences defined by standard domain restrictions. For example, they would be met if agents on one side of the market were required to have single-peaked preferences, or if they were assumed to prefer being matched to somebody on the other side rather than remaining unmatched.

Notice that our definition of richness rules out trivial domains, like those containing only one preference. Yet, it is not exceedingly demanding. This is natural, since for "too large" domains we would fall back to impossibility results.

We can now state our necessity result.

Theorem 4

Let $\mathcal{T}$ be a family of preferences for sets M and W of men and women. Suppose that the preferences in $\mathcal{T}$ are unrestricted for one side of the market, and they respect an anonymous and rich domain restriction on the other side. If there exists a stable and strategy-proof matching rule on $\mathcal{T}$, then the set of preferences for restricted agents must satisfy the top dominance condition.

Proof

Without loss of generality, suppose men are the restricted side of the market, and assume that their sets of admissible preferences, $\mathcal{T}(m_i)$ do not satisfy top dominance. This would mean that we can find preferences $P(m_1)$ and $P'(m_1)$ for man m_1, say, such that, for some $x, y, z \in W \cup \{m_1\}$,

$$\begin{cases} xP(m_1)y \\ yP'(m_1)x \\ x \text{ is individually rational for } P(m_1) \\ y \text{ is individually rational for } P'(m_1) \\ zP(m_1)x \text{ and } zP'(m_1)y \end{cases}$$

Since z is better than some other individual rational option for both preferences, $P(m_1)$ and $P'(m_1)$, z cannot be m_1. Thus, we have $z \in W$. We can distinguish two cases. One, where either x or y equal m_1. The other, where both x and y are also women.

Without loss of generality, let these two cases be:

$$\text{Case 1:} \quad z = w_1, \; x = w_2, \; y = w_3$$
$$\text{Case 2:} \quad z = w_1, \; x = w_2, \; y = m_1$$

If case 1 holds, consider a profile $\underline{P}$ of preferences for 3 men and 3 women satisfying the following specifications:

$$\left.\begin{array}{ll} m_1\colon & w_1P(m_1)w_2P(m_1)w_3P(m_1)m_1 \\ m_2\colon & \ldots w_2P(m_2)\ldots w_1P(m_2)\ldots m_2P(m_2)\ldots \\ m_3\colon & w_3P(m_3)\ldots \\ w_1\colon & m_2P(w_1)m_1P(w_1)\ldots \\ w_2\colon & m_1P(w_2)m_2P(w_2)\ldots \\ w_3\colon & m_1P(w_3)m_3P(w_3)\ldots \end{array}\right\}$$

[Dots indicate possible positions for the unmentioned individuals. These positions are left unspecified because they are irrelevant to our argument].

This preference profile must be admissible. This is because (1) we have no restrictions in choosing preferences for women; (2) we attribute to m_1 one of the admissible preferences that appear in the statement defining necessity; (3) there must exist admissible preferences for m_2 where the order of w_1 and w_2 is reversed with respect to $P(m_1)$, by richness and anonymity; and (4) there must be some admissible preference where m_3 places w_3 above being single (it suffices to take the preferences of m_1 and m_3 to be equivalent). Since the preferences of women are unrestricted, Theorem 3 tells us that the only possible candidate function to be strategy-proof is the one selecting the women's optimal stable matching.

This, in our case, would result in the following matching for profile $\underline{P}$:

$$f_w(\underline{P}) = \mu\colon \begin{Bmatrix} m_1 & m_2 & m_3 \\ w_2 & w_1 & w_3 \end{Bmatrix}$$

Now, consider the profile $\underline{P}\backslash P'(m_1)$, constructed by changing man m_1 preferences to $P'(m_1)$, again an admissible preference for this agent. The reader can check that

the women's optimal matching becomes now:

$$f_w(\underline{P}\backslash P'(m_1)) = \mu' : \begin{Bmatrix} m_1 & m_2 & m_3 \\ w_1 & w_2 & w_3 \end{Bmatrix}.$$

Clearly, man m_1 gains by manipulating the outcome from μ to μ', by declaring preferences $P'(m_1)$ when his true preferences are $P(m_1)$. This contradicts the possibility that f_w might be strategy-proof in domains where top dominance is violated as in case 1.

