Discrete Applied Mathematics 11 (1985) 223-232
North-Holland

SOME REMARKS ON THE STABLE MATCHING PROBLEM

David GALE*
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA

Marilda SOTOMAYOR**
Pontificia Universidade Catolica do Rio de Janeiro, Brazil

Received 1 November 1983
Revised 3 August 1984

The stable matching problem is that of matching two sets of agents in such a manner that no two unmatched agents prefer each other to their mates. We establish three results on properties of these matchings and present two short proofs of a recent theorem of Dubins and Freedman.

1. Introduction

The problem of our title was introduced by Gale and Shapley [2] in 1962. Since then, it has been the subject of numerous research articles and even a short book (Knuth 1976 [3]). The problem continues to be of interest as a rare instance of an exercise in 'pure' mathematics (combinatorial theory of ordered sets) which is actually beng applied to real world situations (assigning medical school graduates to hospitals).

In the next section, we recall the problem briefly and remark on some of its history. The rest of the paper is devoted to proving three new results and giving new proofs of a recent result of Dubins and Freedman [1].

2. The problem, some history, results

The problem of 'college admissions' as described in [2] involves a set of *institutions* and a set of *applicants*. Each applicant lists in order of preference those institutions he wishes to attend and each institution lists in order of preference those applicants it is willing to admit. In addition, each institution has a *quota* giving an upper bound on the number of applicants it can admit. The problem is then to devise

* Partially supported by the National Science Foundation under Grant SES-8113548 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.
** Partially supported by the Conselho Nacional de Desenvolvimento Cientifico e'Tecnologico, Brazil.

some method of assigning applicants to institutions in a way which takes account of their respective preferences. The key notion turns out to be that of *stability*. An assignment is said to exhibit *instability* if there is an applicant α and an institution I such that α prefers I to the institution to which he is assigned and I either has not filled its quota or if it has, it prefers α to some other applicant who is assigned to it. An assignment or *matching*, as we shall call it, is *stable* if it does not exhibit instability. Question: given any set of applicants and institutions, together with their preferences and quotas, can one always find a stable matching? An affirmative answer is given in [2]. The proof is constructed by means of a simple algorithm which starting from the given preference data arrives at a stable matching (in a number of steps roughly proportional to the square of the number of applicants and institutions). The matching obtained in this way turns out to have the rather surprising property that it is the unique stable matching (there may be many) which is preferred by all the applicants to any other such matching.

The above paragraph summarizes the content of [2]. The question of course then arises as to whether these results can be applied 'in practice'. The authors of [2] had expressed some reservation on this point, – and then came another surprise. Not only could the method be applied, it already had been more than ten years earlier! The National Resident Matching Program (located in Evanston, Illinois, founded in 1951) has the task each year of assigning graduates of all the medical schools in the country to hospitals where they are required to serve a year's residency. The method uses by NRMP to do this is exactly the one described in [2] but 'in reverse', that is, the matching obtained is hospital rather than applicant-optimal and hence it is the worst, rather than the best stable matching from the point of view of the students. The two algorithms are easy to describe.

NRMP-Algorithm

Each hospital H *tentatively admits* its quota q_H consisting of the top q_H applicants on its list. Applicants who are tentatively admitted to more than one hospital *tentatively accept* the one they prefer. Their names are then removed from the lists of all other hospitals which have tentatively admitted them. This gives the *first tentative matching*. Hospitals which now fall short of their quota again admit tentatively until either their quotas are again filled or they have exhausted their list. Admitted applicants again reject all but their favorite hospital, giving the *second tentative matching*, etc. The algorithm terminates when, after some tentative matching, no hospitals can admit any more applicants either because their quota is full or they have exhausted their list. The tentative matching then becomes permanent.

Gale–Shapley (G-S)-Algorithm

Each applicant *petitions* for admission to his/her favorite hospital. In general, some hospitals will have more petitioners than allowed by their quota. Such over-

subscribed hospitals now reject the lowest petitioners on their preference list so as to come within their quota. This is the *first tentative matching*. Next, rejected applicants petition for admission to their second favorite hospital and again oversubscribed hospitals reject the overflow, etc. The algorithm terminates when every applicant is tentatively admitted or has been rejected by every hospital on his list.

