European Economic Review 27 (1985) 75–96. North-Holland

COMMON AND CONFLICTING INTERESTS IN TWO-SIDED MATCHING MARKETS

Alvin E. ROTH*

University of Pittsburgh, Pittsburgh, PA 15260, USA

This paper reviews and synthesizes some of what is now known about what I will call *two-sided matching markets*. A number of such models, which arise naturally in the study of certain labor markets, have been shown to have surprising implications about the common and conflicting interests of the agents, and about the incentives they face.

1. Introduction

The purpose of this paper is to review and synthesize some of what is now known about what I will call *two-sided matching markets*. A number of such models, which arise naturally in the study of certain labor markets, have been shown to have surprising implications about the common and conflicting interests of the agents, and about the incentives they face. While it is not yet known precisely which features of these models account for each of the properties they share, their 'two-sidedness' and the 'matching' requirement clearly play a central role.

The phrase 'two-sided' refers to the fact that agents in such markets belong to one of two disjoint sets — e.g. firms or workers — that are specified in advance. This contrasts for example with commodity markets in which the market price determines whether an agent is a buyer or a seller. The term 'matching' refers to the bilateral nature of exchange in these markets — e.g., if I am employed by the University of Pittsburgh, then the University of Pittsburgh employs me. This contrasts with commodity markets in which I may consume some of your initial endowment even though you consume none of mine.

My own interest in such markets was first aroused by the labor market in which graduating medical students seek entry-level positions (internships and residencies) in American hospitals. That market is administered by means of a central clearinghouse, to which students submit a rank-ordering of the

*I have profited from conversation and correspondence on this topic with Charles Blair, who has shared with me some of his work in progress. This work has been supported by grants from the National Science Foundation and the Office of Naval Research, and by Fellowships from the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan Foundation.

hospitals to which they have applied, and hospitals submit a rank-ordering of the students who have applied to them. The clearinghouse then uses this information to produce a matching of students and hospitals, by means of an algorithm adopted by the relevant medical associations in 1951.

In the medical literature describing this procedure, and distributed each year to participating students and hospitals, it is claimed that the algorithm works in such a way that no participant (i.e., no student or hospital) can do better than to submit a rank-ordering corresponding to his *true* preferences. In the course of investigating this claim [Roth (1982a)], and finding it to be false (see Theorem 5.1 of this paper), it became apparent that the algorithm, and the history of the market before and after the adoption of this algorithm in 1951, both cast light on some unusual properties of this market that relate to its two-sided matching structure. Also, many of the issues that arose as practical problems in the course of organizing this market either anticipated by a number of years the discussion of related matters in the literature of economics and game theory,[1] or else required novel game-theoretic analysis [see Roth (1984b)].

The plan of this paper will be as follows. In order to place these results in a specific context, section 2 will briefly describe the history and operation of the labor market for medical interns and residents. Section 3 will introduce a two-sided matching model of this market, which will be the primary model considered here. (A wide variety of related models have recently been shown to have similar properties, and these other models will be referred to informally.) A set of *stable* outcomes will be defined, closely related to the core of the market, and it will be argued that this kind of stability is the appropriate equilibrium concept for this kind of market.[2] Section 4 will consider some of the more striking properties of the set of stable outcomes, as well as some of the underlying structural properties that may eventually provide satisfactory theoretical explanations for these phenomena, which are still poorly understood. Section 5 will consider the incentive properties of procedures designed to produce stable outcomes. Together with each of these results will be a brief description of the related models in which similar results also hold, and those in which they fail to hold, in order to describe what is presently known (and how much more is unknown) about the underlying causes and the generality of these results. Section 6 will consider open questions and extensions suggested by these results, both of a theoretical nature, and of an empirical nature, concerning institutional and procedural features of markets.

[1]For example, an algorithm equivalent to the one adopted in 1951 was independently proposed a decade later by Gale and Shapley (1962).

[2]Instead of 'equilibrium concept' I should perhaps say 'solution concept' to connote that the discussion of stability will be in the framework of cooperative game theory, whereas the discussion of incentives will be in the framework of non-cooperative game theory.

2. The labor market for American medical interns[3]

Hospitals first began offering newly-graduated medical students internship positions around the year 1900. Not until 1945 were the relevant medical associations able to institute a single market for these positions by establishing uniform dates at which such positions could be offered by hospitals, and accepted by students.[4] Once this was accomplished, however, both students and hospitals were dismayed by the chaotic conditions that developed between the time offers of internships were first made, and the time by which students were required to accept or reject them. The situation that developed is described as follows in Roth (1984b):

> 'Basically, the problem was that a student who was offered an internship at, say, his third choice hospital, and who was informed he was an alternate (i.e. on a waiting list) at his second choice, would be inclined to wait as long as possible before accepting the position he had been offered, in the hope of eventually being offered a preferable position. Students who were pressured into accepting offers before their alternate status was resolved were unhappy if they were ultimately offered a preferable position, and hospitals whose candidates waited until the last minute to reject them were unhappy if their preferred alternate candidates had in the meantime already accepted positions. Hospitals were unhappier still when a candidate who had indicated acceptance subsequently failed to fulfil his commitment after receiving a preferable offer. In response to pressure originating chiefly from the hospitals, a series of small procedural adjustments were made in the years 1945–51. The nature of these adjustments, described next, makes clear how these problems were perceived by the parties involved.
>
> For 1945, it was resolved that hospitals should allow students ten days after an offer had been made to consider whether to accept or reject it. For 1946, it was resolved that there should be a uniform appointment date (July 1) on which offers should be tendered..., and that acceptance or rejection should not be required before July 8. By 1949, [the Association of American Medical Colleges] proposed that appointments should be made by telegram at 12:01 AM (on November 15), with applicants not being required to accept or reject them until 12:00 Noon the same day. Even this twelve-hour waiting period was rejected by the American Hospital Association as too long: the joint

[3]References and further details can be found in Roth (1984b), from which the material in this section is drawn.

[4]The earlier difficulties encountered in this market will not concern us here, but see Roth (1984b) for a description of the prisoners' dilemma problem that interfered with the setting of uniform appointment dates prior to 1945.

> resolution finally agreed upon contained the phrase "no specified waiting period after 12:01 AM is obligatory," and specifically noted that telegrams could be filed in advance for delivery precisely at 12:01 AM. In 1950, the resolution again included a twelve-hour period for consideration, with the specific injunction that "Hospitals and/or students shall not follow telegrams of offers of appointment with telephone calls" until after the twelve-hour grace period. [... the injunction against telephone calls was two-way, in order to stem a flood of calls both from hospitals seeking to pressure students into an immediate decision, and from students seeking to convert their alternate status into a firm offer.]'

