MATHEMATICS OF OPERATIONS RESEARCH
Vol. 7, No. 4, November 1982
*Printed in U.S.A.*

# THE ECONOMICS OF MATCHING: STABILITY AND INCENTIVES*†

ALVIN E. ROTH

*University of Illinois*

This paper considers some game-theoretic aspects of matching problems and procedures, of the sort which involve matching the members of one group of agents with one or more members of a second, disjoint group of agents, all of whom have preferences over the possible resulting matches. The main focus of this paper is on determining the extent to which matching procedures can be designed which give agents the incentive to honestly reveal their preferences, and which produce stable matches.

Two principal results are demonstrated. The first is that no matching procedure exists which always yields a stable outcome and gives players the incentive to reveal their true preferences, even though procedures exist which accomplish either of these goals separately. The second result is that matching procedures do exist, however, which always yield a stable outcome and which always give all the agents in *one* of the two disjoint sets of agents the incentive to reveal their true preferences.

**1. Introduction.** The purpose of this paper is to explore the underlying economic structure common to matching problems and procedures. By *matching problems*, I refer to any of the pervasive class of problems which involve matching the members of one group of agents with one or more members of a second, disjoint group of agents, all of whom have preferences over the possible resulting matches. Thus the problems arising from the need to match, e.g., students with educational institutions, athletes with teams, adoptive children with adoptive parents, men with women (in marriage, mixed doubles, or computer dating), civil servants with civil service positions, and authors (via their papers) with scholarly journals, are all, in many of their most important aspects, matching problems.[1]

By *matching procedures*, I refer to the institutional arrangements by which the matching is accomplished. These institutional arrangements may range from completely decentralized procedures, in which agents negotiate directly with one another (as in marriage in contemporary Western societies), all the way to completely centralized procedures, in which all agents state their preferences for possible matches, which

*Received April 6, 1981; revised September 9, 1981.
*AMS 1980 subject classification.* Primary: 90D45.
*OR/MS Index 1978 subject classification.* Primary: 231 Games/group decisions.
*Key words.* Matching; assignment; preference revelation.
†This work has been supported by NSF grant SOC 78-09928 to the University of Illinois. It is also a pleasure to acknowledge stimulating conversations and correspondence on this subject with Lester Dubins, David Gale, Zvi Ritz, and Lloyd Shapley.

[1]The requirement that there be two disjoint (nonintersecting) sets of agents excludes from the class of problems under consideration here those in which there is only one set of agents, who are to be matched to one another; e.g., the problem of matching roommates, bridge partners, etc. The requirement that both sets of agents have active preferences over possible matches will exclude simple assignment problems, such as matching students with dormitories, where the dormitories have no preferences over students. And we will be concentrating on those problems in which the preferences of the participants are related to the matching itself, rather than to other features of the outcome. (Thus we consider matching civil servants to civil service positions, where the wage is part of the job description, rather than considering the general labor market and the problem of matching employees and employers, with the wage to be determined as part of the outcome.) These restrictions will be made precise in the formal model, and relaxation of some of these restrictions will also be discussed.



0364-765X/82/0704/0617$01.25

are then assigned according to some specified algorithm (as in the procedure by which graduating medical students in the United States are matched with the hospitals at which they complete their training). The main focus of this paper will be on determining the extent to which matching procedures can be designed which give agents the incentive to honestly reveal their preferences, and which produce stable matches.[2]

Two principal results will be demonstrated. The first is that no matching procedure exists which always yields a stable outcome and gives players the incentive to reveal their true preferences, even though procedures exist which accomplish either of these goals separately. The second result is that matching procedures do exist, however, which always yield a stable outcome and which always give all the agents in *one* of the two disjoint sets of agents the incentive to reveal their true preferences. That is, it is possible to find matching procedures which produce stable outcomes and which confine to one of the two kinds of agents being matched any possible incentive to misrepresent their true preferences. For instance, in a matching problem which involves matching individuals with institutions, stable matching procedures exist which give every individual the incentive to reveal his true preferences. Why this might be a desirable property of matching procedures will be discussed.

The first result is similar in spirit to a number of impossibility results which have been encountered by investigators seeking to design nondictatorial[3] social-choice procedures which operate in fairly unrestricted domains (cf. Gibbard (1973) or Satterthwaite (1975)). The second result shows that, by exploiting the structure associated with the problem, it may be possible to confine the difficulties associated with misrepresentation of preferences to a manageable subset of the agents involved.[4] Together, the two results will permit us to draw some conclusions about "optimal" matching procedures from the point of view of someone trying to minimize the incentive to misrepresent preferences.

The difficulties associated with misrepresentation are of not only theoretical interest, but also of considerable practical concern. For instance, in the early 1950's, problems associated with the incentives for misrepresentation in an earlier procedure motivated a complete revision of the matching procedure employed by the National Intern and Resident Matching Program (NIRMP), which is the program through which most graduating medical doctors and hospitals are matched in the United States. A future paper is planned to discuss the NIRMP procedure.

In the next section, a formal model of matching problems is introduced. §3 reviews the structure of the set of stable outcomes for such problems, which has been well known since the work of Gale and Shapley (1962). §§4 through 6 present the analysis of the incentive properties of matching procedures, and §7 concludes.

