Econ Theory (1991) 1: 31–44

Economic Theory

Incentives in two-sided matching with random stable mechanisms

Alvin E. Roth and John H. Vande Vate

Faculty of Arts and Sciences, Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received: February 20, 1990

Summary. This paper considers the incentives confronting agents who face the prospect of being matched by some sort of random stable mechanism, such as that discussed in Roth and Vande Vate (1990). A one period game is studied in which all stable matchings can be achieved as equilibria in a natural class of undominated strategies, and in which certain unstable matchings can also arise in this way. A multi-period extension of this game is then considered in which all subgame perfect equilibria must result in stable matchings. These results suggest avenues to explore markets in which matching is organized in a decentralized way.

1. Introduction

The empirical study of two-sided markets as matching processes has so far concentrated on markets in which centralized matching procedures were introduced at some point in the market's history. Loosely speaking, these studies suggest that market organizers turn to centralized procedures to address certain market failures (such as uncontrolled unravelling of appointment dates, and chaotic recontracting), and that those centralized procedures which achieved stable outcomes resolved the market failures that inspired them, while those markets organized through centralized procedures that yielded unstable outcomes continued to fail[1].

[1] Roth (1984a) studies the American market for newly graduated physicians. Prior to 1951 that market experienced a number of failures, having to do with the difficulty of setting uniform dates of appointment, and with the frequency with which contracts were broken. In 1951 a centralized matching procedure that produces stable matchings was adopted, which is still in use, and which resolved these problems. Roth (1990a, b) studies the various different entry level markets for new physicians in the different regions of the National Health Service of the United Kingdom. In response to similar market failures in the late 1960's, centralized matching procedures were adopted in these markets also. But different procedures were adopted in different

Since many entry-level labor markets and other two-sided matching situations don't employ centralized matching procedures, and yet aren't observed to experience such failures, we can conjecture that at least some of these markets may reach stable outcomes by means of decentralized decision making. And decentralized decision making in complex environments will often introduce some randomness into what matchings are achieved (for example if the order in which offers are received depends on the vagaries of the post office, etc.). In this connection, a class of random processes that converge to stable outcomes with probability one was introduced in Roth and Vande Vate (1990).

One of the clearest lessons from the empirical study of centralized (and deterministic) procedures is that understanding the incentives facing the agents is essential for understanding the behavior of the market.[2] There is every reason to believe that the same will be true for decentralized and random processes, and the purpose of the present paper is therefore to begin to study equilibrium behavior for games arising from such processes.

However one of the difficulties that arises in attempting to apply theoretical studies of equilibrium to empirical studies is that the information required for agents to implement some kinds of equilibrium strategies frequently exceeds the information that agents can reasonably be thought to have. One potential course of action is to model such markets as games of incomplete information (see Roth 1989), but as an attempt to model observable markets this has problems of its own, because of the difficulty of observing agents' prior probability distributions. In the present paper we shall instead adopt the tactic of concentrating on a class of plausible, informationally parsimonious strategies that agents can employ even when they know only their own preferences, and that they have been observed to employ in at least one empirical study (namely Mongell and Roth 1990).

It is already known (Roth 1982) that no revelation mechanism both yields a stable matching with respect to the stated preferences and makes it a dominant strategy for all agents to state their true preferences. However it is also known (Roth 1984b) that the mechanism that always yields the optimal stable matching for one side of the market has the property that, although agents may have an incentive to misrepresent their preferences, every equilibrium in undominated strategies will yield a matching that is stable with respect to the true preferences[3]. That such a strong result can be obtained is due in large part to the fact that the restriction to undominated strategies has considerable force in that case. But for

---

markets, and those which produce stable matchings have succeeded in resolving the failures, and remain in use, while in all but the smallest markets those which did not produce stable matchings continued to experience the same problems as when the markets were decentralized, and these centralized schemes were ultimately abandoned. See also Mongell and Roth (1990) for a study of the preferential bidding system used by American sororities. See Crawford and Knoer (1981) and Kelso and Crawford (1982) for theoretical studies emphasizing the connection between two-sided matching models generally and labor markets.

[2] For example one of the matching procedures employed to match graduating medical students to jobs in some regions of the United Kingdom, analyzed in Roth (1990b), gave participants an incentive to pre-arrange matches in a certain way (and thus circumvent the procedure). Over 80% of the matches in one region were observed to be pre-arranged in this way prior to the abandonment of the procedure.

[3] This result applies to models of one-to-one matching of the kind considered here. The situation is more complex in models of many-to-one matching (see Roth 1985; and Roth and Sotomayor 1990, Chap. 5).

the class of random matching mechanisms considered here, we will see that relatively few strategies are dominated.

This paper is organized as follows. Section 2 presents a simple matching model, and describes a class of random mechanisms that, for any (stated) preferences converge to a matching that is stable with respect to those preferences. Section 3 then considers a one period revelation game with random stable matching, and observes that very few strategies are dominated. We then turn our attention to a class of plausible and informationally parsimonious undominated strategies called truncation strategies, and show that every stable matching can be achieved as an equilibrium in truncation strategies. Section 4 then considers a simple dynamic game in which every subgame perfect equilibrium yields a stable matching.

