

Stable matchings and equilibrium outcomes of the Gale–Shapley’s algorithm for the marriage problem

Lin Zhou

Cowles Foundation, Yale University, New Haven, CT 06520, USA

Received 31 October 1990

Accepted 7 January 1991

This note investigates the strategic aspect of the Gale–Shapley’s (1962) ‘deferred acceptance’ algorithm for the marriage problem. We prove that if a stable matching with respect to the true preferences is supported by some preference profile (possibly a non-equilibrium one), then it can be supported by a strategic equilibrium. Our result complements Roth’s result (1984) in showing that even though agents reveal their preferences strategically, the G–S algorithm still yields stable matchings with respect to the true preferences.

1. Introduction

This note investigates the strategic aspect of the Gale–Shapley’s (henceforth G–S) ‘deferred acceptance’ algorithm for the marriage problem. Gale and Shapley (1962) proved that the G–S algorithm always leads to a stable matching when applied to true preferences. However, when agents’ true preferences are privately known, the G–S algorithm has to work with agents’ stated preferences. Dubin and Freedman (1981), and Roth (1982) showed that in the G–S algorithm with men making proposals it is generally not a dominant strategy for each woman to state the truth. This naturally raised the question whether the G–S algorithm can lead to a stable matching with respect to true preferences when it is applied to the stated – probably false – preferences. A partial answer to this question was provided by Roth (1984). He proved: If women’s stated preferences constitute a Nash equilibrium, then the resulting matching of the G–S algorithm must be stable for the true preferences. But he did not prove the existence of such equilibria. In a later paper Gale and Sotomayor (1985) proved for the generalized G–S algorithm that all stable matchings for the true preferences can be supported as a Nash equilibrium. But this result obviously does not hold in the original G–S algorithm. In this note we prove the existence of Nash equilibria in the original G–S algorithm. We shall show that if a stable matching for the true preferences is supported by some preference profile (possibly a non-equilibrium one), then it can be supported by a Nash equilibrium. This immediately implies the nonemptiness of Nash equilibria in the G–S algorithm since the M -optimal stable matching is supported by true preferences. Our result complements Roth’s in showing that even though agents reveal their preferences strategically, the G–S algorithm still yields stable matchings with respect to the true preferences as equilibrium outcomes.

2. The model and main results

There are two disjoint sets of agents – men and women: $M = \{m_1, \dots, m_n\}$ and $W = \{w_1, \dots, w_n\}$. Each man m_i has a strict preference relation $P(m_i)$ over W , and each woman w_j has a strict preference relation $P(w_j)$ over M . A preference profile P is a $2n$ -tuple: $P = \{P(m_1), \dots, P(m_n); P(w_1), \dots, P(w_n)\}$. A *matching* is a one-to-one correspondence from M to W . It is usually denoted by $x = \{(m_1, x(m_1)), \dots, (m_n, x(m_n))\}$, where $x(m_i) = w_j$ is the woman matched with m_i and $x^{-1}(w_j) = m_i$ the man matched with w_j .

A matching x is *unstable* with respect to P if there are m_i and w_j such that (i) $w_j P(m_i) x(m_i)$, and (ii) $m_i P(w_j) x^{-1}(w_j)$. Obviously an unstable matching is not desirable, if it ever occurs, when men and women are free to choose their partners. A matching x is *stable* with respect to P if it is not unstable. Let $S(P)$ denote the set of stable matchings for P .

Gale and Shapley introduced the ‘deferred acceptance’ algorithm as follows. Let P be a preference profile.

Step 1. Each man proposes to his favorite woman. Each woman who receives proposals rejects all but her favorite from among those who propose to her and keeps him as her suitor.

Step k. Each man who has been rejected in the previous step proposes to his favorite from among those who have not rejected him yet. Each woman who receives proposals rejects all but her favorite from among those who propose to her and the man she kept in the previous step, and keeps him as her suitor.

The algorithm ends when every woman receives a proposal. At that point each woman has exactly one suitor, and she is then matched with him. Gale and Shapley showed the algorithm terminates in finite steps for every P . The resulting matching is stable with respect to P and it is M -optimal in the sense that each man weakly prefers (according to P) this matching to any other stable matching with respect to P .

When agents’ true preferences are privately known and agents are free to reveal anything they want, the G–S algorithm defines a game form. The strategy set for each man is the same set of all strict preference relations over the women, and the strategy set for each woman is the set of all strict preference relations over the men. For each strategy profile R , the G–S algorithm produces $M(R)$ – the M -optimal stable matching for R – as the outcome. With each true preference profile $P = \{P(m_1), \dots, P(m_n); P(w_1), \dots, P(w_n)\}$, the G–S algorithm induces a game $\Gamma(P)$. Dubin and Freedman (1981), and Roth (1982) showed: $P(m_i)$ is a dominant strategy for each m_i to play in $\Gamma(P)$ for every P , but $P(w_j)$ is not always a dominant strategy for each w_j . Hence, some women at some situations have incentives to state false preferences. It is then no longer clear that the G–S algorithm will always lead to matchings that are stable with respect to true preferences.