**2. The formal model.** We begin by introducing a specialized model of the matching problem, which will turn out to be sufficiently general to use to explore the general problem. The simplest matching problem to model is the "marriage problem," which consists of two sets of agents $M = \{m_1, \ldots, m_n\}$ and $W = \{w_1, \ldots, w_n\}$ ("men" and "women"). Each $m_i$ in $M$ has a complete and transitive strict preference relation $P(m_i)$

[2] A stable match will be defined as a match-up of agents such that no pair of agents would both prefer to be matched to each other than to their current partners. Such a match is in the core of the cooperative game which would result if the individual agents were able to freely negotiate their own matches. Why this kind of stable match might be a desirable outcome even of procedures in which this is not the case will be discussed.

[3] In the context of matching, the requirement that a matching procedure yield a stable outcome is a strong form of the requirement that it be nondictatorial.

[4] The alternative approach which has been taken in the social choice literature is to look for restrictions on the allowable preferences of the agents which will permit possibility results. See, for example, Kalai and Muller (1977), Maskin (1976), Ritz (1981). Conclusions related to the second result presented here have recently been developed independently by Dubins and Freedman (1980).

defined on $W$. When no confusion will occur, $P(m_i)$ will sometimes be written as $P_i$; thus the statement "$w_k$ is preferred by $m_i$ to $w_j$" can be written as $w_k P(m_i) w_j$ or $w_k P_i w_j$. Each $w_i$ in $W$ has a similar preference $P(w_i)$ defined on $M$.[5] Denote by $P = (P(m_1), \ldots, P(w_n))$ the $2n$-vector of all the agents' preferences, which will sometimes be referred to as the *preference profile*.

An outcome of the (monogamous) marriage problem is a one-to-one matching of men and women, i.e., an invertible function $x: M \to W$. An outcome $x$ can be denoted by $x = [(m_1, x(m_1)), (m_2, x(m_2)), \ldots, (m_n, x(m_n))]$ where $x(m_i) = w_j$ is the woman matched with man $m_i$, and $x^{-1}(w_j) = m_i$ is the man matched with woman $w_j$.

A matching $x$ is *stable* if no man and no woman who are not matched to each other at $x$ prefer each other to their partners at $x$. That is, $x$ is stable if there is no man $m_k$ and woman $w_q$ such that both

(i) $$w_q P_k x(m_k)$$

and

(ii) $$m_k P_q x^{-1}(w_q).$$

If a pair $m_k$ and $w_q$ do exist satisfying (i) and (ii), then $x$ is *unstable* with respect to $m_k$ and $w_q$. The motivation for this terminology should be clear, and it is easily verified that the set of stable outcomes is equal to the core of the cooperative game which results if any man and woman may marry if they both agree (and in which each agent's preference for an outcome is determined solely by his preference for potential partners).

The marriage problem as outlined above differs from the general matching problem in three principal respects. First, in the marriage problem, each agent is to be matched with exactly one partner, but in the general matching problem, different agents may need to be matched with different numbers of partners, so that each agent has a "quota," and an outcome is a function which matches each agent with his quota of partners.[6] For our purposes, this difference between the marriage problem and the general matching problem turns out to be of no consequence, since all the arguments which will be used can be carried over virtually unchanged to the general case. Second, the marriage problem was defined as having equal numbers of men and women, and without the possibility that a feasible outcome could leave any agent unmatched, while in the general matching problem, there may be an excess of one kind of agent, and an outcome need not make a match for every agent. This more general case can be handled by adding suitable "dummy" agents corresponding to the option of being unmatched in the final outcome. The third respect in which the marriage problem as described above is more restrictive than the general matching problem is that indifference between potential partners has been ruled out by the assumption that all preferences are strict. Relaxing this assumption would actually complicate some of the results. To avoid these complications, only strict preferences will be considered.

Thus the marriage problem will be used to represent the general matching problem with strict preferences. This should not obscure the fact that much of the interest of the

[5] That is, the preference relations $P(m_i)$ (or $P(w_j)$) have the following properties:

(1) transitivity: if $w_j P_i w_k$ and $w_k P_i w_l$, then $w_j P_i w_l$;

(2) completeness: for all distinct $w_j$, $w_k$ in $W$, exactly one of the relations $w_j P_i w_k$ or $w_k P_i w_j$ holds.

The interpretation of the preference relations $P_i$ as strict preference relations means that we are ruling out the possibility that an agent will be indifferent between two potential partners. (Also, it is never the case that $w_j P_i w_j$.) See the discussion at the end of this section.

[6] Note that, in the general case, the fact that each agent's preferences are defined over individual agents with whom he might be matched means that the desirability of a particular match is not affected by the other matches made. Thus, for example, a college's desire to have a geographically diverse student body cannot be directly reflected in this model, although it might be possible for the college to achieve the same effect by indicating that the college preferred, say, students from New Mexico to otherwise equally well qualified students from New York.

results to be presented derives from matching problems which are not marriage problems, since perhaps the most common kinds of matching problems are those which involve matching individuals with institutions, with the institutions having quotas greater than 1. The results which will be presented apply unchanged to these cases as well.

**3. The set of stable outcomes.** This section reviews two surprising facts about the set of stable outcomes of matching problems, closely following the discussion of these matters originally given by Gale and Shapley (1962). First, the set of stable outcomes of the general matching problem is always nonempty, i.e., regardless of the preferences of the players, at least one stable outcome always exists. Second, for each of the two disjoint sets of agents, there always exists a stable outcome which is preferred to all other stable outcomes by every member of that set of agents.