2. Random matching in the basic marriage model

We follow Gale and Shapley (1962) in considering the simplest two-sided matching model, known as the marriage problem, with equal numbers of agents on each side of the market[4]. The two sets of agents are $M=\{m_1,\ldots,m_n\}$ and $W=\{w_1,\ldots,w_n\}$, called "men" and "women", and each agent has a complete and transitive strict preference ordering over the n agents on the other side of the market. The preference ordering of a man m, for example, will be a permutation $P(m)$ of w_1 through w_n: if man m prefers w_i to w_j then w_i appears earlier in the list $P(m)$ than does w_j. Let $P\equiv\{P(m_1),\ldots,P(m_n),P(w_1),\ldots,P(w_n)\}$ denote the preference lists of all the agents, so a particular instance of the marriage model is specified by (M, W, P).

An outcome is a *matching* of men to women, i.e., a one-to-one function μ from $M\cup W$ to itself, such that for each m in M and w in W, $\mu(m)=w$ if and only if $\mu(w)=m$, and if $\mu(m)$ is not contained in W then $\mu(m)=m$, and similarly $\mu(w)=w$ if $\mu(w)$ is not contained in M. (If $\mu(m)=w$ then man m is matched to woman w, and if $\mu(m)=m$ then man m is single, or "unmatched".) For a given matching μ, a man m and a woman w are said to form a *blocking pair* if they are not matched to one another ($\mu(m)\neq w$) and if they each prefer one another to their mates at μ (w prefers m to $\mu(w)$ and m prefers w to $\mu(m)$). A matching μ is *stable* if there are no blocking pairs[5]. Note that at a stable matching of this very simple matching model no man or woman is left unmatched.

Gale and Shapley (1962) proved that, for any preferences of the agents, the set of stable matchings is non-empty. Knuth (1976) however, constructed an example in which there are cycles of blocking pairs, so that the process of satisfying blocking pairs in the following way may not lead to a stable matching. If (m', w') is a blocking pair for a matching μ, we say that a new matching ν is obtained from μ by *satisfying* the blocking pair if m' and w' are matched to one another at ν, their mates at μ (if any) are unmatched at ν, and all other agents are matched to the same mates at ν as they were at μ. That is, $\nu(m')=w'$, and

[4] There is now a large literature concerning this and many much more general models of two sided matching. See Blair (1988) for one of the most general of these, or see Roth and Sotomayor (1990) for a comprehensive account.

[5] In this simple model the set of stable matchings coincides with the core of the game. However we refer to stable matchings rather than core outcomes because in more general matching models this exact coincidence is lost.

for all m in M distinct from m' and $\mu(w')$, $\nu(m)=\mu(m)$, and if $\mu(w')=m$ for some m in M, $\nu(m)=m$. So Knuth's example showed that, starting from a given matching μ, a sequence of matchings $\{\mu_i\}$ such that μ_{i+1} is obtained from μ_i by satisfying some blocking pair for μ_i may cycle rather than converge to a stable matching.

Knuth's example raised the question of whether at least one path generated by satisfying blocking pairs exists from any matching to a stable matching, for any preferences of the agents. Roth and Vande Vate (1990) resolved this question in the affirmative. They went on to consider a random process which begins by selecting an arbitrary matching μ, and then proceeds to generate a sequence of matchings $\mu \equiv \mu_1, \mu_2, \ldots$, where each μ_{i+1} is derived from μ_i by satisfying a single blocking pair, chosen at random from the blocking pairs for μ_i. We assume the probability that any particular blocking pair (m, w) for the matching μ_i will be chosen to generate μ_{i+1} is positive, and bounded away from zero (i.e. this probability may be a function of the sequence up to i as well as of the matching $\mu=\mu_i$ itself, but it does not go to zero as i goes to infinity[6]. Let $R(\mu)$ be the random sequence generated in this way from an initial matching μ. The family of random processes beginning from an arbitrary matching and selecting a blocking pair at random to create a new matching will eventually reach a stable matching with probability one.

Theorem 1 (Roth and Vande Vate 1990). *For any initial matching μ, the random sequence $R(\mu)$ converges with probability one to a stable matching.*

3. Strategic considerations

With the exception of the impossibility result to be described in Proposition 3[7], theoretical investigations of strategic considerations in two sided matching have concentrated on deterministic processes, in which either some central clearinghouse arranges the matches, or else they are arranged by the agents acting according to some rigid set of rules designed to achieve a particular stable matching (typically the optimal stable matching for one side of the market)[8]. Similarly, the comparative ease of gathering reliable information about the "rules of the game" when the game involves a centralized matching process has made these the starting point of empirical work (see footnote 1). In contrast, it may be impossible to determine similarly exact rules for markets in which matching is conducted in a decentralized way, even though the behavior of many such markets suggests they achieve stable matchings. And decentralized decision making in complex environments presumably introduces some randomness into what matchings are achieved. We therefore turn to the general class of random processes discussed in Theorem 1. The idea is that the randomness arises from unmodelled (and

[6] The probability that a particular blocking pair (m, w) for a matching μ will be chosen might reflect, for example, factors such as the likelihood that individuals m and w would meet, and the number of other blocking pairs. If, for example, this probability were the same every time the matching μ arises in the sequence of matchings, the necessary condition would be met.