The simplest examples (two applicants, two hospitals) show that the two algorithms need not give the same result. We do not know whether NRMP chose the hospital-optimal method as a matter of policy or whether they were not aware of the alternative possibility. In any case, the discrepancy between the two methods leads at once to the first result of this paper which asserts:

Let μ be any stable matching, let $S(\mu)$ be the set of applicants admitted to some hospital and let $n_H(\mu)$ be the number of applicants admitted to hospital H. Then the set $S(\mu)$ and numbers $n_H(\mu)$ are the same for all stable μ.

In words, although NRMP is worse then G-S for students, it is at least true that any student admitted under G-S will be admitted to some hospital under NRMP. Further, though G-S is worse than NRMP for hospitals, it is still true that each hospital will fill the same fraction of its quota in both cases. This result does not seem obvious although, as we will see, its proof is quite simple.

Our second result is concerned with the question of what happens to the matching if an institution (applicant) extends the list of applicants (institutions) it is willing to accept (enter). Intuitively, one feels that if an institution extends its list, this would be good for the applicants but, possibly, bad for the other institutions. Intuition turns out to be right in the case but the proof is not so easy. We have:

Whether one uses the applicant-optimal (G-S) or institution-optimal (NRMP) matching, it will always be the case that if an institution extends its list no applicant will be made worse off and no institution will be made better off.

Our third result shows that the applicant optimal matching is 'Pareto optimal' for the applicants. That is, there is no matching, stable or not, which is better for all applicants than the applicant-optimal matching.

Several years ago Dubins and Freedman [1] showed that the applicant-optimal matching was 'cheat-proof' for the applicants. This means the following: suppose all preference data is available to all applicants and institutions. One can imagine then that some clever applicant or institution could take advantage of this by suitably falsifying their own preference data in such a way that the final matching would be better for them than if they had been honest. It is shown in [1], by an example, that this is indeed true for the institutions but not for the applicants. More precisely, no set of applicants by falsifying preferences can force a matching which is preferred by all applicants in the set. The proof in [1] is quite long. In the final section, we give a shorter proof of this result, using only the properties of stability and

applicant-optimality. In an appendix, we give an even shorter proof of the result which, however, makes use of the G-S matching algorithm (as did the proof of Dubins and Freedman).

It is interesting to note that these last two results do not hold for the institution optimal matching. Roth [4] has given an example involving 3 institutions and 4 applicants in which there is a matching which is preferred by all institutions to the institution-optimal matching and further the institutions can force this matching by falsifying their preferences.

3. The model

As in previous treatments of the problem, we begin by reducing it to the special case in which each institution has a quota of one. This is done by the following device: we replace institution A by q_A copies of A denoted by $A_1, A_2, \ldots, A_{q_A}$. Each of these A_i has preferences identical with those of A but with a quota of 1. Further, each applicant who has A on his preference list now replaces A by the string $A_1, A_2, \ldots, A_{q_A}$ in that order of preference. It is now easy to verify that the stable matchings for the original problem are in natural one-to-one correspondence with the stable matchings of this modified model. With this modification, the model becomes completely symmetric in the applicants and institutions. To reflect this, we make the usual change of scenario to that of the 'stable marriage problem' in which instead of applicants and institutions, we consider men and women and think of the matchings as (monogamous) marriages.

We now give the formal presentation of the model. There are two sets M and W (men and women). We assume that each man m in M has a total (preference) ordering P_m on the set $W \cup \{m\}$, the set of all women and himself. The position in which he places himself in the ordering has the meaning that the only women he is willing to be matched with are those whom he prefers to himself. This formulation turns out to be convenient for our analysis. Similarly, each w in W has a total ordering P_w on the set $M \cup \{w\}$. We write $w >_m w'$ to mean m prefers w to w'. Similarly, we write $m >_w m'$. We say that w is *acceptable* to m if $w >_m m$ and analogously for W. A pair (m, w) is *compatible* if each is acceptable to the other. The sets M and W together with the orderings will be called a *preference structure*, denoted by $\mathcal{P}$.