As discussed in the previous section, it will be sufficient to prove both results for the case of the marriage problem. It will be convenient to refer to the agents $M = \{m_1, \ldots, m_n\}$ and $W = \{w_1, \ldots, w_n\}$ as "men" and "women," respectively.

THEOREM 1. *The set of stable outcomes is always nonempty.*

PROOF. Following Gale and Shapley (1962) the proof proceeds by displaying a procedure which, for arbitrary preferences of the agents, constructs a stable outcome. The procedure works as follows:

*Step* 1. (a) Each man proposes to his most preferred woman.

(b) Each woman rejects all but her most preferred of the men who have proposed, and keeps the most preferred man as her suitor.

⋮

*Step* $k$. (a) Each man who has been rejected in the previous step proposes to the most preferred of those women who have not yet rejected him (i.e., to whom he has not yet proposed).

(b) Each woman keeps as her suitor the man she most prefers among those who have proposed (including the man she kept as her suitor at the end of step $k-1$), and rejects the rest.

The procedure terminates at the outcome which results when every woman has received at least one proposal (at which point each woman has exactly one suitor). Since the sets of agents are finite, this always occurs in a finite number of steps. The resulting outcome is stable, since any woman preferred by a man to his own final partner has already rejected him, and hence prefers her final partner to him. This completes the proof.[7]

A specific realization of this repeated proposal procedure, when the preference profile of the agents is $P$, will be denoted $G(P)$. Thus we can speak of the women proposed to, or the men rejected, at any step $k$ of $G(P)$. The outcome selected by this procedure when the preference profile is $P$ will be denoted $g(P)$. Gale and Shapley also proved the following.

THEOREM 2. *There is a stable outcome weakly preferred by every man to any other stable outcome, and one weakly preferred by every woman.*

PROOF. We will show that the stable outcome $g(P)$ is weakly preferred by every man to every other stable outcome, i.e., that for any other stable outcome, every man either gets a less desirable match than at $g(P)$, or else gets the same partner at both outcomes. The symmetry of the problem then implies that the corresponding outcome

[7] The only adjustment which would be needed to handle the general case in which each agent is to be matched to some quota of partners would be that each man proposes to his full quota at each step, and each woman rejects men only when her quota is full.

for the women can be constructed by exchanging the roles of $M$ and $W$ in the repeated proposal procedure, i.e., by having women propose, and men accept or reject.

For each man $m_i$, call a woman $w_j$ *possible* for $m_i$ if there is some stable outcome $x$ for which $x(m_i) = w_j$. Suppose that, up to step $k-1$ of the procedure $G(P)$, no man has been rejected by a possible woman, and that, at step $k$, $m_i$ is rejected by $w_j$. Then if we show that $w_j$ is not possible for $m_i$, we will have shown by induction that no man is ever rejected by a possible woman (at any step of $G(P)$), which is sufficient to complete the proof.

Let $m_l$ be the man who $w_j$ did *not* reject at step $k$. Then $w_j$ prefers $m_l$ to $m_i$, and $m_l$ prefers $w_j$ to any $w_i$ who hasn't already rejected him. By the inductive assumption, this means $m_l$ prefers $w_j$ to any (other) $w_i$ who is possible for him. So any outcome which pairs $m_i$ and $w_j$ and pairs $m_l$ with a woman who is possible for him, is unstable with respect to $m_l$ and $w_j$. So $w_j$ isn't possible for $m_i$, which completes the proof.

So far we have discussed the repeated proposal procedure introduced in Gale and Shapley (1962) simply as an algorithm which permits a constructive proof of certain propositions about stable outcomes.[8] The next section concerns questions which arise if we wish to consider the possibility of actually implementing some matching procedure as a mechanism for resolving matching problems, or if we wish to evaluate the effects of some procedure already in use.

**4. Incentives and stability.** Since each agent alone knows his own preferences, any matching procedure which depends on agents' preferences can be thought of as consisting of two parts: a mechanism for eliciting the preferences of the agents, and a mechanism for aggregating these elicited preferences to determine an outcome. This section is concerned with the question of what kinds of matching procedures, if any, can be designed so as to give the agents the incentive to reveal their true preferences.[9] If a procedure does *not* give the players the incentive to reveal their true preferences, then the resulting outcome may not possess certain desirable properties even if the aggregation mechanism produces outcomes which always possess these properties with respect to the stated preferences.

A different kind of problem concerning incentives arises for matching procedures which do not result in stable outcomes, since such procedures give at least one pair of agents the incentive to disregard the matching procedure, and seek an alternative outcome. Of course, it may be possible to *compel* the agents to accept the outcome, in spite of these incentives. For instance, the procedure by which some high school athletes are matched with colleges involves signing "letters of intent," after which athletes are effectively prohibited from further negotiating with other schools. (Professional athletes in several sports are matched with teams under an even more restrictive draft, which prevents a player from negotiating with any team but the one which drafts him.) However in situations in which compulsion plays no part, it is desirable for a matching procedure to yield stable outcomes. A procedure which does so for arbitrary preference profiles will be called a *stable matching procedure*.

[8] These results have been generalized in an illuminating paper by Crawford and Knoer (1981) to a class of labor markets in which wages are determined as part of the matching process. They introduce a "salary adjustment procedure" which operates in much the same way as the repeated proposal procedure. Ritz (1982) demonstrates generalizations of these results and those obtained in §4 on a wide related class of two-sided markets. Knuth (1976) further discusses the structure of the set of stable outcomes, and discusses computational features of various procedures.