To analyse what the G–S algorithm yields when agents state their preferences strategically, one has to make certain assumptions on agents’ behaviors. In Roth (1984) it is assumed that men always state the truth and women state preference relations that constitute Nash equilibria of the game $\Gamma(P)$. This is a reasonable assumption given D–F’s and his findings. Roth then proved that any such equilibrium outcome of $\Gamma(P)$ must be stable with respect to true preference profile P . But he did not show that at least one such equilibrium exists.

The issue of existence of Nash equilibria was later considered by Gale and Sotomayor (1985) for the generalized G–S algorithm. In the generalized G–S algorithm, each agent has a list of ‘unacceptable partners’ whom he or she will never be matched with. Here for each women the extent to which she can manipulate the outcome has been greatly enlarged. A woman can demand a specific

man and reject all her proposers unless the demand is met. It is then direct to verify that in the generalized G–S algorithm any stable outcome x can be supported by an equilibrium in which each woman labels all men but her partner in x as unacceptable. This type of manipulations is totally absent in the original G–S algorithm. We can give a very simple example that shows the result by Gale and Sotomayor does not hold for the original G–S algorithm.

Example 1. The true preference profile $P = \{P(m_1), P(m_2), P(m_3); P(w_1), P(w_2), P(w_3)\}$ is given by the following ranking matrix, in which the first number of each entry indicates the man’s ranking of the woman and the second number the woman’s ranking of the woman.

	w_1	w_2	w_3
m_1	1,3	2,2	3,1
m_2	3,1	1,3	2,2
m_3	2,2	3,1	1,3

There are three stable matchings: $x = ((m_1, w_2), (m_2, w_2), (m_3, w_3))$, $y = ((m_1, w_2), (m_2, w_3), (m_3, w_1))$, and $z = ((m_1, w_3), (m_2, w_1), (m_3, w_2))$. However, x is the only equilibrium outcome. The reason is simple. If men state the truth, then they propose to different women in the first step. Thus the G–S algorithm terminates with x as the outcome.

We now state and prove our main result that guarantees the existence of equilibria. Assume P is the true preference profile. Let $E(I(P))$ be the set of equilibrium outcomes of $I(P)$ and $O(P)$ the set of all matchings that are outcomes of the G–S algorithm applied to strategy profiles of the form $R = (P(m_1), \dots, P(m_n); R(w_1), \dots, R(w_n))$, in which all men state truth but all women can state their preferences in any (non-equilibrium) fashion.

Theorem 1. $E(I(P)) = S(P) \cap O(P)$ for any preference profile P . Therefore $E(I(P)) \neq \emptyset$ since $M(P)$ always belongs to $S(P) \cap O(P)$.

Proof. We know from Roth’s result (1984) that $E(I(P)) \subset S(P) \cap O(P)$. Here we have to show that $S(P) \cap O(P) \subset E(I(P))$.

Take any x in $S(P) \cap O(P)$. There is a preference profile R such that $x = M(R)$ is stable with respect to P . We record all intermediate outcomes when the G–S algorithm is applied to R . When the G–S algorithm is applied to R , it produces for each woman w_j a sequence S_j of men:

$$m_1^j, m_2^j, \dots, m_{i(j)}^j = x^{-1}(w_j),$$

which consists of those men whom w_j has ever kept as her suitors at some step of the algorithm. The order of the men in S_j is the actual order by which w_j accepted (and subsequently rejected) them with the last man being w_j ’s eventual partner in x . Construct a preference relation $Q(w_j)$:

$$x^{-1}(w_j)Q(w_j) \dots Q(w_j)m_{i(j)}^jQ(w_j)m_1^jQ(w_j) \dots,$$

in which men in S_j are ranked in the reverse order as in S_j , and all others are ranked below them. Consider the strategy profile $Q = \{P(m_1), \dots, P(m_n); Q(w_1), \dots, Q(w_n)\}$, in which each man m ,

states the truth $P(m_i)$, and each woman w_j states $Q(w_j)$ as constructed above. It is easy to show by induction that when the G–S algorithm is applied to Q , the intermediate outcome after each step is identical to that occurring after the same step when the G–S algorithm is applied to Q . In particular, the final outcome is still $x = M(Q)$. We now show that Q is a Nash equilibrium.

For each woman w_j , let $\mathcal{M}_j(Q)$ be set of the men who have ever proposed to w_j . We claim that $x = M(Q)$ is stable with respect to P if and only the following is true:

(*) for each w_j , $x^{-1}(w_j)$ is the best man among $\mathcal{M}_j(Q)$ according to $P(w_j)$.