We will need one more concept. As mentioned earlier, we will be interested in studying the effect on stable matchings when a man or woman extends his or her list of acceptable people. In our formulation, this corresponds to people changing their own position in their preference ordering. We will write $P'_m \geq P_m$ if m has possibly lowered his own position in the ordering on $W \cup \{m\}$ leaving the ordering unchanged otherwise. Similarly, we define $P'_w \geq P_w$, and finally we write $\mathcal{P}' \geq_M \mathcal{P}$ if $P'_m \geq P_m$ for all m in M.

We now turn to the matter of matchings. In general, it will not be possible to match all of M and W. We therefore make the convention that a person who is not

matched to someone of the opposite sex is matched with him/herself. Formally, we have

Definition. A *matching* μ is a function from the set $M \cup W$ onto itself of order two, (that is, $\mu^2(x) = x$) such that if $\mu(x) \neq x$, then $\{x, \mu(x)\}$ is a compatible pair. We refer to $\mu(x)$ as the *mate* of x.

There are two natural partial orders on the set of all matchings. If μ and μ' are distinct matchings, we write $\mu >_M \mu'$ if $\mu(m) \geq_m \mu'(m)$ for all m in M. Similarly, we define $\mu >_W \mu'$.

If μ and μ' are matchings, the function $\nu = \mu \vee_M \mu'$ is defined by $\nu(m) = \max(\mu(m), \mu'(m))$ and $\nu(w) = \min(\mu(w), \mu'(w))$. In general, ν will not be a matching. The function $\eta = \mu \vee_W \mu'$ is defined analogously (notice that $\mu \vee_W \mu'$ is the same as $\mu \wedge_M \mu'$ in the usual lattice notation).

Key Definition. If μ is a matching, we say that the pair (m, w) *blocks* μ if m and w prefer each other to their mates, that is, $w >_m \mu(m)$ and $m >_w \mu(w)$. A matching is *stable* if it is not blocked by any pair.

4. Properties of stable matchings

In this section, we prove the first and second theorem described in Section 2.

Lemma 1 (Decomposition). *Let $\mathcal{P}$ and $\mathcal{P}'$ be preference structures with $\mathcal{P}' \geq_M \mathcal{P}$ and let μ and μ' be corresponding stable matchings. Let M_μ ($M_{\mu'}$) be all men who prefer μ to μ' (μ' to μ) and define $W_{\mu'}$ and W_μ analogously. Then μ' and μ are bijections between $M_{\mu'}$ and W_μ.*

Proof. Suppose $m \in M_{\mu'}$. Then $\mu'(m) >_m \mu(m) >_m m$ so $\mu'(m) \in W$ since $\mu' >_m \mu$. Setting $w = \mu'(m)$, we cannot have $\mu'(w) >_w \mu(w)$ for then (m, w) would block μ. Hence, $w \in W_\mu$ and we have

$$\mu'(M_{\mu'}) \subset W_\mu \tag{1}$$

On the other hand, if $w \in W_\mu$, then $\mu(w) >_w \mu'(w) \geq_w w$ so $\mu(w) \in M$. Letting $\mu(w) = m$, we see that we cannot have $\mu(m) >_m \mu'(m)$ or (m, w) would block μ'. Hence,

$$\mu(W_\mu) \subset M_{\mu'}. \tag{2}$$

Since μ and μ' are injective and $M_{\mu'}$ and W_μ are finite, the conclusion follows. □

Corollary 1. *$\mu' >_M \mu$ if and only if $\mu >_W \mu'$.*

Proof. $\mu' >_M \mu$ if and only if M_μ is empty, so μ and μ' agree on $M - M_{\mu'}$ and $W - W_\mu$ which is equivalent to $\mu >_W \mu'$. □

For the special case where $\mathcal{P} = \mathcal{P}'$, the corollary states that the orders $>_M$ and $>_W$ are inverses of each other.

Our first result is an immediate consequence of the lemma.