[7] Which has rather general implications, because of the "revelation principle", which states that equilibria of any mechanism can be achieved by revelation mechanisms.

[8] A notable exception is Kamecke (1989), who studies a decentralized "demand game" version of a matching model in which there is a continuously divisible commodity.

perhaps inherently uncertain) details of the way the game proceeds. For example, randomness might arise from details of the way in which players communicate with one another: there may be randomness in whether a player receives the next proposal someone is planning to make to him before he is able to deliver the next proposal he is otherwise planning to make.

To fix ideas, it may be helpful to think of the strategic decisions of the players, and the random process which follows, as being a model of the sorting process that precedes the actual signing of contracts. For example, workers make strategic decisions about which firms to apply to, and firms about which applicants to interview. The sequence of matchings that arise from the random processes of Theorem 1 can then be thought of as preliminary matchings (i.e. which firms would hire which workers if the negotiating process were arbitrarily stopped at different points in time) which converge to a final stable matching that would be the observable outcome. So for given (true) preferences of the agents, they are each faced with the strategic question of what (revealed) preferences to act on, when the final outcome will be a randomly selected matching that is stable with respect to the revealed preferences.

To consider strategic games in which players may choose not to reveal their true full preferences, we need to slightly generalize our definition of stability to include cases in which players may not reveal that other players are acceptable mates. That is, a player p with true preferences $P(p)$ may choose to behave as if his or her preferences were $Q(p) \neq P(p)$, and $Q(p)$ may not contain as many elements as $P(p)$. (For example, if m does not appear on w's preference list, w has stated that she prefers to remain unmatched rather than be matched to m. Formally, for every player p, the implicit last element of any preference list $Q(p)$ is assumed to be p.) We say that a matching μ is individually rational with respect to some (stated) preferences $\mathbf{Q} \equiv (Q(m_1), \ldots, Q(m_n), Q(w_1), \ldots, Q(w_n))$ if for every m and w such that $\mu(m) = w$, w appears on m's preference list $Q(m)$, and m appears on w's preference list $Q(w)$. A matching is stable if it is individually rational and there are no blocking pairs.

We will also need the following facts about stable matchings of a marriage problem $(M, W, \mathbf{Q})$ with general strict preferences of this kind[9].

Proposition 1. *For any $(M, W, \mathbf{Q})$ the set of stable matchings contains a matching μ_M that is optimal for the men in the sense that no other stable matching ν gives to any man m a mate $\nu(m)$ that he prefers to $\mu(m)$. Similarly, it contains a stable matching μ_W that is optimal for the women.*

Proposition 2. *For any $(M, W, \mathbf{Q})$ any player who is unmatched at some stable matching is unmatched at every stable matching.*

Proposition 3. *No revelation mechanism which chooses a stable matching in terms of stated preferences makes it a dominant strategy for all agents to state their true preferences. However the mechanism which chooses the optimal stable matching for one side of the market makes it a dominant strategy for agents on that side of the market to state their true preferences.*

[9] Proposition 1 is adapted from Gale and Shapley (1962), and Propositions 2 and 3 from Roth (1982). See Roth and Sotomayor (1990) for a full discussion.

We begin by considering the simple case of a one-period game of complete information[10], played by players $M \cup W$ with (true) preferences P, in which every player p states a preference list $Q(p)$ over players on the other side of the market, and then a matching stable with respect to the stated preferences $\mathbf{Q}$ is selected at random, with every stable matching having positive probability. So the strategy set of a player p in this game is the set of all possible preference lists, i.e. the set of all ordered lists consisting of between 0 and n players on the other side of the market.

Observe first that *any* matching μ can be achieved by an equilibrium in which each player p states the preference list $Q(p)$ that consists of the single element $\mu(p)$. But, unless $\mu(p)$ happens to be player p's first choice in his or her true preferences $P(p)$, this is a dominated strategy for each player. We can state this formally as follows.

Lemma 1. *For any player p, the strategy of stating a preference list $Q(p)$ consisting of a single element $\mu(p)$ is a dominated strategy unless $\mu(p)$ is p's true first choice (i.e. unless $\mu(p)$ is the first element in the true preferences $P(p)$).*

Proof. Since the model is symmetric between men and women, it is sufficient to prove the lemma for $p = w$ in W. Let m_1 be the first choice of woman w, and let $Q(w)$ consist of the single element $m \neq m_1$. We will show that the strategy $Q(w) = m$ is dominated by the strategy $Q'(w) = m_1, m$, i.e. the strategy of listing both m_1 and m. Let Q_{-w} denote the set of preferences stated by all players other than w. We need to show that for no preferences Q_{-w} will w ever do worse by stating $Q'(w)$ instead of $Q(w)$, and for some preferences Q_{-w} she will do better.