[9] Thus, in a centralized procedure which elicits preferences by asking the agents to rank-order their preferred matches, we wish to know if procedures can be designed which give agents the incentive to rank the alternatives according to their true preferences. In a decentralized procedure which elicits preferences through the agent's actions (e.g., by supposing that men propose to their most preferred woman), we wish to know if procedures can be designed which give agents the incentive to act in this straightforward manner.

Before proceeding further with this discussion, we need to make precise what is meant by a procedure which gives agents an incentive to reveal their true preferences. Once a given mechanism for aggregating stated preferences is adopted, the matching problem becomes a noncooperative game among the agents, whose payoff consists of the outcome which results, and whose strategy choices consist of what preferences to state (or act according to, cf. footnote 9). Here we define a procedure which gives the agents the incentives to reveal their true preferences as one which aggregates stated preferences in a manner such that, in the resulting noncooperative game, it is a dominant strategy for each player to state his preferences honestly.[10] In such a procedure, no matter what preferences other players may state, a player who misrepresents his preferences can achieve no better outcome than if he had stated them correctly, and he may, of course, do worse.[11]

The following result may now be stated: no stable matching procedure for the general matching problem exists which gives all the players an incentive to reveal their preferences.

THEOREM 3. *No stable matching procedure for the general matching problem exists for which truthful revelation of preferences is a dominant strategy for all agents.*

PROOF. It will be sufficient to demonstrate that a matching problem exists for which no stable matching procedure has truthful revelation as a dominant strategy. Let the two sets of agents be $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$. Let $h$ be an arbitrary stable matching procedure which, for any *stated* preference profile $P$ selects some outcome $h(P)$ contained in the set $C(P)$ of outcomes which are stable with respect to the preference profile $P$. Suppose that the preferences of the players are given by the preference profile $P = (P(m_1), P(m_2), P(m_3), P(w_1), P(w_2), P(w_3))$ defined as follows:[12]

$$
\begin{aligned}
P(m_1)&: \quad w_2 P_1 w_1 P_1 w_3 ,\\
P(m_2)&: \quad w_1 P_2 w_2 P_2 w_3 ,\\
P(m_3)&: \quad w_1 P_3 w_2 P_3 w_3 ,\\
P(w_1)&: \quad m_1 P_1 m_3 P_1 m_2 ,\\
P(w_2)&: \quad m_3 P_2 m_1 P_2 m_2 ,\\
P(w_3)&: \quad m_1 P_3 m_2 P_3 m_3 .
\end{aligned}
$$

Then the set of stable outcomes is

$$
C(P) = \left\{ x = \left[(m_1, w_2), (m_2, w_3), (m_3, w_1)\right], \quad y = \left[(m_1, w_1), (m_2, w_3), (m_3, w_2)\right] \right\}.
$$

That is, this preference profile has exactly two stable outcomes: the outcome $x$, which matches $m_1$ with $w_2$, $m_2$ with $w_3$, and $m_3$ with $w_1$, and the outcome $y$, which matches $m_1$ with $w_1$, $m_2$ with $w_3$, and $m_3$ with $w_2$. Note that the men prefer $x$ while the women prefer $y$ (with $m_2$ and $w_3$ indifferent). Since $h$ is a stable matching procedure, $h(P)$ equals either $x$ or $y$.

[10] When any possible true preferences may occur, an equivalent definition is that a procedure which gives the agents an incentive to reveal their true preferences is one in which, in the resulting noncooperative game, it is always a Nash equilibrium for all the players to state their true preferences (cf. Dasgupta, Hammond, and Maskin (1979)).

[11] All judgments about whether an agent likes one outcome better than another must obviously be made in terms of the agent's *true* preferences.

[12] That is, $m_1$ prefers $w_2$ to $w_1$ to $w_3$; $m_2$ and $m_3$ prefer $w_1$ to $w_2$ to $w_3$; $w_1$ prefers $m_1$ to $m_3$ to $m_2$, etc.

Now suppose that, instead of the preference relation $P(w_1)$, $w_1$ instead stated the alternative preference relation $P'(w_1)$ given by

$$P'(w_1): \quad m_1 P_1' m_2 P_1' m_3 .$$

Let $P' = (P(m_1), P(m_2), P(m_3), P'(w_1), P(w_2), P(w_3))$ be the preference profile which differs from $P$ only in that $P'(w_1)$ replaces $P(w_1)$. Then the outcome $y$ is the unique stable outcome with respect to the preference profile $P'$, i.e., the set of stable outcomes is $C(P') = \{y\}$. Since $h$ is a stable matching procedure, $h(P') = y$.

Similarly, let $P''$ be the preference profile which differs from $P$ only in that $P''(m_1)$ replaces $P(m_1)$, where $P''(m_1)$ is given by

$$P''(m_1): \quad w_2 P_1'' w_3 P_1'' w_1 .$$

Then the outcome $x$ is the unique stable outcome with respect to $P''$, i.e., $C(P'') = \{x\}$, and so $h(P'') = x$.

