Roth, A.E. and Vande Vate, J.H.,
"Random Paths to Stability in Two-Sided Matching,"
Econometrica, 58, 1990, 1475-1480.

Econometrica, Vol. 58, No. 6 (November, 1990), 1475–1480

RANDOM PATHS TO STABILITY IN TWO-SIDED MATCHING

BY ALVIN E. ROTH AND JOHN H. VANDE VATE[1]

1. INTRODUCTION

EMPIRICAL STUDIES OF TWO SIDED MATCHING have so far concentrated on markets in which certain kinds of market failures were addressed by resorting to centralized, deterministic matching procedures. Loosely speaking, the results of these studies are that those centralized procedures which achieved stable outcomes resolved the market failures, while those markets organized through procedures that yielded unstable outcomes continued to fail.[2] So the market failures seem to be associated with instability of the outcomes.

But 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. So 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 presumably introduces some randomness into what matchings are achieved. However, as far as we are aware, no nondeterministic models leading to stable outcomes have yet been studied.

The present paper demonstrates that, starting from an arbitrary matching, the process of allowing randomly chosen blocking pairs to match will converge to a stable matching with probability one. (This resolves an open question raised by Knuth (1976), who showed that such a process may cycle.) Furthermore, every stable matching can arise with positive probability from an initial situation in which all agents are unmatched.

2. RANDOM MATCHING IN THE MARRIAGE MODEL

We follow Gale and Shapley (1962) in considering the simple two-sided matching model, known as the marriage problem, in which matchings are one-to-one.[3] The two sets of agents are $M = \{m_1, \ldots, m_n\}$ and $W = \{w_1, \ldots, w_p\}$, called "men" and "women," and each agent has a complete and transitive preference ordering over the agents on the other side of the market and the prospect of remaining single. The preference ordering

[1] This work has been partially supported by grants from the National Science Foundation and the Alfred P. Sloan Foundation. We have received helpful comments from Robert Foley, David Gale, Donald Knuth, Jack Ochs, and Uri Rothblum.

[2] 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 (1990) 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 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.

[3] There is now a large literature concerning this and many much more general models of two-sided matching. See Roth and Sotomayor (1990) for a comprehensive account.

of a man m, for example, can be represented by an ordered list $P(m)$: if man m prefers w_i to w_j then w_i appears earlier in the list than does w_j, and so man m's preferences might be given by

$$P(m) = w_1, w_3, w_5, m, w_2, \ldots, w_k,$$

indicating that his first choice is to be matched to woman w_1, his second choice is to be matched to woman w_3, his fourth choice is to remain single, etc.[4] For our purposes it will be sufficient to describe only those people that an agent prefers to being single, so that the above preferences can be abbreviated by

$$P(m) = w_1, w_3, w_5.$$

Let $P \equiv \{P(m_1), \ldots, P(m_n), P(w_1), \ldots, P(w_p)\}$ 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 $\mu(m)$). For our purposes here, it will also be convenient to say that a man m forms a blocking pair with himself if $\mu(m) = w$ such that man m prefers being single to being matched with w, and similarly w forms a blocking pair with herself if she prefers being single to being matched with $m = \mu(w)$. A matching μ in which no man or woman forms a blocking pair with him or herself is called *individually rational*. A matching μ is *stable* if there are no blocking pairs.

Gale and Shapley (1962) proved that, for any preferences of the agents, the set of stable matchings is nonempty. Knuth (1976) however, observed that there may be cycles of blocking pairs, so that the process of allowing blocking pairs to form may not lead to a stable matching.[5] A bit of terminology will prove useful: 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 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, $\nu(m) = m$.

Knuth considered the example, with $n = 3$, with preferences:

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

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

Consider the unstable matching $\mu_1 = [(w_1, m_1), (w_2, m_2), (w_3, m_3)]$ with blocking pair (w_2, m_1), which when satisfied leads to the matching

$$\mu_2 = [(w_1, w_1), (w_2, m_1), (w_3, m_3), (m_2, m_2)],$$

blocked by (w_1, m_2), which when satisfied leads to

$$\mu_3 = [(w_1, m_2), (w_2, m_1), (w_3, m_3)],$$

[4] If man m were indifferent between, say, w_3 and w_5, we could indicate this by enclosing them in brackets in the preference list $P(m)$.

[5] Gale and Shapley and Knuth looked at models in which no agent remains single, but the slightly more general model considered here does not change any of the conclusions we discuss.

blocked by (w_2, m_3), which leads to

$$\mu_4 = [(w_1, m_2), (w_2, m_3), (w_3, w_3), (m_1, m_1)]$$

blocked by (w_3, m_1) and leading to

$$\mu_5 = [(w_1, m_2), (w_2, m_3), (w_3, m_1)],$$

blocked by (w_1, m_3), yielding

$$\mu_6 = [(w_1, m_3), (w_2, w_2), (w_3, m_1), (m_2, m_2)],$$

blocked by (w_2, m_2), leading to

$$\mu_7 = [(w_1, m_3), (w_2, m_2), (w_3, m_1)],$$

blocked by (w_1, m_1), which leads to

$$\mu_8 = [(w_1, m_1), (w_2, m_2), (w_3, w_3), (m_3, m_3)]$$

blocked by (w_3, m_3) and leading to μ_1, completing the cycle.