Proposition 2 implies that when a matching is selected at random from those stable with respect to stated preferences $\mathbf{Q} = (Q_{-w}, Q(w))$, player w either will be unmatched with certainty, or matched with man m with certainty. If w is unmatched at $\mathbf{Q}$ then she can certainly do no worse by stating $Q'(w)$, so suppose w is matched at $\mathbf{Q}$, i.e., suppose that w is matched to m at every stable matching with respect to $\mathbf{Q}$. Then (again by Proposition 2) the only way that w could do worse by stating $Q'(w)$ instead of $Q(w)$ would be if w were unmatched at every stable matching with respect to $\mathbf{Q}' = (Q_{-w}, Q'(w))$. But in this case w would be unmatched at the W-optimal stable matching with respect to $\mathbf{Q}'$, and Proposition 3 implies that this is not the case, since if it were then it would not be a dominant strategy for a player w whose true preferences were $Q'(w)$ to state them when the W-optimal stable matching would be chosen (i.e. such a player would prefer to state $Q(w)$). So there exists no Q_{-w} such that w does worse by stating $Q'(w)$ than by stating $Q(w)$.

To see that there exist some Q_{-w} for which w does strictly better by stating $Q'(w)$ rather than $Q(w)$, suppose $Q(m_1) = w$, i.e., man m_1 lists only woman w. Then at $\mathbf{Q}'$ w is matched to her true first choice with certainty, but at $\mathbf{Q}$ she has zero probability of being matched with him. This completes the proof of Lemma 1.

[10] Looking at the complete information case will allow us to separate the unavoidable strategic questions from the further complications which arise in the case of incomplete information. In matching with incomplete information, the strong impossibility results obtained in Roth (1989) via the revelation principle show that the situation will necessarily be more complex than the one we explore here.

Although Lemma 1 establishes that a class of equilibrium strategies is dominated, the next lemma shows that, in contrast to the stable mechanisms considered in Proposition 3, when we consider random stable mechanisms there are very few dominated strategies.

Lemma 2. *For any player p, the strategy $Q(p)$ is undominated if the first element in $Q(p)$ is player p's true first choice (i.e. the first element of $P(p)$).*

Proof. As in the previous lemma, it is sufficient to prove the lemma for some woman w. We need to show that if $Q(w)$ is a stated preference list that lists w's first choice, m_1, first, and if $Q'(w) \neq Q(w)$ is any other preference list, then there exist preferences Q_{-w} for the other players such that w prefers the lottery that results from $\mathbf{Q} = (Q_{-w}, Q(w))$ to that which results from $\mathbf{Q}' = (Q_{-w}, Q'(w))$. We will consider three cases.

Case 1. $Q'(w)$ doesn't list m_1 first. Then suppose man $m \neq m_1$ is listed first in $Q'(w)$, and let Q_{-w} be such that $Q(m_1) = Q(m) = w, w'$ for some woman w', and $Q(w') = m, m_1$. Suppose further that no other men list w or w', and no other women list m or m_1. Then it is straightforward to verify that w is matched to m_1 with probability one when the stated preferences are $\mathbf{Q}$, and that she is matched to m with probability one when they are $\mathbf{Q}'$, so she does better by stating $Q(w)$.

Case 2. $Q'(w)$ contains a different set of men than does $Q(w)$. If there is a man m in $Q(w)$ but not in $Q'(w)$, then let Q_{-w} be such that $Q(m) = w$ and no other man lists w. Then w is matched to m at $\mathbf{Q}$, but unmatched at $\mathbf{Q}'$. If there is a man m in $Q'(w)$ but not in $Q(w)$, let Q_{-w} be such that $Q(m_1) = w', w$ for some woman w', $Q(m) = w, w'$, and $Q(w') = m, m_1$. Then it is straightforward to show that w is matched to m_1 at every matching stable with respect to $\mathbf{Q}$ but is matched to m at the M-optimal stable matching with respect to $\mathbf{Q}'$, so that when w states $Q(w)$ she is certain to be matched to her first choice, but has a positive probability of doing worse when she states $Q'(w)$.

Case 3. $Q'(w)$ contains the same elements as $Q(w)$, but in a different order. Then there exits m and m' such that m is preferred to m' according to $Q(w)$, but m' is preferred to m according to $Q'(w)$. Let Q_{-w} be such that there is a woman w' having preferences $Q(w') = m', m_1, m$, and $P(m_1) = w', w$; $P(m) = P(m') = w, w'$; and no other agents list w, w', m, m' or m_1. Then it is straightforward to show that w is matched to m_1 under every stable matching with respect to $\mathbf{Q}$. But the M-optimal stable matching with respect to $\mathbf{Q}'$ matches w to m'. So when w states $Q(w)$ she is certain of being matched to her true first choice, but when she states $Q'(w)$ she has a positive probability of being matched to a less preferred choice. This completes the proof.