Theorem 1. *The set of people who are matched with themselves is the same for all stable matchings.*

Proof. Suppose, say, m was matched under μ' but not under μ (assume now $\mathcal{P} = \mathcal{P}'$). Then $m \in M_{\mu'}$, but from the lemma, μ maps W_μ onto $M_{\mu'}$ so m is also matched under μ, contradiction. □

Lemma 2. *With assumptions and notations of Lemma* 1, *we have*

$$\nu = \mu \vee_M \mu' \text{ is a matching and is stable for } \mathcal{P}. \tag{3}$$

$$\eta = \mu \vee_W \mu' \text{ is a matching and is stable for } \mathcal{P}'. \tag{4}$$

Proof. By definition, $\mu \vee_M \mu'$ must agree with μ' on $M_{\mu'}$ and W_μ and with μ otherwise. By Lemma 1, ν is therefore bijective. Further, for $m \in M_{\mu'}$, $\mu'(m) >_m \mu(m) \geq_m m$ in $\mathcal{P}$ so μ' is a permissible matching in $\mathcal{P}$. Suppose now that some (m, w) blocks ν and $m \in M_{\mu'}$ so $w >_m \mu'(m)$. Then certainly $w >_m \mu(m)$ so if $w \in W_\mu$, then $m >_w \mu'(w)$ and μ' would be blocked, and if $w \in W - W_\mu$, then $m >_w \mu(w)$ and μ would be blocked. On the other hand, if $m \in M - M_{\mu'}$, then $m >_m \mu(m) \geq_m \mu'(m)$ so (m, w) would block μ or μ' according as w is in $W - W_\mu$ or W_μ. This proves (3).

Next, we have η agrees with μ on $W_\mu \cup M_{\mu'}$ and with μ' otherwise. Again, η is a matching from Lemma 1. The stability argument is as in the previous paragraph. □

Corollary 2 (Conway-Knuth). *The set of stable matchings for* $\mathcal{P}$ *form a lattice.*

Proof. Take $\mathcal{P}' = \mathcal{P}$ above. □

Corollary 3. *There are stable matchings* μ_M (μ_W) *preferred by all man* (*women*) *to any other stable matching.*

Proof. Every finite lattice has a maximum and minimum element. □

Corollary 3, also proved by a different method in [2], is especially surprising when applied to the original model with applicants and institutions. It says that if the institution-optimal matching is used (as in the case of NRMP), then the applicants that each institution gets are 'preferred' by that institution to those applicants it

would get under any other stable matching. Thus, suppose institution A gets applicants $\alpha_1 \succ_A \alpha_2 \cdots \succ_A \alpha_k$ under the institution-optimal matching and A gets $\beta_1 \succ_A \beta_2 \cdots \succ_A \beta_r$ under some other stable matching. Then $\alpha_1 \succeq_A \beta_1, \alpha_2 \succeq_A \beta_2, \ldots$. This is a very strong form of optimality.

The matchings μ_M and μ_W are called the *M-optimal* and *W-optimal* matchings.

We now prove our second main result.

Theorem 2. *Suppose* $\mathscr{P}' \succeq_M \mathscr{P}$ *and let* μ'_M, μ_M *and* μ'_W, μ_W *be corresponding optimal matchings. Then*

$$\mu_M \succeq_M \mu'_M \qquad (\text{so } \mu'_M \succeq_W \mu_M, \text{ Corollary 1}) \tag{5}$$

and

$$\mu'_W \succeq_W \mu_W \qquad (\text{so } \mu_W \succeq_M \mu'_W, \text{ Corollary 1}). \tag{6}$$

In words, the men are better off and women worse off under $\mathscr{P}$ than under $\mathscr{P}'$ *no matter which of the two optimal matchings are used.*

Proof. By Lemma 2, $\mu_M \vee_M \mu'_M$ is $\mathscr{P}$-stable so $\mu_M \succeq_M \mu_M \vee_M \mu'_M \succeq_M \mu'_M$.