Note that in this example there is a path that leads to a stable matching. (If from μ_1 we begin by satisfying the blocking pair (w_2, m_3), we reach the stable matching $[(w_1, m_1), (w_2, m_3), (w_3, m_2)]$ in one more step.) The open question Knuth raised was whether at least one such path exists from any matching to a stable matching, for any preferences of the agents.[6] The theorem below resolves this question in the affirmative. A consequence is that a fairly large 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. This result is presented below as a corollary of the theorem.

THEOREM: *Let μ be an arbitrary matching for (M, W, P). Then there exists a finite sequence of matchings $\mu_1, \ldots, \mu_k$, such that $\mu = \mu_1$, μ_k is stable, and for each $i = 1, \ldots, k-1$, there is a blocking pair (m_i, w_i) for μ_i such that μ_{i+1} is obtained from μ_i by satisfying the blocking pair (m_i, w_i).*

PROOF: Let μ_1 be an arbitrary individually rational matching (if μ_1 is not individually rational, we can initiate the sequence with a string of blocking pairs formed by individuals, until we reach an individually rational matching). Suppose that μ_1 has a blocking pair (m_1, w_1). (If no blocking pairs exist, $k = 1$ and we are done.) Let μ_2 be the next matching in the sequence described in the Theorem, with $\mu_2(m_1) = w_1$, and define the set $A(1) \equiv \{m_1, w_1\}$. Note that if (m_2, w_2) is any blocking pair for μ_2, then $\{m_2, w_2\}$ is not contained in $A(1)$. The proof will proceed by constructing the required sequence of matchings in such a way that it can be associated with an increasing sequence of sets $A(q)$ which contain no blocking pairs, until a matching with no blocking pairs has been achieved.

Inductively, suppose we have a set $A(q)$ such that there are no blocking pairs for μ_{q+1} contained in $A(q)$, and such that μ_{q+1} does not match any agent in $A(q)$ to any agent

[6] Knuth (1976, problem 8) formally stated an open problem for the case in which there are equal numbers of men and women, all mutually acceptable. So in his problem all men and women were always matched, and when he satisfied a blocking pair he required that the "divorced" spouses should be matched to each other. Here we speak of the related question in the more general model we consider, and without any "forced" marriages, i.e. without requiring that the divorced spouses be matched to one another. Our theorem does not resolve the question of under what circumstances paths to stable matchings can always be found having the property that at each step the spouses of the blocking pair to be satisfied will always be matched to one another. (Clearly such paths cannot always exist when there are different numbers of men and women.)

not in $A(q)$. Then if μ_{q+1} is not stable, there exists a blocking pair (m', w') such that at most one of m' and w' is contained in $A(q)$. We consider three cases.

First, suppose there is a blocking pair (m_{q+1}, w_{q+1}) such that m_{q+1} is contained in $A(q)$, and choose it to be the blocking pair with the property that, among all such blocking pairs (m, w_{q+1}), m_{q+1} is w_{q+1}'s most preferred mate[7] in $A(q)$. Let the next matching in the sequence, μ_{q+2}, be formed using this blocking pair, and define $A(q+1) \equiv A(q) \cup \{w_{q+1}\}$. If m_{q+1} was unmatched at μ_{q+1}, then $A(q+1)$ is a set such that no blocking pair for μ_{q+2} is contained in $A(q+1)$. Otherwise there may be a blocking pair (m_{q+2}, w_{q+2}) for μ_{q+2}, with $w_{q+2} = \mu_{q+1}(m_{q+1})$ and m_{q+2} both contained in $A(q+1)$, in which case we choose it to be the blocking pair with the property that, among all such blocking pairs (m, w_{q+2}), m_{q+2} is w_{q+2}'s most preferred mate in $A(q+1)$. The next matching in the sequence, μ_{q+3}, is formed using this blocking pair, and the process continues until we reach a matching μ_r, $r > q$, such that no blocking pairs for μ_r are contained in the set $A(r) \equiv A(q+1)$. (This must eventually happen, since no man ever receives a worse mate and hence no blocking pair appears twice in the sequence $\mu_{q+2}, \ldots, \mu_r$.) The set $A(r)$ is the set we require: it strictly contains $A(q)$, and contains no blocking pairs for μ_r.