It was eventually recognized that these problems could not be solved by compressing the last stage of the matching process into a shorter and shorter time period, and it was agreed to instead try a centralized matching algorithm, to be used on a voluntary basis. Students and hospitals would continue to exchange information via applications and interviews as before, but then both students and hospitals would submit rank-orderings of their potential assignments,[5] and the algorithm would be used to suggest a matching of students to hospitals, who would then, it was hoped, find no reasons not to sign employment contracts with their suggested assignments.

The first algorithm to be proposed was abandoned after a year because it was observed to give students the incentive to submit a rank-ordering different from their true preferences. The algorithm that was proposed in its place was used for the first time in 1951, and remains in use to this day. (This algorithm, which is described in appendix 1, will be called the NIMP algorithm, where NIMP stands for National Intern Matching Program, which is the name under which the algorithm was initially administered.)

Note that this system of arranging matches was conceived and implemented as a *voluntary* procedure — students and hospitals were free to try to arrange their own matches outside of the system, and there was no means of enforcing compliance on those who did participate.[6] This makes it all the more remarkable that, in the first years of operation, over 95% of eligible students and hospitals participated in the system, and these high rates of participation continued until the early 1970's. (Since then, increasing numbers of students, particularly those among the growing number of medical students who are married to other medical students, have begun to seek to

[5]Regarding the problem of formulating a rank-ordering, note that the complete job-description offered by a hospital program in a given year was customarily specified in advance. Thus the responsibilities, salary, etc. associated with a given internship, while they might be adjusted from year to year in response to a hospital's experience in the previous year's market, were not a subject of negotiation with individual job candidates.

[6]The experience prior to 1950 amply demonstrated that no amount of moral suasion was effective at preventing participants from acting in what they perceived as their own best interests.

arrange their own matches, without going through the centralized clearinghouse.)

In the remainder of this paper, a model will be studied that illuminates a number of remarkable features of this market. Of course, how interesting such a model ultimately is will depend not merely on its mathematical properties, but on how well it explains the observable features of markets like the one described above, and on how widespread and important such markets are in the economy. In particular, with respect to the historical development of the market for American medical interns, a successful model should be able to explain the chaotic conditions prior to 1951, the high degree of orderly participation after 1951, and the nature of the changes in the market that contributed to a decline of these high rates in the 1970's.

3. A model of the hospital–intern market

The agents in the hospital–intern market consist of two disjoint sets $H=\{h_1,\ldots,h_n\}$ and $S=\{s_1,\ldots,s_p\}$ ('hospitals' and 'students').[7] Each hospital h_i has a *quota* q_i which is the number of students for which it has places. Each student s has a complete preference ordering $P(s)$ over the set $H\cup\{u\}$, and each hospital h has a complete preference ordering $P(h)$ over the set $S\cup\{u\}$, where u denotes the possibility of remaining unmatched.[8] That is, each agent can compare the desirability of each of his or her potential assignments, which are the agents from the opposite set and the possibility of remaining unmatched. An agent's preferences are called *strict* if the agent is not indifferent between any two distinct potential assignments. It will simplify the exposition in this paper to only consider the case in which all agents have strict preferences, and this will henceforth be assumed.[9] Let $h_jP(s)h_k$ denote that student s prefers hospital h_j to hospital h_k, and let $h_jR(s)h_k$ denote that he either prefers h_j to h_k or else is indifferent. (Note that he can only be indifferent if $j=k$, since all preferences are strict.) Similar notation will be used for the preferences of the hospitals, and $P=(P(h_1),\ldots,P(h_n),P(s_1),\ldots,P(s_p))$ will denote the vector of preference orderings of each agent for agents on the other side of the market.

[7]Formally, each h in H should be called a 'hospital program', since each hospital divides its available internships into separate programs consisting of identical positions, and these programs are administered separately.

[8]Hospitals which have some of their positions unmatched, and students who are unmatched by the NIMP algorithm, have the opportunity to enter a decentralized 'after-market' that will not be analyzed here.

[9]But see Roth (1948b, c) for a treatment of non-strict preferences. Curiously, while it is important for some of the results in section 4 that preferences be strict, similar results are known to hold for models that allow sidepayments, in which preferences essentially cannot be strict. Some examples of such models are those of Shapely and Shubik (1972), Crawford and Knoer (1981), and Demange and Gale (1984). These models arise naturally, when salary is modelled as a continuous variable, in labor markets in which salary is negotiable, for example.

An *outcome* of the market is defined by a correspondence $x: H \cup S \to H \cup S \cup \{u\}$ such that $|x(s)| = 1$ for all s in S, $|x(h_i)| = q_i$ for all h_i in H, and, for any h in H and s in S, $x(s) = h$ if and only if s is an element of $x(h)$. That is, an outcome assigns a subset of the students to a subset of the places, and leaves the rest of the students and places unmatched. (If a hospital h with quota q is assigned some number $k < q$ of students at an outcome x, then $q-k$ elements of $x(h)$ are equal to u.) No student is assigned to more than one place, and no hospital is assigned more than its quota of students.

Students' preferences over outcomes correspond precisely to their preferences over hospitals, so that student s prefers outcome x to outcome y if and only if he prefers hospital $x(s)$ to hospital $x(y)$. The preferences of hospitals over outcomes are necessarily related to their preferences over students in a more complex way, since a hospital h with quota $q > 1$ receives different *sets* of students and vacancies at different outcomes.

Specifically, let $P^{\#}(h)$ denote the preference relation of hospital h over all assignments $x(h)$ that it could receive at some outcome x. A hospital h's preferences $P^{\#}(h)$ will be called *responsive* to its preferences $P(h)$ over individual assignments if $y(h) P^{\#}(h) x(h)$ whenever $y(h)$ is obtained from $x(h)$ by replacing some student s_j (or u) in $x(h)$ with a preferred student s_k who is not in $x(h)$; i.e., whenever $y(h) = x(h) \cup \{s_k\} \setminus \{\sigma\}$ for σ in $x(h)$ and s_k not in $x(h)$ such that $s_k P(h) \sigma$. That is, a hospital h has responsive preferences over assignments if, for any two assignments that differ in only one student, it prefers the assignment containing the more preferred student. For example, if $x(h)$ assigns hospital h with a quota of $q_h = 2$ its 3rd and 4th choice students, and $y(h)$ assigns it its 2nd and 4th choice students, then hospital h prefers $y(h)$ to $x(h)$ if its preferences are responsive. Hospitals will henceforth be assumed to have preferences over sets of students that are responsive as well as strict.