If condition (*) is not satisfied, then there exist some w_j , and some m_i in $\mathcal{M}_j(Q)$ such that $m_i P(w_j)x^{-1}(w_j)$. But since m_i is rejected by w_j before the finally ends up with $x(m_i)$, it must be true that $w_j P(m_i)x(m_i)$. Therefore, x is unstable for P . On the other hand, if x is unstable for P , then we can find m_i and w_k such that $w_k P(m_i)x(m_i)$, and $m_i P(w_k)x^{-1}(w_k)$. But $w_k P(m_i)x(m_i)$ implies that $w_k \in \mathcal{M}_k(Q)$. This, together with $m_i P(w_k)x^{-1}(w_k)$, violates condition (*).

To show that Q is a Nash equilibrium, we need to check that no woman can make a unilateral profitable deviation. Suppose any woman w_j makes a deviation $Q'(w_j)$ from Q while others still state the same things as in Q . Denote this preference profile by Q' . Let $\mathcal{M}_j(Q')$ be the set of her proposers when the G–S algorithm is applied to Q' . If we can show that $\mathcal{M}_j(Q')$ is a subset of $\mathcal{M}_j(Q)$, then condition (*) means that w_j cannot make a profitable deviation $Q'(w_j)$ from Q . The reason why $\mathcal{M}_j(Q') \subset \mathcal{M}_j(Q)$ is as follows. If w_j did not receive proposal from m_i in Q , then m_i must have been accepted by his eventual partner $x(m_i) = w_k$, whom m_i prefers to w_j . Therefore, m_i still proposes to w_k before he will propose to w_j even in Q' . Since in Q' w_k states $Q(w_k)$ that has $m_i = x^{-1}(w_k)$ on top, she will not let m_i go. Hence m_i will not propose to w_j in Q . \square

Theorem 1 shows that although not all stable matchings are supported as equilibrium outcomes in the G–S algorithm, any stable matching that is supported by some misstatement of preferences can be supported as an equilibrium outcome. In particular, the M-optimal stable matching is such a stable matching. Hence the G–S algorithm always yields stable matchings as equilibrium outcomes. To summarize, we have the inclusion relationship:

$$\{M(P)\} \subset E(\Gamma(P)) \subset S(P), \text{ for any } P.$$

Example 1 shows that second inclusion can be strict. The first one also can be strict, and in fact it is frequently so as the following result indicates.

Theorem 2. For any preference profile P , if the truth is not an equilibrium strategy profile for $\Gamma(P)$, then $\{M(P)\} \neq E(\Gamma(P))$.

Proof. If the truth P is not an equilibrium strategy profile for $\Gamma(P)$, then some woman, say w_j , will find it profitable to deviate (no man wants to deviate since it is always a dominant strategy for him to state the truth). Let $R(w_j)$ be any of her best responses to P . Denote R the preference profile in which everyone has the true preferences except w_j has $R(w_j)$.

When the G–S algorithm is applied to R , it yields $x = M(R)$ as the outcome, which is different from $M(P)$ by assumption. If we show that R satisfies condition (*) in the proof of Theorem 1, then Theorem 1 implies $x \in E(\Gamma(P))$. It is obvious that $x^{-1}(w_k)$ is the best man among $\mathcal{M}_k(P)$ according to $P(w_k)$ for each woman $w_i \neq w_j$ since she states the truth in R . Suppose this is not true for w_j . Then there is some man m_i who proposed to w_j and was nevertheless rejected even though $m_i P(w_j)x^{-1}(w_j)$. Now if w_j states the preference relation $R'(w_j)$ that is identical to $R(w_j)$ but has m_i on top while all others still state the truth, the G–S algorithm will yield an outcome that matches w_j with m_i . But this contradicts the assumption that $R(w_j)$ is one of w_j 's best responses to P . \square

3. Discussions

In this note we have proved the nonemptiness of equilibria $E(I(P))$ in the G–S algorithm for every true preference profile P . Combined with Roth’s result, it shows that the G–S algorithm always yields stable matchings. But $E(I(P))$ usually contains many stable matchings other than the M-optimal one. It will be interesting to have some plausible selection of equilibria. We have tried the following line of reasoning. Since it is the women who manipulate in the G–S algorithm, it is natural to ask that if they can coordinate their manipulations in an optimal fashion? More precisely, does there exist a matching $E(I(P))$ that weakly dominates all other matchings in $E(I(P))$ from the women’s point of view? Such a matching, if exists, seems more likely to emerge than others. Although we have verified that this is true for the simplest marriage problem of three men and three women, we still do not know enough about the structure of $E(I(P))$ to provide an answer for general marriage problems. We leave it as a conjecture for future research.

References

- Dubin L E and D A Freedman, 1981, Machiavelli and the Gale–Shapley Algorithm, *American Mathematical Monthly* 88, 485–494
- Gale, D. and L. Shapley, 1962, College admissions and the stability of marriage, *American Mathematical Monthly* 69, 9–15
- Gale, D. and L. Sotomayor, 1985, Mr Machiavelli and the stable matching problem, *American Mathematical Monthly* 92, 261–268.
- Roth, A E., 1982, The economics of matching: Stability and incentives, *Mathematics of Operations Research* 7, 617–628.
- Roth, A.E., 1984, Misrepresentation and stability in the marriage problem, *Journal of Economic Theory* 34, 383–387