Lemma 2 implies that quite complex strategies may be undominated, so that concentrating on equilibria in undominated strategies does not rule out equilibria in which, say, every agent lists all the agents in the opposite set, but in an order quite different from their true order. Thus random stable matching mechanisms present technical difficulties of a kind not found in the strategic analysis of the (deterministic) mechanisms which select the optimal stable matching for one side of the market, in which case only one side of the market has many undominated strategies (cf. Gale and Sotomayor 1985; Roth and Sotomayor 1990, Theorem 4.21). But even if a complicated equilibrium $\mathbf{Q}$ of this sort exists, it is difficult

to imagine in most applications that any player p would have sufficient information about the preferences Q_{-p} of the other players to know that a strategy of this form would be a best response. And complex strategies are unlikely to be best responses to a wide range of possible preferences Q_{-p} that other players might state.

However we can identify a class of undominated strategies for each player, called truncations, with the property that for any strategies of the other players, each player will always have a truncation as a best response. Define a *truncation* of a player p's true preference list $P(p)$ to be a list $Q(p)$ containing k elements (for k between 0 and n) such that the k elements of $Q(p)$ are the first k elements of $P(p)$, in the same order. (Of course if $k=0$ the strategy is dominated.) Truncations are plausible kinds of strategies, that have been observed in practice[11]. We show next that truncations have the best response property mentioned above.

Theorem 2. *For any collection of stated preferences Q_{-p} for players other than an arbitrary player p, player p always has a best response that is a truncation of $P(p)$.*

Proof. Let $Q(p)$ be an arbitrary preference list for player p (not necessarily a truncation) and let μ be the optimal stable matching with respect to $\mathbf{Q}=(Q(p), \mathbf{Q}_{-p})$ for players on p's side of the market. Let $Q^*(p)$ be the truncation of $P(p)$ whose last element is $\mu(p)$. Then for no matching μ^* that is stable with respect to $\mathbf{Q}^*=(Q^*(p), \mathbf{Q}_{-p})$ is $\mu(p)$ preferred to $\mu^*(p)$. (This follows since otherwise p must be unmatched at some stable matching of $\mathbf{Q}^*$, and hence at every stable matching by Proposition 2. But then an agent whose true preferences were $Q^*(p)$ could do better by stating $Q(p)$ than by stating $Q^*(p)$ when the optimal stable matching for players on his side of the market will be chosen, which contradicts Proposition 3.) So any lottery over the matchings stable with respect to $\mathbf{Q}^*$ gives player p at least as good a final mate as any lottery over the matchings with respect to $\mathbf{Q}$. Since $\mathbf{Q}$ is arbitrary, this completes the proof.

For the game at hand, we can state the following results concerning truncation strategies.

Theorem 3. i. *Any stable matching can be achieved by equilibrium in undominated strategies via truncation strategies.*
ii. *No equilibrium in truncation strategies will have more than one stable matching in stated preferences.*
iii. *No set of truncation strategies can result in an unstable matching at which everyone is matched.*
iv. *But unstable matchings having some players unmatched can arise at equilibria in truncation strategies.*

Proof. Part (i): Let μ be any stable matching for $(M, W, \mathbf{P})$, and let $\mathbf{Q}$ be the vector of truncations in which each player p reveals his preferences down to $\mu(p)$ (i.e. $Q(p)$ contains $\mu(p)$ but does not contain any less preferred matches). Then μ is the unique stable matching with respect to $\mathbf{Q}$. To see that $\mathbf{Q}$ is an equilibrium, we need to show that no player p can profitably state $Q'(p)$ different from $Q(p)$ (where $Q'(p)$ need not be a truncation). But if ν is a stable matching with respect

[11] See Mongell and Roth, 1990, in which agents had an incentive to truncate after their first choice, and in which very high percentages of submitted preferences were observed to contain only a single choice.

to $\mathbf{Q}' = (Q'(p), \mathbf{Q}_{-\mathbf{p}})$ then the stability of μ with respect to $\mathbf{P}$ and the fact that $\mathbf{Q}_{-\mathbf{p}}$ are all truncations imply that $v(p)$ is not preferred to $\mu(p)$, since otherwise $(p, v(p))$ would have formed a blocking pair for μ. So $\mathbf{Q}$ is an equilibrium.

Part (ii). Suppose μ and v are both stable with respect to Q, with $\mu(p) \neq v(p)$ for some player p Then player p faces a lottery which gives positive probability to both $\mu(p)$ and $v(p)$. But if he stated a truncation $Q'(p)$ which ended at the optimal stable matching with respect to $\mathbf{Q}$ for agents on p's side of the market, he would face a lottery all of whose outcomes are at least as preferred as the best outcome from $\mathbf{Q}$. So $\mathbf{Q}$ is not an equilibrium.

Part (iii): Suppose $\mathbf{Q}$ is a vector of truncations, and μ is a stable matching with respect to $\mathbf{Q}$, and all players are matched at μ. Then every player p has listed $\mu(p)$ in $Q(p)$, together with all preferred players. So μ is stable with respect to $\mathbf{P}$, since any blocking pair under preferences $\mathbf{P}$ would also be a blocking pair under $\mathbf{Q}$.