So if, in the original problem, $h(P)$ equals $x$, then $w_1$ has an incentive to state the preference relation $P'(w_1)$ instead of the true preference $P(w_1)$, in order to change the outcome from $x$ to $y$ (which changes $w_1$'s partner from $m_3$ to $m_1$). And if, instead, $h(P)$ equals $y$, then it is $m_1$ who has an incentive to misrepresent his preferences as $P''(m_1)$, to change the outcome from $y$ to $h(P'') = x$. Since $h$ was an arbitrary stable matching procedure, this completes the proof.

To see the role played by stability in Theorem 3, observe that there are efficient matching procedures, i.e., procedures which always yield Pareto optimal (but not necessarily stable) outcomes, which do not give players any incentive to misrepresent their preferences.

THEOREM 4. *Efficient matching procedures exist for which truthful revelation of preferences is a dominant strategy for every agent.*

PROOF. Consider the procedure which, for any stated preference profile $P$, yields the outcome $f(P) = x$ in which $x(m_1)$ is the most preferred partner of $m_1$, and $x(m_k)$ is the most preferred partner of $m_k$ in the set $W \setminus \{x(m_1), \ldots, x(m_{k-1})\}$ for $k = 1, \ldots, m$. That is, this procedure matches $m_1$ with his (stated) first choice, $m_2$ with his (stated) first choice of the remaining $w_i$, and so forth. Truthful revelation of preferences is clearly a dominant strategy for the men, and it is also a (degenerately) dominant strategy for the women, whose preferences have no influence on the outcome of this procedure (which bears some resemblance to the football draft). Although the outcome $f(P)$ need not be stable, it is always Pareto optimal with respect to $P$, since at any other outcome some man would do worse. This completes the proof.

So there are matching procedures which always yield stable outcomes, and there are efficient matching procedures in which truthful revelation is a dominant strategy for every agent, but no matching procedure exists which meets both these requirements. However, it is possible to find stable matching procedures which confine any incentive for misrepresentation to either one of the two sets of agents, and which constrain the scope for misrepresentation by those players. Specifically, we will prove the following results, which make use of the fact that Theorem 1 permits us to identify, for each set of agents, a unique optimal stable outcome, which they each like at least as well as any other stable outcome.

THEOREM 5. *In the matching procedure which always yields the optimal stable outcome for a given one of the two sets of agents (i.e., for $M$ or for $W$), truthful revelation is a dominant strategy for all the agents in that set.*

COROLLARY 5.1. *In the matching procedure which always yields the optimal stable outcome for a given set of agents, the agents in the other set have no incentive to misrepresent their first choice.*

Note that both results are phrased in terms of "the" matching procedure which yields a particular outcome. There are obviously different procedures which yield the same outcome, but from the point of view of incentivies such procedures are equivalent, and can be regarded as a single procedure. Note also that Theorem 5 implies that, although an agent can in general change the set of stable outcomes by misrepresenting his preferences, no agent can manipulate his preferences in such a way that he prefers his best outcome in the altered set of stable outcomes to his best outcome in the original set of stable outcomes. The next section is devoted to the proof of these results, which is somewhat more complex than the proof of the earlier results. §6 presents some additional results about the structure of the set of stable outcomes, which arise in the course of the proofs.

**5. Proofs of Theorem 5 and its corollary.** In this section it will be shown that the repeated proposal procedure $G$ used in the proof of Theorem 1 has the properties stated in Theorem 5 and Corollary 5.1. As discussed in section 2, to establish these results for the general matching problem with strict preferences, it will be sufficient to demonstrate them for the marriage problem. Throughout this section, therefore, the sets of agents will be $M = \{m_1, \ldots, m_n\}$ and $W = \{w_1, \ldots, w_n\}$, the true preferences of the players will be given by the arbitrary preference profile $P$, and $x = g(P)$ will denote the outcome which results from the repeated proposal procedure when the true preferences are stated.

To prove Theorem 5, we need to show that truthful revelation is a dominant strategy for each $m_i$ in $M$. Since the (true) preference profile $P$ is arbitrary, it will be sufficient to show that, if $P'$ is a preference profile which differs from $P$ only in that $P'(m_i)$, say, replaces man $m_i$'s true preferences $P(m_i)$, then man $m_i$ doesn't prefer the outcome $y = g(P')$ to $x = g(P)$. That is, we need to show that no *successful* misrepresentation of preferences is possible, where a misrepresentation $P'(m_i)$ is defined to be a successful misrepresentation by man $m_i$ if $y(m_i)P(m_i)x(m_i)$. That is, a misrepresentation $P'(m_i)$ is successful if $m_i$ prefers (according to his true preferences) the partner $y(m_i)$ he's matched with when he misrepresents his preferences to the partner $x(m_i)$ he's matched with when he states his true preferences. (Throughout this section, $y = g(P')$ will denote an outcome resulting from misrepresentation by $m_i$.)

We first show that only a certain kind of simple misrepresentation need be considered, since if any successful manipulation is possible, then it can be achieved by a simple misrepresentation. Specifically, if $P'(m_i)$ is an arbitrary misrepresentation as discussed above, then an equivalent *simple* misrepresentation $P''(m_i)$ is one in which $y(m_i)P''(m_i)w_j$ for all $w_j \neq y(m_i)$. That is, $P''(m_i)$ is a preference relation which has $y(m_i)$ as the most preferred match of $m_i$. The justification for calling $P''(m_i)$ an *equivalent* misrepresentation to $P'(m_i)$ is given by the following lemma, which states that $m_i$ will end up matched to the same partner, $y(m_i)$, whether he misrepresents using $P'(m_i)$ or $P''(m_i)$. (The preference profile $P''$, of course, denotes the one which differs from $P$ only in that $P''(m_i)$ replaces $P(m_i)$.