The remaining two cases are now simple to consider. If there is no blocking pair (m_{q+1}, w_{q+1}) with m_{q+1} contained in $A(q)$, but there is one with w_{q+1} contained in $A(q)$, then we proceed as above, reversing the role of men and women. If every blocking pair (m_{q+1}, w_{q+1}) is disjoint from $A(q)$, then select any such blocking pair to form the next matching μ_{q+2}, and define the set $A(q+1) \equiv A(q) \cup \{m_{q+1}, w_{q+1}\}$. $A(q+1)$ contains $A(q)$, and contains no blocking pairs for μ_{q+2}, so it is the required set in this case. The process must stop in finitely many steps (since $A(q)$ can be strictly increased until a stable matching is reached, but it cannot grow larger than $M \cup W$), so a stable matching is eventually reached. This completes the proof.

We can now 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 depends only on the matching $\mu = \mu_i$ (and not on i).[8] Let $R(\mu)$ be the random sequence generated in this way from an initial matching μ. We can now state the following corollary, whose proof is immediate from the theorem and the positive probability of every blocking pair.

COROLLARY: *For any initial matching μ, the random sequence $R(\mu)$ converges with probability one to a stable matching.*

3. CONCLUDING REMARKS

The process by which we have constructed the sequence of matchings in the proof of the Theorem is very closely related to the deferred acceptance algorithm proposed by Gale and Shapley (1962) to prove that the set of stable matchings is always nonempty.

[7] At any point in which an agent is indifferent between more than one most preferred mate, ties may be broken arbitrarily.

[8] 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. And the Corollary holds even if we relax the assumption that this probability must be the same every time the matching μ arises, so long as the probability is bounded away from zero.

Indeed, if we begin with the matching μ_1 at which all agents are single, and choose the set $A(1)$ to be the set of all men, then the sequence constructed in the Theorem is precisely the sequence of matchings which occur in the deferred acceptance algorithm with women proposing (and it converges to the stable matching that is best for the women—see Gale and Shapley (1962), Roth and Sotomayor (1990)).

Note that, if we begin with the matching μ_1 at which all agents are single, *every* stable matching can be reached by a sequence of matchings as in the Theorem.[9] For example, if μ is an arbitrary stable matching, then the sequence formed by letting

$$A(1) = \{m_1, \mu(m_1)\}, \ldots A(i) = A(i-1) \cup \{m_i, \mu(m_i)\}, \ldots$$

leads to μ. So the class of random processes $R(\mu_1)$ discussed in the corollary yields every stable matching with positive probability.

The special structure of the marriage model implies that the set of stable matchings precisely equals the core, when the rules are that every man and woman is free to remain single, and every mutually consenting couple consisting of one man and one woman is free to marry.[10] However it is possible that our results may also have a bearing on more general situations in which pairwise optimality has global implications: see, e.g., Feldman (1973) who studies conditions under which pairwise optimal outcomes in a pure exchange economy with money are Pareto optimal.

A natural direction to pursue further research will be into the incentives facing agents who face the prospect of being matched by some sort of random stable mechanism. It is already known (Roth (1982)) that there exists no revelation mechanism which 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 deferred acceptance algorithm is a mechanism with 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. Some simple models in which a similar result can be obtained for random stable mechanisms are explored in Roth and Vande Vate (1990).

Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.
and
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

Manuscript received August, 1989; final revision received December, 1989.

REFERENCES

CRAWFORD, VINCENT P., AND ELSIE MARIE KNOER (1981): "Job Matching with Heterogeneous Firms and Workers," *Econometrica*, 49, 437–450.

FELDMAN, ALLAN M. (1973): "Bilateral Trading Processes, Pairwise Optimality, and Pareto Optimality," *Review of Economic Studies*, 40, 463–473.

GALE, DAVID, AND LLOYD SHAPLEY (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–15.

[9] The Theorem thus gives a new proof of the nonemptiness of the set of stable matchings, which does not introduce any asymmetries between the two sets of agents.

[10] And when we generalize the model to the case of many-to-one matching, so that e.g. firms may employ many workers, the set of stable matchings is contained in the core.

KELSO, ALEXANDER S., JR., AND VINCENT P. CRAWFORD (1982): "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica*, 50, 1483–1504.

KNUTH, DONALD E. (1976): *Mariages Stables*. Montreal: Les Presses de l'Universite de Montreal.

MONGELL, SUSAN, AND ALVIN E. ROTH (1990): "Sorority Rush as a Two-Sided Matching Mechanism," *American Economic Review*, forthcoming.

ROTH, ALVIN E. (1982): "The Economics of Matching: Stability and Incentives," *Mathematics of Operations Research*, 7, 617–628.

——— (1984a): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92, 991–1016.

——— (1984b): "Misrepresentation and Stability in the Marriage Problem," *Journal of Economic Theory*, 34, 383–387.

——— (1990): "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians in the U.K.," *American Economic Review*, forthcoming.

ROTH, ALVIN E., AND MARILDA SOTOMAYOR (1990): *Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis*, Econometric Society Monograph Series. New York: Cambridge University Press.

ROTH, ALVIN E., AND JOHN H. VANDE VATE (1990): "Incentives in Two-Sided Matching with Random Stable Mechanisms," *Economic Theory*, 1, forthcoming.