Also, by Lemma 2, $\mu_W \vee_W \mu'_W$ is $\mathscr{P}'$-stable so $\mu'_W \succeq_W \mu_W \vee_W \mu'_W \succeq_W \mu_W$. □

5. Pareto optimality[1]

We wish to show that there is no matching μ which is (strictly) preferred by all men to the matching μ_M. To prove this, we introduce some terminology. If μ is a matching, we say that m *admires* w if m and w are compatible and m prefers w to his mate $\mu(m)$ (thus, m and w block μ if each admires the other).

Proposition. *If* $|M| \leq |W|$, *then there is a woman* w *in* $W' = \mu_M(M)$ *who has no admirers.*

For the proof, we need the following facts about finite sets.

Lemma 3. *Let* f *and* g *be functions from a finite set* X *into a set* Y *where* f *is bijective. Then there is a non-empty subset* $A \subset X$ *such that* f *and* g *are bijections from* A *to* $f(A)$.

Proof. Let $h = f^{-1} \circ g$. Then h maps X into X and since X is finite and $h^{n+1}(X) \subset h^n(X)$, we must have $h^k(X) = h^{k+1}(X)$ for some k. The set $A = h^k(X)$ has the desired property. □

[1] For those familiar with the matching algorithm of [2], a very short proof of the material of this section is given in the appendix.

To prove the proposition, suppose every w in W' has an admirer and let $\alpha(w)$ be her favorite admirer. Applying Lemma 3 to the functions α and μ_M gives a set $\hat{W}$ such that $\mu_M(\hat{W}) = \alpha(\hat{W})$. Now define $\hat{\mu}$ to agree with α on $\hat{W}$, with α^{-1} on $\mu(\hat{W})$ and with μ_M otherwise. Now $\hat{\mu}$ is stable for if (m, w) were to block $\hat{\mu}$, then w would be in $\hat{W}$ and m would admire w, but w is matched by $\hat{\mu}$ to her favorite admirer whom she therefore prefers to m. But $\hat{\mu}$ is preferred to μ_M by all m in $\mu(\hat{W})$ contradicting the fact that μ_M is M-optimal.

Theorem 3. *There is no matching μ (stable or not) such that $\mu \succ_m \mu_M$ for all m in M.*

Proof. If $|M| > |W|$, then the result is immediate since, in that case, at least one man would be matched with himself by μ. If the conclusion were false, then every man would have to be matched under μ with someone he admires under μ_M, but from the proposition there is at least one woman who no one admires under μ_M. □

6. The Dubins–Freedman Theorem

We need the following result whose formulation is due to J.S. Hwang:

Key Lemma. *Let μ be any matching on $\mathscr{P}$ and let M' be all men who prefer μ to μ_M. Then there is a pair (m, w) which blocks μ where $m \in M - M'$.*

Proof. *Case* I: $\mu(M') \neq \mu_M(M')$. Choose w in $\mu(M') - \mu_M(M')$, say, $w = \mu(m')$. Then m' admires w under μ_M so w prefers $\mu_M(w) = m$ to m', and m is not in M' since w is not in $\mu_M(M')$, hence m prefers w to $\mu(m)$ so (m, w) blocks μ.

Case II:[2] $\mu(M') = \mu_M(M') = W'$. We now define a preference structure $\mathscr{P}'$ for M', W'. First, P'_m is the same as P_m restricted to $W' \cup \{m\}$ for all m in M'. For w in W', P'_w agrees with P_w restricted to $M' \cup \{w\}$ except that w is now ranked just below $\mu(w)$. In other words, the only men in M' who are acceptable to w are those m such that $m \geq_w \mu(w)$. Note that μ_M restricted to $M' \cup W'$ is still stable for $\mathscr{P}'$, because any pair which blocks in $\mathscr{P}'$ would also block in $\mathscr{P}$. Letting $\mu_{M'}$ be the M'-optimal matching for $\mathscr{P}'$, we see that $\mu_{M'} \succ_{M'} \mu_M$, that is, there is at least one m in M' who prefers $\mu_{M'}$ to μ_M because by hypothesis, $\mu \succ_m \mu_M$ for all m in M' and if $\mu_M = \mu_{M'}$, this would contradict Theorem 3. We now define $\hat{\mu}$ on $M \cup W$ by

$$\hat{\mu} = \mu_{M'} \quad \text{on } M' \cup W',$$
$$= \mu_M \quad \text{on } (M - M') \cup (W - W').$$

[2] A shorter proof of this part of the lemma is given in the appendix for readers familiar with the matching algorithm of [2].