Part (iv): Consider the example in which $n=2$ and the preferences are $P(m_1) = w_1, w_2$; $P(m_2) = w_1, w_2$; $P(w_1) = m_1, m_2$; $P(w_2) = m_1, m_2$. Let $\mathbf{Q}$ be the vector of stated preferences at which each player states only his or her first choice. Then the only stable matching with respect to $\mathbf{Q}$ is the matching $\mu = [(w_1, m_1), (w_2, w_2), (m_2, m_2)]$ that leaves w_2 and m_2 unmatched. Yet $\mathbf{Q}$ is an equilibrium, since w_2 cannot do better by listing m_2 so long as m_2 doesn't list w_2, and vice versa.

Part (iv) of Theorem 3, and its proof, suggest that it is the one period nature of the decentralized game under consideration that permits some unstable matchings to arise at equilibrium[12]. If, after stating truncated preferences and finding themselves unmatched, as in the proof of part (iv), agents had a further opportunity to extend their preferences, this kind of instability might be avoided. We turn now to consider such a multi-period game.

4. A simple dynamic model

Here we consider a multi-period analog of the above game. It can be thought of as a simple model of a situation in which players make proposals to each other, and rearrange themselves into potential matchings which are stable with respect to the preferences revealed by the proposals, until no player wishes to make any new proposals, at which point the last potential matching under consideration becomes the actual matching. The rules of the game are as follows.

The game begins with every player p declaring a list of mates he or she is willing to accept, by announcing a preference list $Q(p)$. This preference list must

[12] The parallel result (Roth, 1984b) for the centralized game which selects one side's optimal stable matching in terms of the stated preferences is that every equilibrium in undominated strategies yields a stable matching with respect to the true preferences (see also Roth and Sotomayor 1990).

be a truncation of the player's true preference list[13]. When all players have declared their acceptable mates in this way, a potential matching μ is selected, by the random process of Theorem 1. So the potential matching μ is stable with respect to the revealed preferences $\mathbf{Q} \equiv (Q(m_1), \ldots, Q(m_n), Q(w_1), \ldots, Q(w_n))$.

Following the selection of a potential matching μ, an order of all the players is selected at random, and the players have the opportunity, in order, to extend their revealed preference list to include one or more additional players to whom they are now willing to be matched. (Again, their extended preference list must be a truncation of their true preferences.) If no player wishes to extend his list, the game ends, and μ becomes the final matching of the game. If some player extends his revealed preference list, then a new potential matching is immediately chosen, by randomly choosing blocking pairs, starting from the matching μ, and stopping when the process converges to a matching ν that is stable with respect to the current set of revealed preferences. The game then continues according to the rules laid out in this paragraph. (Note that the game must stop in a finite number of moves, since each player can extend his preferences at most n times.)

The game is played under perfect as well as complete information: i.e. all moves, including chance moves, become common knowledge as soon as they occur. Players' preferences for outcomes of the game depend only on the final matching, not on the path by which it was reached. (Thus a given strategy $2n$-tuple determines a lottery over final matchings, and at each point in the game players can evaluate alternative decisions in terms of the expected utility of the resulting lotteries.) Players may condition their decisions at subsequent stages of the game on any of the events that precede their decision, starting from the beginning of the game as described above.

It turns out that for any marriage problem $(M, W, \mathbf{P})$ as described in Sect. 2, only stable matchings can arise as subgame perfect equilibria of the game just described. That is, we have the following result.

Theorem 4. *In the dynamic game played by players $M \cup W$ with preferences $\mathbf{P}$, only stable matchings μ can arise with positive probability as the final outcome of a subgame perfect equilibrium.*

The proof of Theorem 4 will proceed by a series of lemmas.

Lemma 4.1. *For a given $(M, W, \mathbf{P})$, if a matching μ arises as stable with respect to the revealed preferences at any stage of the dynamic game, μ is stable with respect to $\mathbf{P}$ if and only if no player is unmatched at μ.*

Proof. If μ leaves any players unmatched, then it is not stable with respect to $\mathbf{P}$, since there are equal numbers of men and women, all of whom are mutually acceptable. So suppose μ leaves no player unmatched, and let $\mathbf{Q}$ be the set of preferences revealed by the players. Then since μ is stable with respect to $\mathbf{Q}$, every player p has included $\mu(p)$ in his or her revealed preference list $Q(p)$. But since $Q(p)$ is a truncation of $P(p)$ it coincides with $P(p)$ for the mates player p prefers to $\mu(p)$, and so the fact that there are no blocking pairs for μ with respect to $\mathbf{Q}$ implies that there are none with respect to $\mathbf{P}$, and so μ is stable with respect to $\mathbf{P}$.

[13] So truncated strategies enter this game as part of the rules, and the results would be different if we allowed players to announce more complicated misrepresentations of their preferences (see e.g. footnote 14). This can be regarded as an attempt to model the information that the players have through a restriction on the complexity of the strategies they can effectively employ.