LEMMA 1. *If $y = g(P')$ and $z = g(P'')$ then $z(m_i) = y(m_i)$.*

PROOF. The outcome $y$ is stable with respect to the preference profile $P''$ (i.e., $y$ is in $C(P'')$), since $y$ is in $C(P')$ and since no new potential instabilities for $y$ arise in changing from $P'$ to $P''$. So $y(m_i)$ is "possible" for $m_i$ under the preference profile $P''$, in the sense defined in the proof of Theorem 2. Since $y(m_i)$ is $m_i$'s most preferred match according to $P''(m_i)$, it is the best possible match for $m_i$ with respect to $P''$. But,

by Theorem 2, $z = g(P'')$ gives every man his best possible match with respect to the profile $P''$, and so $z(m_i) = y(m_i)$ as was to be proved.

So Lemma 1 shows that, to prove Theorem 5, it is sufficient to prove that no simple misrepresentation $P'(m_i)$ (i.e., no manipulation in which $m_i$ proposes to $y(m_i)$ in step 1 of $G(P')$) can be successful. The following lemma states that if a misrepresentation by $m_i$ leaves $m_i$ at least as well off as at $x = g(P)$, then no man will suffer, i.e., *every* man likes the outcome $y = g(P')$ resulting from the misrepresentation at least as well as the outcome $x = g(P)$.

LEMMA 2. *If $P'(m_i)$ is a simple misrepresentation such that $y = g(P')$ and either $y(m_i)P(m_i)x(m_i)$ or $y(m_i) = x(m_i)$ then, for each $m_j$ in $M$, either $y(m_j)P(m_j)x(m_j)$ or $y(m_j) = x(m_j)$.*

PROOF. Suppose, to the contrary, that $x(m_j)P(m_j)y(m_j)$ for some $m_j$ in $M$, i.e., $m_j$ does worse in the outcome $y$ than in $x$. Since every agent other than $m_i$ states the same preferences in the profiles $P$ and $P'$, it must be that $m_j$ is rejected by $x(m_j)$ at some step of the procedure $G(P')$. Let $l$ be the *first* step of $G(P')$ at which some $m_j$ is rejected by $x(m_j)$. Then $x(m_j)$ must have received a proposal in step $l$ of $G(P')$ from some $m_k$ who did *not* propose to her in $G(P)$, such that $m_k P(x(m_j))m_j$, i.e., from an $m_k$ who $x(m_j)$ prefers to $m_j$. The fact that $m_k$ didn't propose to $x(m_j)$ in $G(P)$ means $x(m_k)P(m_k)x(m_j)$, and so $m_k$ must have been rejected by $x(m_k)$ in $G(P')$ prior to step $l$, which contradicts the choice of $l$ as the first such period. Consequently, no $m_j$ is rejected by $x(m_j)$ in $G(P')$, which completes the proof.

We can now proceed to prove Theorem 5, by showing that no man can successfully misrepresent his preferences in the repeated proposal procedure (in which men do the proposing).

PROOF OF THEOREM 5. Let $x = g(P)$, and suppose that $x$ results from the repeated-proposal procedure in $t$ steps, i.e., $G(P)$ terminates at step $t$. Let $P'$ be the preference profile which results from $P$ when one agent, $m_i$, makes a simple misrepresentation to obtain $y = g(P')$. We want to show that this misrepresentation cannot be successful, and we will proceed by assuming that either $y(m_i)P(m_i)x(m_i)$ or $y(m_i) = x(m_i)$ and then showing that only the latter alternative can occur. That is, we consider only manipulations which don't actually harm the manipulator, and then show that they don't help him either.

For any $m_j$ in $M$, say that $m_j$ *makes a match at step $k$* of $G(P)$ if $m_j$ proposes to his ultimate partner $x(m_j)$ at step $k$. Note that each $m_j$ makes a match exactly once.

Now we will show that, for any $m_j$ who makes a match at period $t$ of $G(P)$ (the final period), $y(m_j) = x(m_j)$. This follows from the fact that, since $t$ was the final step of $G(P)$, $m_j$ was the only man who proposed to $x(m_j)$ in $G(P)$ (since otherwise at least one more step would have occurred). But, by Lemma 2, no man does worse at $y$ than at $x$, so no man proposes to $x(m_j)$ in $G(P')$ who didn't propose to $x(m_j)$ in $G(P)$. Consequently only $m_j$ proposes to $x(m_j)$ in $G(P')$ (since $x(m_j)$ receives at least one proposal), and so $y(m_j) = x(m_j)$. The same conclusion holds for any $m_j$ who is the only one to propose to $x(m_j)$ in $G(P)$, regardless of the period at which he makes his match. So if the manipulator $m_i$ made a match at the final step $t$ of $G(P)$, or if he made a match with someone who received no other proposals in $G(P)$, then his manipulation can't be successful, and we're done.

Suppose instead that the manipulator $m_i$ makes a match at some other step $k$ of $G(P)$ $(1 \leq k < t)$. We will show by induction that, for every $m_j$ (including $m_i$) who makes a match at step $k$ or later, $y(m_j) = x(m_j)$.