Note that many different responsive preference orderings $P^{\#}(h)$ exist for any preference $P(h)$, since, for example, responsiveness does not specify whether a hospital with a quota of 2 prefers to be assigned its 1st and 4th choice students instead of its 2nd and 3rd choice students. However, the preference ordering $P(h)$ over individual students can be derived from $P^{\#}(h)$ by considering a hospital h_i's preferences over assignments $x(h_i)$ containing no more than a single student (and $q_i - 1$ copies of u). [Since $P(h)$ is completely determined by $P^{\#}(h)$, it can be thought of as a summary of the full preferences.]

The preferences of the hospitals over different outcomes x and y can now be specified as corresponding precisely to their preferences over their own assignments at x and y; i.e., a hospital h_i prefers x to y if and only if $x(h_i) P^{\#}(h_i) y(h_i)$. Denote by $P^{\#}$ the vector of preferences $P^{\#} = (P^{\#}(h_1), \ldots, P^{\#}(h_n), P(s_1), \ldots, P(s_p))$, which defines the preferences of the agents over all feasible outcomes.

An outcome x is *individually rational* if for every student s, $x(s)R(s)u$, and if for every hospital h and σ in $x(h)$, $\sigma R(h)u$. An outcome x is *unstable* if it is not individually rational or if there exist a hospital h and a student s who are not matched at $x[x(s) \neq h]$ and who each prefer one another to one of their assignments; i.e., such that $hP(s)x(s)$ and $sP(h)\sigma$ for some σ in $x(h)$. An outcome x that is not unstable will be called *stable*, and the set of stable outcomes with respect to any vector P of preference orderings will be denoted $S(P)$.[10]

An important special case of this hospital-intern model, which has received the most attention in the game-theoretic literature, is the symmetric market that results when all quotas are equal to 1; in this case the model is called the 'marriage problem' (and the two sets of agents are referred to as 'men' and 'women'). In this literature, the situation where one set of agents has quotas that may be greater than 1 is referred to as the 'college admissions problem', and it became customary to specify college admissions problems only up to the point of specifying the colleges' preferences over individual students, without considering preferences over outcomes.[11] For a long time it went unnoticed that this level of description failed to specify the college admissions problem as a well-defined game, even though the marriage problem is perfectly well defined in this manner.[12] For this reason, all theorems proved for the marriage problem were thought to carry over virtually unchanged to the college admissions problem. This turns out not to be the case: while many properties of the marriage problem do carry over to the case where preferences are responsive and quotas on one side of the market need not equal 1, other important properties do not.[13] In this paper, Theorems 4.2d and 5.3 illustrate this point.

The set of stable outcomes is closely related to the core of the game that arises from the hospital–intern market when the rules are that any student and hospital may sign an employment contract if they both agree. It is proved in appendix 2 that the set of stable outcomes equals the core defined by *weak* domination. (In the special case of the marriage problem, this coincides with the core defined by strong domination.)

[10]Based on the definition as given here, and in view of the fact that hospitals have preferences defined over sets of students and vacancies, it might seem better to call such outcomes *pairwise* stable, and to also consider some form of *group* stability, whose statement would involve groups of hospitals and students that might be larger than pairs, and hospitals' preferences $P^{\#}$ rather than P. In appendix 2 it is shown that these two definitions are equivalent when preferences are responsive.

[11]I.e., P was specified, but not $P^{\#}$.

[12]When all quotas equal 1, $P^{\#} = P$.

[13]This mis-specification and its consequences was first noted in Roth (1984e), which also contains some more detailed references to erroneous statements in the literature. For a simple model with sidepayments, Rochford (1984) observes further important differences between the case of one-to-one matching and the case of many-to-one matching in which firms seek to fill multiple, heterogeneous positions with groups of heterogeneous workers.

Before discussing the properties of stable outcomes in this market, let us first consider why this definition of stability might be an appropriate equilibrium notion for markets of this kind, and how it might be used to explain, for example, the history of the particular market outlined in the previous section. In Roth (1984b), this was approached in the following way:[14]

> 'Consider now a set of job offers from hospitals to (acceptable) students which, if the students each accept the best of the offers they have received (including the possibility of remaining unmatched), would result in an unstable outcome x. The fact that x is unstable means that there is a hospital h_i and a student s_j who would both prefer to x another outcome at which s_j was employed by h_i. So these two agents have an incentive to try to locate each other, and we might expect to witness the kind of last-minute turmoil observed in the intern market prior to 1951. Looking at the other side of the coin, we might expect that any voluntary system of organizing the market would experience similar turmoil if it produced unstable outcomes. Since the NIMP is a voluntary system that has maintained a high degree of orderly participation for many years, it is reasonable to conjecture that it produces stable outcomes, i.e. outcomes in the core of the market.'

This conjecture turns out to be correct, as will be discussed below, and so this formulation of stability turns out to have a reasonable amount of explanatory power for the behavior of the hospital–intern market immediately before and after the adoption of the NIMP algorithm. We will also see that the same theoretical construct can be used to explain the decreasing rates of participation as married couples started to make up a noticeable fraction of the medical student population, since the procedure used to match married couples produced unstable outcomes. Thus the explanatory power of this notion of stability is not limited to one episode in the history of this market. In the next section we will see that it will also illuminate why this decreasing rate of participation is likely to prove difficult to resolve.

4. Stable outcomes

The results stated formally as theorems in this section all apply to the hospital–intern model of the previous section, in which preferences are strict and responsive. The first of these is an existence theorem, which also speaks to the question of why the NIMP algorithm was so successful in the labor market for interns.

[14]A student s is *acceptable* to h if $sP(h)u$; i.e., if the hospital prefers employing s to leaving a position vacant.

4.1. Existence of stable outcomes

Theorem 4.1. For any vector P of preferences, the set S(P) of stable outcomes is non-empty. Furthermore, if students and hospitals submit rank-order lists P, then the NIMP algorithm produces an outcome in S(P).

The first formal existence result of this kind to appear in the literature is due to Gale and Shapley (1962), who were not aware of the NIMP algorithm, but proposed an equivalent algorithm.[15] A number of authors have constructively proved the existence of stable outcomes in more general two-sided matching markets, by developing related algorithms. Crawford and Knoer (1981) accomplished this for a market in which workers' salaries and job descriptions are determined endogenously, and Kelso and Crawford (1982) generalized this to a case in which the number of workers employed by a given firm is also determined endogenously, instead of being determined by a quota that is independent of which workers are employed.[16] The common features of these algorithms are well captured in a formulation due to Blair (1984a,b). It is worth noting that these algorithms have nice computational properties [see McVitie and Wilson (1971), Knuth (1976), Jones (1983)].