Lemma 4.2. *If* $\mathbf{Q}$ *is a set of truncated preferences,* μ *is a stable matching with respect to* $\mathbf{Q}$*, and the set of unmatched players is non-empty, then there must be an unmatched player* p *who has not fully revealed his or her preferences, i.e. for whom* $Q(p) \neq P(p)$.

Proof. Suppose the lemma is false, so that $Q(p) = P(p)$ for every player p who is unmatched at μ. Since there are equal numbers of men and women, there are the same number of unmatched men as unmatched women. But any unmatched m and w form a blocking pair for μ, since $P(m)$ contains w and $P(w)$ contains m. This contradicts the stability of μ.

Lemma 4.3. *If* μ *is the final matching resulting from some subgame perfect equilibrium, then* μ *leaves no player unmatched.*

Proof. Starting from the end of the game, suppose the preferences revealed so far are $\mathbf{Q}$, μ is the matching that will occur if no player extends his preferences, and there is exactly one unmatched player who has not fully revealed his preferences, player p. Then it is not a subgame perfect equilibrium for every player to refuse to extend his preferences, since if player p were to extend his preference list from $Q(p)$ to $P(p)$ then he or she would be matched at the final outcome, which is preferable to being unmatched. (This follows since Lemma 4.2 implies that player p will be matched at the new matching, ν, which is stable with respect to the true preferences, by Lemma 4.1. Since strategies must be truncations, no blocking pairs can appear at any subsequent stage of the game, so ν must be the final outcome of the game.)

Now suppose, inductively, that it has been shown that if there are exactly k unmatched players who have not fully revealed their preferences, then there is no subgame perfect equilibrium at which every player will refuse to extend his preference list, and consider the case when there are exactly $k+1$ such players. Then at a subgame perfect equilibrium it cannot be that no player will extend his preferences, since the inductive hypothesis insures that if one of them did, he would eventually be matched. This completes the proof of the lemma.

The proof of Theorem 4 is now immediate from Lemmas 4.1. and 4.3.

An open question that remains about the dynamic game described above is whether *every* stable matching can be achieved with positive probability as the outcome of a subgame perfect equilibrium. One reason that this seems to be a difficult question to settle is that best response strategies can be complex.

Consider the following example, with $n = 3$:

$$P(m_1) = w_1, w_2, w_3; P(m_2) = w_2, w_1, w_3; P(m_3) = w_3, w_1, w_2$$

$$P(w_1) = m_2, m_1, m_3; P(w_2) = m_1, m_2, m_3; P(w_3) = m_3, m_1, m_2.$$

There are two stable matchings, $\mu_M = [(m_1, w_1), (m_2, w_2), (m_3, w_3)]$ and $\mu_W = [(m_1, w_2), (m_2, w_1), (m_3, w_3)]$. We assume that the order in which players are given the opportunity to extend their preferences is chosen independently at each stage of the game, with positive probability for each ordering, so subgame perfect equilibria will give positive probability to both stable matchings. Note that if m_1 and m_2 both extend their preferences to include their first two choices before w_1 and w_2 do, the game will end with μ_W, while if w_1 and w_2 reveal their first two choices before m_1 and m_2, then the final matching will be μ_M. Suppose now that at some stage k of the game m_1 has revealed his first two choices and all other

players have revealed only their first choice, so that the potential outcome is $\mu_k = [(m_1, w_2), (m_2, m_2), (m_3, w_3), (w_2, w_2)]$ which leaves m_2 and w_2 unmatched. Suppose the random order selected for players to have the opportunity to extend their preferences is $w_3, w_2, w_1, m_3, m_1, m_2$. If no player has extended his preferences when m_2's turn comes, his best response will be to extend, and the game will end with the woman optimal stable matching $\mu_{k+1} = \mu_W$[14]. So if no player has extended when m_1's turn comes, his best response will be to extend, which will yield $\mu_{k+1} = \mu_k$ and a new random ordering, since this will leave a positive probability that the game will terminate in the man optimal stable matching.

Thus this example shows that it may be an error to be too choosy (by not extending your preferences soon enough), just as it can be an error not to be choosy enough. Consequently the strategic problems facing players in decentralized two-sided matching games are complex, even when those games are played under conditions that tend towards stability.

5. Concluding discussion

To put these results in perspective it may help to consider their relation to some of the prior work concerning centralized and deterministic matching mechanisms. In view of the impossibility result contained in the first part of Proposition 3, a question arose in explaining the observed success of stable matching mechanisms in comparison with unstable mechanisms in the various empirical studies already referred to. Specifically, the fact that a mechanism produces stable matchings with respect to stated preferences does not necessarily imply that the matching it produces will be stable with respect to the *true* preferences, since agents may have an incentive to misrepresent their preferences[15]. But if the matching is unstable (with respect to agents' true preferences) then some agents would still have an incentive to arrange their matches by circumventing the matching mechanism, just as in the case of the unstable mechanisms which have failed in the face of the problems this creates. Nevertheless, the stable mechanisms observed in practice have not been observed to fail[16].