Since $\hat{\mu} \succ_M \mu_M$, we know that $\hat{\mu}$ is not stable for $\mathscr{P}$ so let (m, w) be the blocking pair. Now we cannot have (m, w) in $M' \cup W'$ because if m was acceptable to w in $\mathscr{P}'$, then (m, w) would block $\mu_{M'}$ and if m is not acceptable to w in $\mathscr{P}'$, then by construction of P'_w, w prefers $\mu_{M'}(w) = \hat{\mu}(w)$ to m. Further, if $m \in M'$ and $w \in W - W'$, then $\{m, w\}$ does not block $\hat{\mu}$ because then $\{m, w\}$ would block μ_M, since m is no better off under μ_M than under $\hat{\mu}$. Therefore, we must have $m \in M - M'$ and $w \in W'$, but then $\{m, w\}$ also blocks μ_M because again, by construction of P'_w, w is at least as well off under $\mu_{M'}$ as under μ. □

Theorem 4 (Dubins-Freedman). *Let $\mathscr{P}$ be a preference structure and let $\tilde{\mathscr{P}}$ differ from $\mathscr{P}$ in that some set $\tilde{M} \subset M$ falsify their preferences. Then there is no matching μ, stable for $\tilde{\mathscr{P}}$ which is preferred to μ_M by all members of $\tilde{M}$.*

Proof. Suppose μ is a matching preferred to μ_M by $M' \supset \tilde{M}$. Then from the Key Lemma, there is a pair (m, w) which blocks μ in $\mathscr{P}$. But $m \in M - M'$, so neither m nor w is falsifying preferences, so $\{m, w\}$ also blocks μ in $\tilde{\mathscr{P}}$ so μ is not $\tilde{\mathscr{P}}$-stable. □

7. Appendix

In order to make this article self-contained, we did not assume the reader was familiar with the matching algorithm of [2]. For those who are, the proofs of Theorems 3 and 4 can be considerably shortened.

Theorem 3 (Pareto Optimality). *There is no matching which makes all the men better off than they are under μ_M.*

Proof. If μ were such a matching it would match every man to some woman who had rejected him in the algorithm, hence all of these women, $\mu(M)$, would have been matched under μ_M, hence all of M would have been matched under μ_M. But the last woman to get a proposal has not rejected anyone, contradiction. □

Proof of Key Lemma. *Case* I: Same proof as in Section 6.

Case II: $\mu_M(M') = \mu(M') = W'$. Let w be the last woman in W' to receive a proposal from a member of M' in the algorithm. Since all w in W' have made rejections, we see that w had a tentative mate m when she received this last proposal. We claim (m, w) is the desired blocking pair. First, m is not in M' for if so, after having been rejected by w, he would have proposed again to a member of W' contradicting the fact that w received the last such proposal. But m prefers w to his mate under μ_M and since he is no better off under μ, he prefers w to $\mu(m)$. On the other hand, m was the last man in M' to be rejected by w so she must have rejected her mate under μ, call him m', *before* she rejected m and hence she prefers m to m' so (m, w) blocks μ as claimed.

References

[1] L.E. Dubins and D. Freedman, Machiavelli and the Gale-Shapley Algorithm, Amer. Math. Monthly 88 (7) (1981) 485–494.

[2] D. Gale and L.S. Shapley, College Admissions and the Stability of Marriage, Amer. Math. Monthly 69 (1962) 9–14.

[3] D.E. Knuth, Marriages Stables (Montreal Univ. Press, Montreal, 1976).

[4] A.E. Roth, Incentives in the college admissions problem and related two-sided markets, Working paper #173, Department of Economics, University of Pittsburgh (1983).