Gale and Shapley also showed that the existence of stable outcomes depends on the two-sided structure of the market, by observing that in the one-sided analog of the marriage problem, which they called the 'roommate problem', the set of stable outcomes may be empty.[17] However, a recent paper by Quinzii (1984) presents a non-constructive proof that suggests connections between the existence of stable outcomes in these two-sided markets, and existence results for certain one-sided markets with discrete goods.[18]

We turn now from questions of existence to an examination of the structure of the set of stable outcomes.

[15]The proofs of Theorems 4.1 and 4.2a in Roth (1984b) essentially proceeded by demonstrating this equivalence.

[16]In the model of Kelso and Crawford, some additional assumptions about firms' preferences are required to assure the existence of stable outcomes. They also find it convenient to model salary as a discrete variable, e.g., as when salary can be specified only up to the nearest penny. A closely related model and algorithm is explored in Roth (1984a).

[17]The roommate problem is defined by a single set N of players, each of whom has preferences over all the other players. The set of feasible outcomes is the set of all partititions of the players into pairs, and an outcome is unstable in case there are two players not matched to one another each of whom prefers the other to the player he is matched with.

[18]These markets, which are studied by Shapley and Scarf (1974), Roth and Postlewaite (1977) and Roth (1982b), lack even the bilateral 'matching' property of the other markets considered here. However, each trader in these markets has an initial endowment of a single unit of an individible commodity, so there is a sense in which each trader is 'identified' with his commodity. Kaneko and Wooders (1982) study the question of how the existence of stable outcomes is related to two-sidedness, by considering what other restrictions on which coalitions can productively form will also yield a non-empty core for every characteristic function.

4.2. Conflict and coincidence of interest

Perhaps the most striking feature of these two-sided matching markets is the way the apparent patterns of common and conflicting interest among and between the agents on different sides of the market are reversed when we confine our attention to the set of stable outcomes. When we look at the entire set of feasible outcomes (unstable as well as stable), it is natural to think of students as competing with one another for desirable places in hospitals, while hospitals compete with one another for desirable students. Thus while agents on opposite sides of the market have a common interest in arranging matches with one another, agents on the same side of the market have interests that conflict at least to some degree.[19] The following two results show how dramatically the situation changes when we look at the common and conflicting interests of the agents on the set of stable outcomes.

For a given preference profile P, a student s and a hospital h will be called *achievable* for one another if there is some stable outcome at which s is employed by h [i.e., if there is an outcome x in $S(P)$ such that $x(s)=h$]. For each h_i, let r_i be the maximum number of students employed at any stable outcome ($r_i \leqq q_i$).

Theorem 4.2a. In the set $S(P)$ of stable outcomes, there is a Hospital-optimal stable outcome x^* with the property that every hospital is assigned its r_i most-preferred achievable students, and a Student-optimal stable outcome y^* with the property that every student is assigned his most-preferred achievable hospital. (The NIMP algorithm selects the H-optimal stable outcome.)

Theorem 4.2b. At the H-optimal stable outcome x^*, each student is matched with his least-preferred achievable hospital, while at the S-optimal stable outcome y^*, each hospital is matched with its r_i least preferred achievable students.

Since hospitals' preferences are responsive, this implies that all hospitals agree that x^* is the most preferable stable outcome and y^* the least preferable, while students all agree that y^* is the best and x^* the worst.[20] Thus, contrary to the picture painted above of students competing with one another for the best hospitals and hospitals for the best students, when we

[19]This conflict is greatest the more agents on the same side of the market have common preferences over agents on the other side. An extreme example is when all students have the same first-choice hospital, and all hospitals have the same first-choice student. In this case all students are competing for the same position, and all hospitals for the same student.

[20]Because of the definition of r_i used here, these results are a little different from the similar results proved in Roth (1984b), but in view of Theorem 4.3a they are equivalent. The earliest results of this form for the marriage problem (in which all $r_i=1$) are found in Gale and Shapley (1962) and Knuth (1976), respectively, while similar results for a sidepayment model are found in Shapley and Shubik (1972).

look at stable outcomes, hospitals have a common interest in securing x^* and avoiding y^*, and students have exactly the opposite preference in common. To put it another way, when we ask students and hospitals to consider only achievable matches, the apparent common interest between the two sides vanishes, as does the apparent conflict of interest among the agents on each side.[21]

Similar optimal stable outcomes for each side of the market have been observed in each of the more general two-sided matching markets referred to in section 4.1. The following result gives some insight into why such optimal stable outcomes exist in the hospital–intern market.[22] For any hospital h with quota q, let its *choice set* from any set T of acceptable students be h's q most preferred students in T if $|T| \geqq q$, and be the entire set T if $|T| \leqq q$.

Theorem 4.2c. Let x and y be outcomes in $S(P)$. Then there is a feasible outcome z at which each student s is matched with whichever of $x(s)$ or $y(s)$ he prefers. Similarly, there is a feasible outcome w at which each hospital h is matched with its choice set from $\{x(h) \cup y(h)\}$. Furthermore, both z and w are stable; i.e., both are in $S(P)$.

Theorem 4.2c says the following. Take any two stable outcomes x and y, and invite each student s to point to whichever of the two assignments $x(s)$ or $y(s)$ he prefers. Then *no hospital is pointed to by more than its quota of students*, and so there is a feasible outcome z at which every student is assigned to the hospital he pointed to. Similarly, when the hospitals are invited to point to their most preferred students from among those assigned to them at either x or y, *no two hospitals point to the same student*, so there is a feasible outcome w at which each hospital is assigned the students it pointed to. Finally, both these outcomes are stable.

This means that, for every two stable outcomes x and y, we have a way of producing stable outcomes z and w such that there is a consensus among the students that z is as good[23] or better than both x and y, and a consensus among the hospitals (since they have responsive preferences) that w is as good or better than both x and y. Furthermore, it is straightforward to show, the hospitals all agree that z is at least as bad as the worst of x and y, and the students all agree that w is at least as bad as the worst of x and y.

[21] In the example discussed in fn. 19, in which each side has a common first preference, the first choice student is matched with the first choice hospital at every stable outcome, so that other hospitals are not 'really' involved in competition for him in a market that yields stable outcomes.

[22] A version of the following property, called the 'consensus property' in Roth (1984d), was noted for the marriage problem by Knuth (1976), who attributed it to John Conway, and also observed in the sidepayment model of Shapley and Shubik (1972).

[23] A student who is assigned to the same hospital (or who is unmatched) at both x and y will of course get the same assignment at z.