We go now to case 2. Here, consider the 2-men, 2-women profile $\underline{P}$ given by:

$$\left.\begin{array}{ll} m_1: & w_1 P(m_1) w_2 P(m_1) m_1 \\ m_2: & w_2 P(m_2) w_1 P(m_2) m_2 \\ w_1: & m_2 P(w_1) m_1 P(w_1) w_1 \\ w_2: & m_1 P(w_2) m_2 P(w_2) w_2 \end{array}\right\}$$

This preference profile is clearly admissible. Its associated women's optimal matching is:

$f_w(\underline{P}) = \mu : \begin{Bmatrix} m_1 & m_2 \\ w_2 & w_1 \end{Bmatrix}$, whereas the preference profile $\underline{P}\backslash P'(m_1)$ will lead to $f_w(\underline{P}) = \mu : \begin{Bmatrix} m_1 & m_2 \\ w_1 & w_2 \end{Bmatrix}$, and again m_1 will be in a position to manipulate, a contradiction.

To complete the proof of theorem 4, let us remark that our constructs for 2 and 3 agents on each side of the market do not only apply to the case where this is actually the number of agents. For any larger number on any side, one can easily construct markets where any stable matching would be unambiguously defined for the rest of the market, and where all relevant changes would occur on the 2 or 3 agents for each side that we have concentrated upon in our proof.

6. An extension to the college admissions problem

In this section we briefly describe the extension of our results to many-to-one matching markets. These are particularly suited to describe the assignment of individuals to institutions: students and colleges, workers and firms, etc. One can show that, under appropriate conditions, stable mechanisms for many-to-one matching markets can be identified with stable mechanisms for an associated one-to-one market.

We simply present the basics of the college admissions problem and discuss the connections with the marriage problem that allow us to extend our previous result to this larger setting[9].

Let $C = \{c_1, c_2, \ldots, c_n\}$, be the set of *colleges*, and
$S = \{s_1, s_2, \ldots, s_m\}$, be the set of *students*..

[9] We follow very closely the presentation by Roth and Sotomayor (1991). Given the expository character of this section, we keep formalism to a minimum. For a more detailed presentation, we refer the reader to chapter 5 of the above mentioned book.

Each college has preferences over students, and students have preferences over colleges. For each college c, there is a positive integer, q_c called the quota of college c, which indicates the maximum number of positions it may fill.

An outcome of the college admissions model is a matching of students to colleges, such that each student is matched to at most one college and each college is matched to at most its quota of students.

A matching is stable if it is individually rational for all agents, and no college-student pair would prefer to be matched to each other rather than to some of their present mates. Again, stability is a condition for the durability of matchings. Even if its definition only involves bilateral relations, stable arrangements can also be shown to be robust against more complex, multilateral arrangements (Lemma 5.5 in Roth and Sotomayor (1991)).

We start by assuming that the preferences of colleges are described by a ranking of individual students. But matchings will assign sets of students to colleges: thus preferences over sets are also important. Some consistency between the preferences over students and the preferences over sets of students is required, if we want some of the main results for one-to-one matchings (including the existence of stable matchings), to hold in the many-to-one case. We'll make the following assumption of responsiveness.

Let P be the preferences of agents over individual students, and $P^{\#}$ the preferences of this same college over sets of students. The preferences of a college c over sets of students will be called *responsive* to its preferences over individual students if, for any two assignments that differ in only one student, it prefers the assignment containing the most preferred student (and is indifferent between them if it is indifferent between the students).

Given a particular college admissions problem, we can consider a related marriage market, in which each college c with quota q_c is broken into q_c "pieces" of itself, so that in the related market, the agents will be students and college positions, each having a quota of one: We replace college c with quota q_c by $c_1, c_2, \ldots, c_{q_c}$. Each of these positions has preferences over individuals that are identical to those of c. Since each position has a quota of one, we do not need to consider its preferences over groups of students. The student's preferences for this marriage problem will be derived as follows: each student's list is modified by replacing c, wherever it appears in his or her list, by the string $c_1, c_2, \ldots, c_{q_c}$, in that order. If the preferences over individuals are strict (an assumption we have made throughout the paper), there is a natural one-to-one correspondence between matchings in the original college admissions problem and matchings in the marriage market derived from it in this way. What's essential for us is that a matching of the college admissions problem is stable if, and only if, the corresponding matching of the related marriage problem is also stable. (This is shown for strict preferences by Lemma 5.1 of Roth and Sotomayor (1991)).