Let $r$ be a step of $G(P)$ such that $k \leq r < t$. We have already demonstrated the desired conclusion for any $m_j$ who makes a match at step $t$. The inductive part of the proof is to show that, if $y(m_j) = x(m_j)$ for every $m_j$ who makes a match at steps $r + 1$

through $t$ of $G(P)$, then $y(m_j) = x(m_j)$ for every $m_j$ who makes a match at step $r$ of $G(P)$.

Let $m_q$ be a man who makes a match at step $r$ of $G(P)$. Let $M'$ be the subset of men who were rejected by $x(m_q)$ in $G(P)$, i.e., $M' = \{m_j \text{ in } M | x(m_q) P(m_j) x(m_j)\}$ is the subset of men $m_j$ who prefer $x(m_q)$ to their ultimate partner. If $M'$ is empty, then $y(m_q) = x(m_q)$ by the argument of the previous paragraph. If not, let $m_u$ be the man in $M'$ such that $m_u P(x(m_q)) m_s$ for all $m_s$ other than $m_u$ in $M'$. That is, $m_u$ is preferred by $x(m_q)$ to all the other men she rejected in $G(P)$.

Then $m_u$ makes his match *after* step $r$ of $G(P)$, since he's not rejected by $x(m_q)$ until step $r$. So $y(m_u) = x(m_u)$ by the inductive hypothesis.

Since $m_u$ isn't the manipulator (i.e., $m_u \neq m_i$), this means that $m_u$ proposes to $x(m_q)$ in $G(P')$ and is rejected in favor of someone who $x(m_q)$ prefers. But since no man proposes to $x(m_q)$ in $G(P')$ who didn't also propose to her in $G(P)$, this means that $x(m_q)$ rejects $m_u$ in favor of $m_q$, so $y(m_q) = x(m_q)$. Thus, by induction, $y(m_j) = x(m_j)$ for every $m_j$ who makes a match at step $k$ or later. In particular, $y(m_i) = x(m_i)$, so the manipulation is not successful. This completes the proof.

To prove Corollary 5.1, it is sufficient to note that, since the repeated proposal procedure has men proposing and women accepting or rejecting, the only opportunity for misrepresentation which a woman has is to reject a more preferred man in favor of someone less preferred at some step of the procedure. Theorem 3 showed that this can sometimes lead to a more preferable final match, but obviously if some $w_j$ receives a proposal at any step of the procedure from her *first* choice, she can do no better than to accept, which establishes Corollary 5.1.

**6. Further results.** Note that Theorem 5 and its proof leave open the possibility that men who make a match *before* $m_i$ may profit from his misrepresentation, even though $m_i$ cannot himself gain any benefit from misrepresentation. The following example shows that this is indeed possible. Let $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$, with preferences

$$\begin{aligned} P(m_1)&: w_2 P_1 w_1 P_1 w_3, & P(w_1)&: m_1 P_1 m_2 P_1 m_3, \\ P(m_2)&: w_1 P_2 w_2 P_2 w_3, & P(w_2)&: m_3 P_2 m_1 P_2 m_2, \\ P(m_3)&: w_1 P_3 w_2 P_3 w_3, & P(w_3)&: m_1 P_3 m_2 P_3 m_3. \end{aligned}$$

Then $g(P) = [(m_1, w_1), (m_2, w_3), (m_3, w_2)]$.

If $m_2$ misrepresented his preferences as $P'(m_2)$: $w_3 P_2 w_1 P_2 w_2$ then

$$g(P') = [(m_1, w_2), (m_2, w_3), (m_3, w_1)]$$

which leaves $m_2$ no worse than at $g(P)$, but which benefits $m_1$ and $m_3$.

Another consequence of the argument used to establish Theorem 5 is the following result, which compares the best stable outcome for the men with the set of *all* feasible outcomes (stable or not).

THEOREM 6. *No feasible outcome is strictly preferred by all $m_i$ in $M$ to the outcome $g(P)$.*

The fact that no *stable* outcome is strictly preferred to $g(P)$ isn't news: Theorem 2 gives a stronger result. What Theorem 6 says is that in fact, $g(P)$ is weakly Pareto optimal from the point of view of the men, with respect to any possible outcome. The example above shows that this can't be strengthened to strong Pareto optimality.

PROOF OF THEOREM 6. I am indebted to David Gale for pointing out that the proof follows almost immediately from the observation (in the proof of Theorem 5) that if $m_i$ makes a match in the final period $t$ of $G(P)$, then $x(m_i)$ receives only one proposal in $G(P)$. So if $y$ is any outcome which $m_i$ prefers to $x$ (i.e., $y(m_i) P(m_i) x(m_i)$) then some

other $m_i \neq m_j$ must be matched with $x(m_j)$ at $y$ (i.e., $y(m_i) = x(m_j)$). But the fact that $m_i$ didn't propose to $x(m_j)$ in $G(P)$ means $m_i$ prefers $x(m_i)$ to $x(m_j)$, which completes the proof.

Taken together, Theorems 3 and 5 and Corollary 5.1 provide bounds on how much misrepresentation we can hope to preclude in any stable matching procedure. Theorem 3 shows that it isn't possible to remove all incentive for misrepresentation, but Theorem 5 shows that such incentives can be removed from one of the two sets of agents, and Corollary 5.1 shows that the incentive to misrepresent can simultaneously be somewhat constrained in the other set of agents. In fact, the procedure discussed in Theorem 5 and its corollary take us as far as we can go in this direction. The following result formalizes the sense in which this is the case.