Thus we have a way to build up the two optimal stable outcomes described in the previous theorems. If x is not the H-optimal stable outcome, there must be another stable outcome y that some hospital prefers, but we can find w that is better still for all hospitals (and worse for all students). Proceeding in this way, the outcome w is eventually the hospital-optimal stable outcome, and the worst for the students.

Note that the agents on one side of the market do not have identical interests over the set of stable outcomes — students, for example, may disagree over which of two stable outcomes x or y is more desirable — but in that case all students can agree that there is a third stable outcome z that it is their *common* interest to pursue. Similarly, although hospitals may not agree on which of x or y is less desirable, they have a common interest in avoiding z (and in pursuing w). In the language of abstract algebra, this pattern of common interests makes the set of stable outcomes a *lattice* under the partial ordering of the common interests of the students or the hospitals. The optimal stable outcome for one side of the market is the highest element of this lattice in the common preferences of the agents on that side of the market, and the lowest element in the common preferences of the agents on the other side of the market.[24]

The consensus property and its associated lattice structure give the appearance of providing a good explanation for the existence of optimal stable outcomes for each side of the market. But it was shown in Roth (1984d) that in more general models, such as that of Kelso and Crawford (1982), the consensus property fails to hold, even though optimal stable outcomes continue to exist for both sides of the market. At least part of the resulting mystery appears to have been cracked by Blair (1984b), who shows that at least some of the lattice structure carries over to these more general models.

The final result of this subsection shows that there is a weak sense in which the student-optimal stable outcome exhausts the *common* interests of the students, although this is not true of the hospitals and the hospital-optimal stable outcome (except in the special case of the marriage problem, when the two sides of the market are symmetric).

Theorem 4.2d. There is no outcome y (stable or not) that every student strictly prefers to the student-optimal stable outcome y^*. However, there may exist unstable outcomes x that every hospital strictly prefers to the hospital-optimal stable outcome x^*. (But no such x exists in the special case of the marriage problem.)

[24]Knuth (1976) further observed that the set of stable outcomes of the marriage problem falls in the class of what are called *distributive* lattices, and Blair (1984a) showed that no further refinement can be found, since all distributive lattices can be generated as a set of stable outcomes.

The first part of this theorem was proved in Roth (1982a), where it was also shown that there may exist an unstable outcome y that gives some students the same assignment as y^*, and gives all other students strictly preferable assignments. This is why the common interests of the students can only be said to be exhausted at y^* in a weak sense. The latter part of the theorem was proved in Roth (1984e), and is the first result so far mentioned here that illustrates a property of the marriage problem that does not generalize to the hospital intern market.

4.3. Further questions raised by the medical market

In any theoretical study of a set of phenomena as complicated as a real market, some questions arise directly out of the study of the market, some arise unexpectedly from the model, and many fall somewhere in between. Of the theorems considered above, Theorems 4.1. and 4.2a and b answer questions from this latter class, while Theorems 4.2c and d, and the associated algebraic results, answer questions that would probably not have been asked except in the context of a formal model. The following two results address questions about the medical market that were first raised in some form in the medical literature.

The first of these arose when, on the basis of experience and some simple examples, it was noted that the NIMP algorithm seemed to give some sort of advantage to hospitals over students (recall Theorems 4.2a and b). In defense of the NIMP algorithm, it was noted that certain rural hospitals failed to fill their full quota of positions even under the current system, and any attempt to shift the advantage from hospitals to students would surely worsen this situation. The following theorem shows that, as long as we consider only procedures yielding stable outcomes, this is not correct.[25]

Theorem 4.3a. When all preferences are strict, the set of hospital positions filled is the same at every stable outcome, as is the set of students who are assigned positions.

The second of these questions concerns the way to handle married couples who wish to go through the matching process together to ensure that they will both receive positions in the same geographic area. The procedure used to accommodate the increasing numbers of such couples during most of the years covered here was observed to give both couples and certain hospital directors the incentive to try to arrange matches outside of the centralized

[25]References to the relevant medical literature, as well as the proofs to these two theorems, can be found in Roth (1984b).

system; i.e., it was observed to produce unstable outcomes.[26] It was proposed that new procedures be explored to correct this, and indeed a new procedure has been implemented with this in mind. However the following result shows that there is a limit to how far any procedure can correct this problem.

Theorem 4.3b. In a market in which some agents are couples, the set of stable outcomes may be empty.

The results in the next section, concerning incentives, also arises in answer to claims made in the medical literature.

5. Incentives

As mentioned earlier, the first algorithm proposed for a centralized matching process in the hospital–intern labor market was discarded when it was observed that it gave some agents an incentive to state a rank-ordering different from their true preferences. In particular, it was noted that students might have an incentive not to rank first their true first choice. The NIMP algorithm was proposed as a replacement to solve this problem, and the instructions distributed to participants in the matching process claim that no student or hospital can ever achieve a better assignment by submitting a rank-ordering different from the true preferences.

Define a *stable matching procedure* to be a function from the set of all preference profiles P (or $P^{\#}$) to the set of stable outcomes $S(P)$. The adoption of any particular matching procedure creates a non-cooperative game, in which the strategies of the agents are the possible rank orderings they might submit. The following theorem, from Roth (1982a), shows that the above claim is not correct.

Theorem 5.1. No stable matching procedure exists that makes it a dominant strategy for all agents to state their true preferences.

The theorem says that no procedure that yields stable outcomes in the stated preferences can have the property that no agent can ever improve his match by misstating his preferences.[27]

It should be noted, however, that although the NIMP algorithm ran up against the limits of the possible in trying to remove all incentives for agents

[26]This problem was observed to arise from the fact that couples have preferences over pairs of positions that may not be representable in terms of their individual preferences over single positions.

[26]This impossibility theorem holds even on the restricted domain of the marriage problem. It was independently proved by Bergstrom and Manning (1982).

to misrepresent their preferences, it did represent an improvement over the algorithm it replaced.[28]

Theorem 5.2. The NIMP algorithm gives no student any incentive to misrepresent this true first choice.

The incentive situation for students would be better still if a procedure yielding the student-optimal stable outcome in terms of the stated preferences were employed. The following theorem was proved in Roth (1984e).

Theorem 5.3. A procedure that yields the student-optimal stable outcome y^* (in terms of the stated preferences) makes it a dominant strategy for each student to state his true preferences. However, a procedure yielding the hospital-optimal stable outcome x^* does not make it a dominant strategy for all hospitals to state their true preferences (except in the special case of the marriage problem).