As for strategy-proofness, the major result we build upon is the following: a stable matching procedure that yields the students' optimal stable matching makes it a dominant strategy for all students to state their true preferences (Roth and Sotomayor (1991), Th, 5.16). Notice, however, that no parallel statement holds true regarding procedures that would favour colleges rather than students.

Since the marriage problem is a special case of the college admissions problem, all negative results for the particular case apply to the general framework. What about possibility results?. Notice that the situation is asymmetric now, since only the students' optimal matching mechanism will make it a dominant strategy for this side of the market to declare the truth, even if preferences are unrestricted. Therefore, we shall impose our domain restrictions on the preferences of colleges.

We say that a class of preferences for colleges satisfies top dominance if the admissible rankings of students by colleges satisfy the top dominance condition, as defined in section 4. In general, this definition could be unsatisfactory, since it has no direct bearing on the way how colleges rank sets of students. However, we shall only use it in association with the assumption that colleges' preferences over sets of students are responsive to their preferences over individuals.

We can now state our results for the college admissions problem.

Theorem 5

Let $\mathscr{T}$ be a family of preferences for sets C and S of colleges and students. If the preferences of every college over sets of students are responsive to their preferences over individuals and satisfy the top dominance condition, then, there exist stable and strategy-proof matching rules on $\mathscr{T}$. Moreover, if the preferences of every student are unrestricted, then the matching rule obtained by application of the student's optimal algorithm is the *unique* stable and strategy-proof rule on $\mathscr{T}$.

This theorem follows from all our preceding remarks and Theorems 2 and 3 in Section 5, and we leave its proof to the reader.

7. Conclusion

Since requiring the sets of admissible preferences for one side of the market to satisfy top dominance is a strong condition, each reader may want to decide whether to look at our results as positive or negative. But their practical implications are clear. Whenever designing a mechanism to solve a specific type of matching problem, it is well worth analyzing whether the preferences of one side of the market may reasonably be expected to satisfy restrictions resembling top dominance. If so (and this will of course depend on the problem at hand), then you have very good reasons to choose the available stable and strategy-proof rules. Otherwise, these two criteria will not help. In particular, whatever stable algorithm you choose will be manipulable. This may or may not have bad consequences. For some markets, you may still be able to choose a mechanism implementing a nice selection of matchings for a relevant equilibrium concept[10]. At any rate, we think it's good to know, from our results, exactly when you should and when you shouldn't worry about strategic distortions in the design of matching rules.

[10] Some results in this topic are analyzed in Alcalde (1992).

References

Alcalde, J.: Implementing stable solutions to the marriage problem. Mimeographed, Barcelona 1992
Barberà, S., Jackson, M.: A characterization of strategy-proof social choice functions for economies with pure public goods. Soc. Choice Welfare **11** (1994) forthcoming
Barberà, S., Sonneschein, H., Zhou, L.: Voting by committees. Econometrica **59**, 595–609 (1991)
Gale, D., Shapley, L.: College admissions and the stability of marriage. Am. Math. Monthly **69**, 9–15 (1962)
Gibbard, A.: Manipulation of voting schemes: a general result. Econometrica **41**, 587–601 (1973)
Green, J., Laffont, J.-J.: Incentives in public decision-making. North Holland: Amsterdam 1979
Moulin, H.: On strategy-proofness and single-peakedness. Public Choice **35**, 437–56 (1980)
Roth, A. E.: The economics of matching: stability and incentives. Math. Operat. Res. **7**, 617–628 (1982)
Roth, A. E., Sotomayor, M.: Two-sided matching: a study in game-theoretic modeling and analysis. Econometric Society Monograph Series. New York: Cambridge University Press 1990
Satterthwaite, M. A.: Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J. Econ. Theory **10**, 187–217 (1975)
Sprumont, Y.: The division problem with single-peaked preferences: a characterization of the uniform allocation rule. Econometrica **59**, 509–519 (1991)