The paper of Roth (1984b) provided a direction in which an answer to this puzzle might be sought, since it showed (for a simple matching model of the kind considered here) that at equilibria in undominated strategies, even though the stated preferences might differ substantially from the true preferences, the resulting matching would be stable with respect to the true preferences, when the optimal stable mechanism for one side of the market is employed. Thus stability could be achieved not only by straightforward behavior on the part of the agents, but also by sophisticated strategic behavior, and this suggested that stability in the actual markets studied might therefore be robust to a range of plausible behavior.

---

[14] Note that if the rules of the game allowed m_2 to state something other than a truncation, he could extend his stated preference to $Q(m_2) = w_2, w_3$ without causing the game to end, and leaving a positive probability that the man optimal stable matching would be chosen.

[15] And the incentive to misrepresent is endemic: it exists whenever there is more than one stable matching (Roth and Sotomayor 1990, Theorem 4.6).

[16] The particular stable mechanisms observed to date all chose the optimal stable matching for one or the other side of the market.

The results presented in this paper can be similarly interpreted, in connection with decentralized mechanisms that might give rise to the kind of random matching process considered here. That is, as particular observable markets in which matching is decentralized are studied, it will be necessary to model the matching process, either in detail as an extensive form game (which will very likely include random elements), or more abstractly. In either case, the nature of the outcomes that result from straightforward play and from sophisticated strategic play will have to be analyzed. Our results suggest that, although straightforward play will rarely be in equilibrium, there may continue to be a strong observed connection between the stability of a mechanism with respect to straightforward behavior and the observed stability of its market outcomes[17].

In closing, and in order to put the first step taken here in the context of the further work that suggests itself, it may be helpful to review the assumptions made in the simple models we have explored. The assumptions we have made about complete information are perhaps not quite as strong as they appear, in view of the fact that truncation strategies can be implemented even in low information environments. But by modelling decentralized matching as the kind of random process of Theorem 1, we are implicitly making strong assumptions about low transaction costs and speedy formation of blocking pairs[18], together with a limit on how the process may be history dependent[19]. Since these assumptions influence both the random process and the strategic problems facing the players, it remains important to analyze models in which they are relaxed.

Acknowledgements. This work has been supported by grants from the Alfred P. Sloan Foundation and the National Science Foundation. We have received helpful comments from Charles Blair and Jack Ochs.

References

Blair Ch (1988) The lattice structure of the set of stable matchings with multiple partners. Math Operations Res 13: 619–628

Crawford VP, Knoer EM (1981) Job matching with heterogeneous firms and workers. Econometrica 49: 437–450

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

Gale D, Sotomayor M (1985) Ms. Machiavelli and the stable matching problem. Am Math Monthly 92: 261–8

[17] In the other direction the connection may be less strong. The mechanism studied in Mongell and Roth (1990) does not produce a stable matching when agents behave straightforwardly, but does produce a stable matching when they behave strategically, and agents appear to have responded to the incentives to play strategically.

[18] Perhaps part of the market for new economics professors has features which approach some of these properties: following a large professional meeting at which information about candidates and jobs is exchanged, candidates quickly go through a series of interviews, receiving job offers and turning them down as better ones arrive.

[19] The limit on history dependence makes itself felt through the assumption that the probability that a given blocking pair will form is bounded away from zero. If m and w were tentatively matched at some stage of the procedure, and subsequently one of them chose to entertain a better offer, one can readily imagine that the other might be reluctant to form the same blocking pair later in the process, so that in some circumstances it might be desirable to assume that this probability goes to zero.

Kamecke U (1989) Non-cooperative matching games. Int J Game Theory 18: 423–431
Kelso AS Jr, Crawford VP (1982) Job matching coalition formation, and gross substitutes. Econometrica 50: 1483–1504
Knuth DE (1976) Mariages stables. Montreal, Les Presses de l'Universite de Montreal
Mongell S, Roth AE (1990) Sorority rush as a two-sided matching mechanism. Am Econ Rev (forthcoming)
Roth AE (1982) The economics of matching: stability and incentives. Math Operations Res 7: 617–628
Roth AE (1984a) The evolution of the labor market for medical interns and residents: a case study in game theory. J Political Econ 92: 991–1016
Roth AE (1984b) Misrepresentation and stability in the marriage problem. J Econ Theory 34: 383–387
Roth AE (1985) The college admissions problem is not equivalent to the marriage problem. J Econ Theory 36: 277–288
Roth AE (1989) Two sided matching with incomplete information about others' preferences. Games Econ Behav 1: 191–209
Roth AE (1990a) New physicians: a natural experiment in market organization. Science (forthcoming)
Roth AE (1990b) A natural experiment in the organization of entry level labor markets: regional markets for new physicians in the U.K. Am Econ Rev (forthcoming)
Roth AE, Sotomayor M (1990) Two-sided matching: a study in game-theoretic modelling and analysis. Econometric Society Monograph Series, Cambridge University Press (forthcoming)
Roth AE, Vande Vate JH (1990) Random paths to stability in two-sided matching. Econometrica (forthcoming)