Note that for the special case of the marriage problem, the theorem says that a procedure that yields the optimal stable outcome for one side of the market makes it a dominant strategy for agents on that side of the market to state their true preferences. This was proved for the marriage problem in Roth (1982a) and Dubins and Freedman (1981).[29] While the proof of Theorem 5.3 in Roth (1984e) shows that the NIMP algorithm does not make it a dominant strategy for hospitals to state their true preferences, it does not shed any light on how often, or under what circumstances and in what manner it would be profitable for a hospital to submit a rank-ordering different from its preferences. Some work aimed at these questions has been begun by Wood (1984).

Given that agents may have the incentive to submit rank-orderings that differ from their true preferences, we must consider whether it is still reasonable to regard the NIMP algorithm as yielding an outcome that is stable with respect to the *true* preferences. That is, if P^* represents the vector of agents' true preferences over the agents on the other side of the market, and P is the vector of stated preferences, the fact that the NIMP algorithm

[28]The statement of the following theorem is from Roth (1984b), but it was proved in Roth (1982a) in connection with an equivalent algorithm.

[29]Related results are obtained in markets with sidepayments by Demange (1982) and Leonard (1983). Dubins and Freedman actually proved a slightly stronger result for the marriage problem: a procedure that yields the optimal stable outcome for one side of the market gives no *group* of agents on that side of the market the opportunity to all strictly profit by misrepresenting their preferences. However, this group result does not generalize to models that allow any sidepayments to be made among agents on the same side of the market, because it may be possible for a group to obtain an outcome that all members at least weakly prefer to the truth-telling outcome, and then to make transfers among themselves that leave them all strictly better off than at the truth-telling outcome. (See the paragraph after Theorem 4.2d).

produces an outcome in $S(P)$ does not ensure that this outcome is in $S(P^*)$. Nevertheless, the orderly operation of the market following the introduction of the NIMP algorithm makes it extremely plausible that the outcome is indeed in $S(P^*)$. This question may be addressed in two different ways.

First, if (most) agents state their true preferences, then the stated preferences and true preferences coincide, and no problem arises. This could be expected to be the state of affairs if either the incentives to misrepresent occur only rarely, or (more likely) if the kind of information needed about other agents' preferences to know how to profitably and safely misrepresent one's true preferences is of a more detailed sort than is readily available.

Alternatively, if we think of the agents in this market as highly rational and well informed, then we would expect that the vector P of rank order lists they submit would constitute a Nash equilibrium — i.e., it would exhaust the possibilities for further profitable manipulations. In this case P would typically differ from P^*, but the following theorem shows that the final outcome produced by the NIMP algorithm could still be expected to be stable with respect to the true preferences.[30]

Theorem 5.4. There exist Nash equilibria P of submitted rank order lists such that the outcome produced by the NIMP procedure is stable with respect not only to P, but also to the true preferences P^*. That is, the NIMP outcome is contained in $S(P^*)$ as well as in $S(P)$.

6. Extensions and open questions

A number of theoretical questions are raised by the surprising structure of the set of stable outcomes in this kind of market. Whenever similar results can be obtained for a variety of related models, as is the case here, it seems likely that some general theory exists that would clarify the origin and generality of these results. Certainly no theory of this kind has yet been proposed that is adequate to explain, for example, in what class of discrete markets the set of stable outcomes is non-empty, and when there will exist optimal stable outcomes of the kind found here. More generally, examination of the set of stable outcomes reveals that the markets discussed here exhibit a very pronounced degree of common interests among agents on one side of the market and conflicting interests between agents on opposite sides of the market, that is not at all apparent when looking at the entire set of feasible outcomes. This permits us to make welfare comparisons between different equilibrium outcomes — e.g., there is one that is best for the hospitals and worst for the students — of a kind that are unavailable in most markets. Perhaps it will be fruitful to investigate in what other kinds of markets, and

[30]Related results are found in Roth (1984c, e) and Gale (1983).

for what other notions of equilibrium, will the structure of the market be mirrored in this way in the set of equilibrium outcomes. And, of course, the difficulties observed here with married couples are likely to show up in various forms in increasing numbers of labor markets, as two-career households become more commonplace in the population, so it would be useful to understand more about the causes, frequency, and consequences of the associated instability.

There are a great many practical questions that remain unanswered about the opportunity for an agent to profitably misrepresent his preferences in such markets. The theorems discussed above indicate only that such opportunities exist, but do not speak at all to the questions of how often, and in what circumstances, using what information, such strategic manipulations may be made. Participants I have encountered in such markets often express an urgent interest in the answers to these questions.

Finally, there are empirical questions about the variety of institutions that have developed to mediate two-sided matching problems, which would appear to arise in a moderately wide variety of circumstances. (Note that many of the results discussed here apply as well to decentralized markets as they do to the quite centralized clearinghouse employed in the particular labor market discussed here.) Institutions that involve some degree of centralization have arisen in certain British medical labor markets, in the matching of undergraduates to American sororities, in the matching of Fulbright Scholars to host universities, in the assignments of officers and enlisted personnel in the American armed forces, and undoubtedly in the internal personnel assignments of many firms, and the practices of union hiring halls.[31] Less centralized institutions abound in a variety of more or less well-defined labor markets, particularly those involving workers seeking employment at approximately the same time — e.g., new graduates. Whether these markets reach stable outcomes, and if so how, and if not with what consequences, are all questions that will require a more detailed understanding of particular cases.

Appendix 1: The NIMP Algorithm

The algorithm is described as follows in Roth (1984b):

> 'The NIMP algorithm (cf. Stalnaker, 1953; Darley, 1959; NIRMP Directory, 1979) works as follows. Each hospital program rank orders the students who have applied to it (marking "X" any students who are unacceptable) and each student rank orders the hospital programs to

[31] I would be glad to learn the details of any matching procedures that readers might be acquainted with.

which he has applied (similarly indicating any which are unacceptable). These lists are mailed to the central clearinghouse, where they are edited by removing from each hospital program's rank-order list any student who has marked that program as unacceptable, and by removing from each student's list any hospital which has indicated he is unacceptable. (I am indebted to J.S. Graettinger, M.D., for clarifying this initial editing.) The edited lists are thus rank orderings of acceptable alternatives.

These lists are entered into what may be thought of as a list-processing algorithm consisting of a matching phase and a tentative-assignment-and-update phase. The first step of the matching phase (the 1:1 step) checks to see if there are any students and hospital programs which are top-ranked in one another's ranking. (If a hospital h_i has a quota of q_i then the q_i highest students in its ranking are top-ranked). If no such matches are found, the matching phase proceeds to the 2:1 step, at which the second ranked hospital program on each student's ranking is compared with the top-ranked students on that hospital's ranking. At any step when no matches are found, the algorithm proceeds to the next step, so the generic k:1 step of the matching phase seeks to find student–hospital pairs such that the student is top-ranked on the hospital's ranking and the hospital is kth ranked by the student. At any step where such matches are found, the algorithm proceeds to the tentative-assignment-and-update phase.