THEOREM 7. *No stable matching procedure exists which never gives any agent an incentive to misrepresent his kth choice, for* $k \neq 1$.

PROOF. The result follows from the proof of Theorem 3. Examples of the kind used there can obviously be constructed, such that an agent in one of the two sets will have an incentive to misrepresent his $k$th choice, for any $k > 1$.

**7. Discussion.** The theorems presented in §§3, 4 and 6 demonstrate that the structure of the matching problem allows powerful conclusions to be drawn about the set of possible outcomes and the procedures which can be used to select a particular outcome. Consider, for example, the matching problem which involves students and colleges; specifically, the problem of matching students with the colleges at which they will matriculate.

Theorem 1 shows that the set of stable outcomes is nonempty, so that, when the preferences of students for colleges and colleges for students are known, it is always possible to assign students to colleges in a way which gives colleges the incentive to admit the students they were assigned, and students the incentive to attend the college to which they were assigned, since neither can hope to find a more preferable match. Furthermore, Theorem 2 shows that the set of stable outcomes has a structure which reflects common interests of students or of colleges. It is somewhat surprising that common interests of this kind can be identified, since the nature of the problem is that students compete with each other for the best (i.e., the most widely preferred) colleges, and colleges compete with each other for the best students. But when attention is confined to the set of stable outcomes, these causes of competition and conflicting interests disappear, and all students have a common interest in the "student-optimal" stable outcome, while all colleges prefer the "college-optimal" stable outcome. Theorem 6 shows that this common interest is not in conflict with the requirement of stability, i.e., even if stability were not required, students could not all do better than the student optimal stable outcome.

A similar structure remains when, in §4, the assumption that the preferences are known is abandoned. Although Theorem 3 shows that it isn't possible to find a stable matching procedure which doesn't potentially give some agent an incentive to misrepresent his preferences, Theorem 5 shows that it *is* possible to confine this incentive to misrepresent to either one of the two sets of agents. This latter result suggests that, despite the result of Theorem 3, it may be possible to largely avoid the distortions introduced by misrepresentation in matching problems like the problem of matching students and colleges, in which one set of agents consists of institutions rather than individuals.[13]

[13] In certain respects this result may have some resemblance to the results of Wilson (1978) concerning competitive exchange markets in which one player is assigned the role of the passive auctioneer.

In particular, suppose that the matching procedure is used which yields the student-optimal stable assignment of students to colleges, and which gives no student any incentive to misrepresent his preferences. Since it is a dominant strategy for each student to reveal his true preferences, the only potential source of distortion of the procedure lies in the stated preferences of the colleges. But the preferences of colleges (and institutions in general) are likely to be much more regular than the preferences of students (and individuals in general), so that colleges may have less scope for (undetectable) misrepresentation. For example, the kinds of preferences which colleges may exhibit are already influenced by legislation and regulation designed to eliminate racial discrimination. The enforcement of these laws and regulations presupposes that the preferences exercised by a college can be examined (e.g., through litigation) with sufficient reliability to determine which choices result from "legitimate" preferences and which from discriminatory preferences. And, to the extent that colleges rank students through objective criteria like grades or exam scores, the degree to which "strategic" opportunities arise from misrepresentation of preferences over other factors is reduced.

## References

[1] Crawford, Vincent P. and Knoer, Elsie Marie. (1981). Job Matching with Heterogeneous Firms and Workers. *Econometrica* **49** 437–450.

[2] Dasgupta, Partha, Hammond, Peter and Maskin, Eric. (1979). The Implementation of Social Choice Rules: Some General Results on Incentive Compatibility. *Rev. Econom. Stud.* **XLVI** 185–216.

[3] Dubins, L. E. and Freedman, D. A. (1980). Machiavelli and the Gale-Shapley Algorithm. Mimeo.

[4] Gale, David and Shapley, Lloyd S. (1962). College Admissions and the Stability of Marriage. *Amer. Math. Monthly* **69** 9–15.

[5] Gibbard, Allan. (1973). Manipulation of Voting Schemes: A General Result. *Econometrica* **41** 587–601.

[6] Kalai, Ehud and Muller Eitan. (1977). Characterization of Domains Admitting Nondictatorial Social Welfare Functions and Nonmanipulable Voting Procedures. *J. Econom. Theory* **16** 456–469.

[7] Knuth, Donald. (1976). *Mariages Stables*. Les Presses de l'Université de Montréal.

[8] Maskin, Eric. (1976). On Strategy Proofness and Social Welfare Functions When Preferences are Restricted. Mimeo. Harvard University.

[9] Ritz, Ziv. (1981). On Arrow Social Welfare Functions and on Nonmanipulable and Noncorruptible Social Choice Functions. Unpublished Ph.D. dissertation, Northwestern University.

[10] ———. (1982). Incentives and Stability in Some Two-Sided Economic and Social Models. Mimeo.

[11] Satterthwaite, Mark. (1975). Strategy-Proofness and Arrow's Conditions: Existence and Correspondence Theorems for Voting Procedures and Social Welfare Functions. *J. Econom. Theory* **10**.

[12] Wilson, Robert. (1978). Competitive Exchange. *Econometrica* **46** 577–585.

DEPARTMENT OF BUSINESS ADMINISTRATION, UNIVERSITY OF ILLINOIS, URBANA, ILLINOIS 61801