When the algorithm enters the tentative-assignment-and-update phase from the k:1 step of the matching phase, the k:1 matches are tentatively made; i.e., each student who is a top-ranked choice of his kth choice hospital is tentatively assigned to that hospital. The rankings of the students and hospitals are then updated in the following way. Any hospital which a student s_j ranks lower than his tentative assignment is deleted from his ranking (so the updated ranking of a student s_j tentatively assigned to his kth choice now lists only his first k choices) and student s_j is deleted from the ranking of any hospital which was deleted from s_j's ranking (so the updated rankings of each hospital now include only those applicants who haven't yet been tentatively assigned to a hospital they prefer). Note that, if one of a hospital's top-ranked candidates is deleted from its ranking, then a lower-ranked choice moves into the top-ranked category, since the hospital's updated ranking has fewer students, but the same quota, as its original ranking. When the rankings have been updated in this way, the algorithm returns to the start of the matching phase, which examines the updated rankings for new matches. Any new tentative matches found in the matching phase replace prior tentative matches involving the same student. (Note that new tentative matches can only improve a student's tentative assignment, since all lower ranked hospitals have been deleted from his

ranking.) The algorithm terminates when no new tentative matches are found, at which point tentative matches become final. That is, the algorithm matches students with the hospitals to which they are tentatively matched when the algorithm terminates. Any student or hospital position which was not tentatively matched during the algorithm is left unassigned, and must make subsequent arrangements by directly negotiating with other unmatched students or hospitals.'

(See fig. 1 for a schematic of the algorithm.)

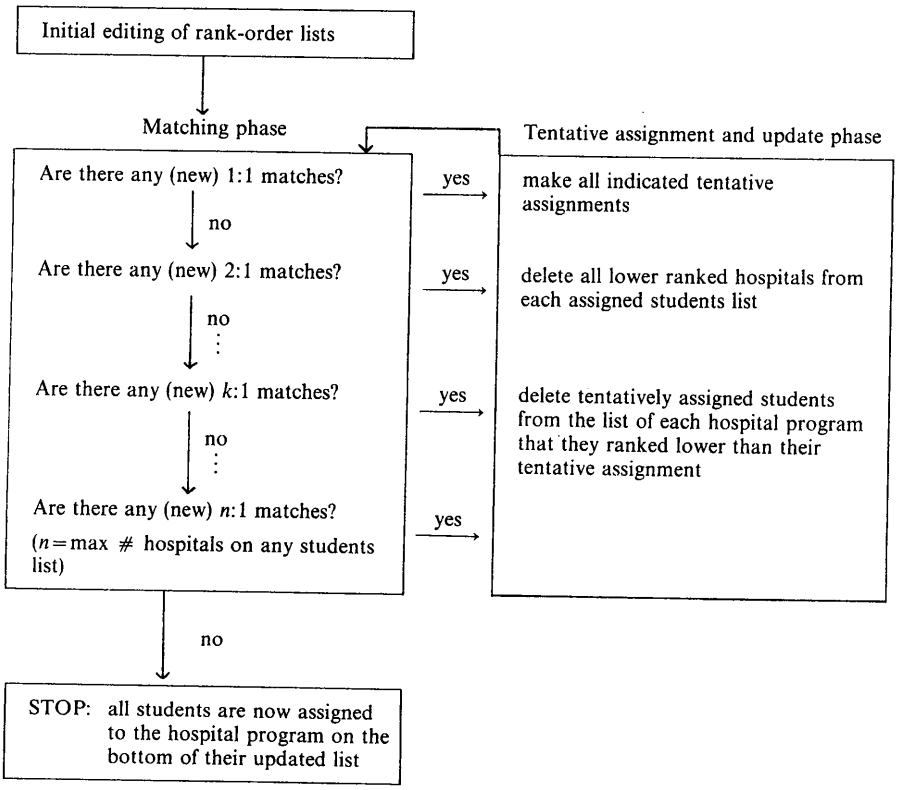

Fig. 1. The NIMP algorithm.

Appendix 2: Stability, group stability, and the core

Throughout this appendix, an *outcome* will be as defined in section 3.

An outcome x will be called *group unstable* if there exists an outcome y and a coalition $A \subset H \cup S$ such that, for all students s in A, and for all

hospitals h in A,

$$y(s) \in A \cap H, \qquad y(s)P(s)x(s),$$

$$y(h) \subset \{A \cap S\} \cup x(h) \quad \text{and} \quad y(h)P^{\#}(h)x(h).$$

That is, x is group unstable if there exists some coalition A of hospitals and students that, by matching among themselves, could give each student and hospital in A an assignment preferable to x. Note that the definition allows a hospital to include among its interns at the new outcome y some of those assigned to it at x, who are therefore indifferent between x and y.

A *group stable* outcome is an outcome that is not group unstable. It turns out that this definition of group stability is equivalent to the (pairwise) definition of stability given in section 3.

Theorem A2.1. An outcome is group stable if and only if it is stable.

Proof. If x is (pairwise) unstable via a student s and hospital h, then it is group unstable via the coalition $A = \{s\} \cup \{h\}$ and y such that $y(s) = h$ and $y(h) = x(h) \cup \{s\} \setminus \sigma$ for σ in $x(h)$ such that $sP(h)\sigma$. In the other direction, if x is group unstable via coalition A and outcome y, let h be in $A \cap H$. Then the fact that $y(h)P^{\#}(h)x(h)$ implies that there exist students s in $y(h)$ and σ in $x(h)$ such that $sP(h)\sigma$. [Otherwise, $\sigma R(h)s$ for all σ in $x(h)$ and s in $y(h)$, and since $x(h) \neq y(h)$ and preferences are strict, $\sigma P(h)s$ for at least one such σ and s. But since $P^{\#}(h)$ is responsive, this would imply $x(h)P^{\#}(h)y(h)$.] Since s prefers h to $x(s)$, s is unstable via s and h. Q.E.D.

We now turn to the relationship between the set $S(P)$ of stable outcomes and the core of the game. Following conventional game-theoretic usage, an outcome y is said to *dominate* another outcome x via a coalition $A \subset H \cup S$ if for all students s in A, and for all hospitals h in A,

$$y(s) \in A \cap H, \qquad y(s)P(s)x(s),$$

$$y(h) \subset \{A \cap S\}, \quad \text{and} \quad y(h)P^{\#}(h)x(h).$$

Similarly, an outcome y is said to *weakly dominate* x via a coalition $A \subset H \cup S$ if for all students s in A, and for all hospitals h in A,

$$y(s) \in A \cap H, \qquad y(s)R(s)x(s),$$

$$y(h) \subset \{A \cap S\} \qquad y(h)R^{\#}(h)x(h),$$

and either

$$y(s)P(s)x(s) \quad \text{for some} \quad s \text{ in } A \cap S, \quad \text{or}$$

$$y(h)P^{\#}(h)x(h) \quad \text{for some} \quad h \text{ in } A \cap H.$$

That is, if y dominates x via A, then every member of the effective coalition A strictly prefers y to x, while if y weakly dominates x via A, then every member of A likes y at least as much as x, and at least one member of A strictly prefers y to x.

The *core* of the game, $C(P)$ is the set of outcomes that are not dominated by any other outcome. The *core defined by weak domination*, $C_W(P)$, is the set of outcomes that are not weakly dominated by any other outcome. Since domination implies weak domination, $C_W(P)$ contains $C(P)$. When preferences are strict, the two cores coincide in the marriage problem, but not in the general hospital–intern market. However, when preferences are responsive, the set of stable outcomes coincides with the core defined by weak domination.

Theorem A2.2. $S(P) = C_W(P)$.

Proof. If x is not in $S(P)$ then x is unstable via some student s and hospital h with $sP(h)\sigma$ for some σ in $x(h)$. Then x is weakly dominated via the coalition $h \cup s \cup x(h) \setminus \sigma$ by any outcome y with $y(s) = h$ and $y(h) = s \cup x(h) \setminus \sigma$. In the other direction, if x is not in $C_W(P)$, then x is weakly dominated by some outcome y via a coalition A. If, for some h in $A \cap H$, h is indifferent between $y(h)$ and $x(h)$, then $x(h) = y(h)$, since $P^{\#}(h)$ is strict. This implies that x is also weakly dominated by y via the coalition $A' = A \setminus \{h \cup x(h)\}$, and by repetition if necessary, x is dominated by y via a coalition A such that every hospital h in $A \cap H$ strictly prefers $y(h)$ to $x(h)$. But $P^{\#}(h)$ is responsive, so $y(h)P^{\#}(h)x(h)$ implies that there exists a student s in $y(h)$ but not in $x(h)$ such that $sP(h)\sigma$ for σ in $x(h)$ (see the previous proof). So x is unstable via h and s. Q.E.D.

References

Bergstrom, Theodore and Richard Manning, 1982, Can courtship be cheatproof? Mimeo.

Blair, Charles, 1984a, Every finite distributive lattice is a set of stable matchings, Journal of Combinatorial Theory, forthcoming.

Blair, Charles, 1984b, Stable matching with multiple partners and lattice structure, Mimeo. (University of Illinois, Urbana, IL).

Crawford, Vincent P. and Elsie Marie Knoer, 1981, Job matching with heterogeneous firms and workers, Econometrica 49, 437–450.

Darley, Ward, 1959, The seventh annual report of the National Intern Matching Program, Journal of Medical Education 34, 38–46.

Demange, Gabrielle, 1982, Strategyproofness in the assignment market game, Mimeo. (Laboratoire d'Econometrie de l'Ecole Polytechnique, Paris).

Demange, Gabrielle and David Gale, 1983, A strategy-proof allocation mechanism for two-sided matching markets, Mimeo., presented at the IMSSS-economics workshop (Stanford University, Stanford, CA).

Dubins, L.E. and D.A. Freedman, 1981, Machiavelli and the Gale–Shapley algorithm, American Mathematical Monthly 485–494.

Gale, David, 1983, Ms. Machiavelli and the Gale–Shapely algorithm, Mimeo.

Gale, David and Llyod Shapley, 1962, College admissions and the stability of marriage, American Mathematical Monthly 69, 9–15.
Jones, Philip C., 1983, A polynomial time market mechanism, Journal of Information and Optimization Sciences 4, 193–203.
Kaneko, Mamoru and Myrna Holtz Wooders, 1982, Cores of partitioning games, Mathematical Social Sciences 3, 313–327.
Kelso, A.S., Jr. and V.P. Crawford, 1982, Job matching, coalition formation, and cross substitutes, Econometrica 50, 1483–1504.
Knuth, Donald E., 1976, Mariages stabiles (Les Presses de l'Université de Montréal, Montreal).
Leonard, Herman B., 1983, Elicitation of honest preferences for the assignment of individuals to positions, Journal of Political Economy 91, 461–479.
McVitie, D.G. and L.B. Wilson, 1971, The stable marriage problem, Communications of the ACM 14, 486–492.
NIRMP directory, National Resident Matching Program 1979 (Evanston, IL).
Quinzii, Martine, 1984, Core and competitive equilibria with indivisibilities, International Journal of Game Theory 13, 41–60.
Rochford, Sharon C., 1984, Market power in assignment markets, Mimeo. (Department of Economics, University of Georgia, Athens, GA).
Roth, Alvin E., 1982a, The economics of matching: Stability and incentives, Mathematics of Operations Research 7, 617–628.
Roth, Alvin E., 1982b, Incentive compatibility in a market with indivisible goods, Economics Letters 9, 127–132.
Roth, Alvin E., 1984a, Stability and polarization of interests in job matching, Econometrica 52, 47–57.
Roth, Alvin E., 1984b, The evolution of the labor market for medical interns and residents: A case study in game theory, Journal of Political Economy, forthcoming.
Roth, Alvin E., 1984c, Misrepresentation and stability in the marriage problem, Journal of Economic Theory 34, 383–387.
Roth, Alvin E., 1984d, Conflict and coincidence of interest in job matching: Some new results and open questions, Mathematics of Operations Research, forthcoming.
Roth, Alvin E., 1984e, The college admissions problem is not equivalent to the marriage problem, Journal of Economic Theory, forthcoming.
Roth, Alvin E. and Andrew Postlewaite, 1977, Weak versus strong domination in a market with indivisible goods, Journal of Mathematical Economics 4, 131–137.
Shapley, Lloyd S. and Herbert Scarf, 1974, On cores and indivisibility, Journal of Mathematical Economics 1, 23–28.
Shapley, Lloyd S. and Martin Shubik, 1972, The assignment game I: The core, International Journal of Game Theory 1, 111–130.
Stalnaker, John M., 1953, The matching program for intern placement: The second year of operation, Journal of Medical Education 28, 13–19.
Wood, Robert O., 1984, A note on incentives in the college admissions market, Mimeo. (Stanford University, Stanford, CA).