Two-Sided Matching: The Marriage Model

Jonathan Levin

September 2012

These notes describe two "marriage model" of two-sided matching. Because there is no textbook, these notes are meant to supply the technical details of the model, and some examples, but I haven't included the chatty stories from the lecture slides.

1 The Marriage Model

In the marriage model, there are two sides of the market. We refer to them as the men and the women. We let M denote the set of men, and W denote the set of women. Let m, w denote typical members of the two groups. We are interested in finding a one-to-one matching of men and women, where each man is matched with a single woman, and each woman with a single man.

Each man has a strict preference ranking over the women, and conversely each woman has a strict preference ranking over the men. We say that woman w is *acceptable* to man m if m prefers being matched to w to being unmatched.

Example 1 *Suppose $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$. Then a possible preference ranking for m_1 is:*

$$m_1 : w_2 \succ w_3 \succ \varnothing \succ w_1$$

which means m_1's top choice is w_2, his second choice is w_3, and he would prefer to be unmatched ($\varnothing$) to matching with w_1. Sometimes we will write the preference ranking simply as a list:

$$m_1 : w_2, w_3, \varnothing, w_1$$

A *matching* is a set of pairs (m, w) such that each individual has one partner. We allow for men or women to be unmatched by writing $(m, \varnothing)$ or $(\varnothing, w)$.[1]

Example 2 *Suppose $M = \{m_1, m_2, m_3\}$ and $W = \{w_1, w_2, w_3\}$. Then two possible matchings are:*

$$\begin{aligned} \textit{Matching 1} \quad &: \quad (m_1, w_2), (m_2, w_3), (m_3, w_1) \\ \textit{Matching 2} \quad &: \quad (m_1, w_3), (m_2, \varnothing), (m_3, w_2), (\varnothing, w_1) \end{aligned}$$

[1] If m is unmatched, we alternatively can write (m, m) — i.e. m is matched to himself, rather than no one.

2 Stable Matchings

A matching is *stable* if (a) every individual is matched with an acceptable partner; and (b) there is no man-woman pair, each of whom would prefer to match with each other rather than with their assigned partner. If such a pair exists, we call it a *blocking pair* and we say that the match is *unstable.*

Example 3 *Suppose* $M = \{m_1, m_2, m_3\}$ *and* $W = \{w_1, w_2, w_3\}$*, and preferences are*

$$\begin{array}{ll} m_1 : w_1, w_2, w_3 & w_1 : m_1, m_2, m_3 \\ m_2 : w_1, w_3, w_2 & w_2 : m_2, m_3, m_1 \\ m_3 : w_2, w_3, w_1 & w_3 : m_3, m_1, m_2 \end{array}$$

In this example, everyone prefers being matched to being unmatched so we omit preferences over not matching ($\varnothing$*).*

There are two stable matches:

$$\begin{array}{lll} \textit{Matching 1} & : & (m_1, w_1), (m_2, w_2), (m_3, w_3) \\ \textit{Matching 2} & : & (m_1, w_1), (m_2, w_3), (m_3, w_2) \end{array}$$

To see why these are both stable, note that in the first match, w_1, w_2 *and* w_3 *are all getting their top choice, so no woman wants to form a blocking pair, making the match stabel. In the second match,* m_1 *and* m_3 *get their top choices, so neither of them would want to block. In contrast* m_2 *would like to convince* w_1 *to form a blocking pair, but she likes* m_1 *best so she will not be part of a blocking pair. Given that* m_2 *will want to stay with* w_2*, making the matching stable.*

Now consider the alternative matching:

$$\textit{Matching 3: } (m_1, w_2), (m_2, w_1), (m_3, w_3).$$

This matching is not stable. Why? Because m_1 *and* w_1 *would want to form a blocking pair.*

3 Deferred Acceptance Algorithm

We can use the deferred acceptance algorithm to find stable matchings. The (man-proposing) deferred acceptance algorithm works as follows:

- Each man proposes to the highest woman on his list
- Women make a "tentative match" based on their preferred offer, and reject other offers, or reject all their offers if none are acceptable.
- Each rejected man removes the woman who rejected him from his list, and makes a new offer.
- The process continues until their are no more rejections or offers, at which point we freeze the tentative matches.

The algorithm also can be run with the women proposing to the men.

Theorem 4 *The outcome of the deferred acceptance algorithm is a stable one-to-one matching.*

Proof. First, we note that the algorithm must terminate so long as the number of men and women is finite. When it terminates, we have a matching. Let's prove it is stable. To see this, suppose the DA algorithm matches m and w. However, m prefers w' to w. During the algorithm, m proposes to women in order of preference, so if he ended up with w, he must have proposed to w' at some point, and was rejected. Now, when w' rejected m, she must have had an offer in hand that she preferred to m. As the algorithm proceeded, she may have gotten still better offers, but she could not have gotten worse off, so at the end of the algorithm, she must have ended up with someone she likes better than m. Therefore, she will not form a blocking pair with m. It follows that at the end of the algorithm, there will be no man m who can form a blocking pair with a woman he likes better than his DA partner. Hence the matching is stable. *Q.E.D.*

Note that if we run the woman-proposing version of the DA algorithm, we will also get a stable matching, but not necessarily the same one!

Example 5 *(Same as above) Suppose* $M = \{m_1, m_2, m_3\}$ *and* $W = \{w_1, w_2, w_3\}$*, and preferences are*

$$\begin{array}{ll} m_1 : w_1, w_2, w_3 & w_1 : m_1, m_2, m_3 \\ m_2 : w_1, w_3, w_2 & w_2 : m_2, m_3, m_1 \\ m_3 : w_2, w_3, w_1 & w_3 : m_3, m_1, m_2 \end{array}$$

Suppose we run the man-proposing DA algorithm:

	Round 1	*Round 2*
Offers	$m_1 \to w_1$	
	$m_2 \to w_1$	$m_2 \to w_3$
	$m_3 \to w_2$	
Matches	$(m_1, w_1), (m_3, w_2)$	$(m_1, w_1), (m_2, w_3)\,(m_3, w_2)$

and the algorithm ends after two rounds.

If alternatively we run the woman-proposing DA algorith:

	Round 1
Offers	$w_1 \to m_1$
	$w_2 \to m_2$
	$w_3 \to m_2$
Matches	$(m_1, w_1), (m_2, w_2)\,, (m_3, w_3)$

the algorithm ends after one round, at a different match.

4 Optimal Stable Matchings

In the example above, note that if we compare the stable matching that comes from the man-proposing DA algorithm and the stable matching that comes from the woman-proposing DA algorithm, the men like the first one better and the women like the second one better. This turns out to be a general property of the algorithm.

We say that a stable matching is *man-optimal* if every man prefers his partner to any partner he could have in some other stable matching. Similarly a stable matching is *woman-optimal* if every woman prefers her partner to any partner she might have in some other stable matching. (Note: A man-optimal stable matching does not mean that every man gets his first choice woman, because matching some man m with his first choice woman w might lead to instability.)

Theorem 6 *The man-proposing deferred acceptance algorithm results in a man-optimal stable matching. The woman-proposing deferred acceptance algorithm results in a woman-optimal stable matching.*

Proof. Fix a set of men and women and their preferences. Suppose we enumerate all stable matchings (at least one exists, from the above result). We say that w is *possible* for m if m is matched with w in one of the stable matchings. To prove the result, we will show that if we run the man-proposing DA, no man is ever rejected by a woman who is possible for him. Since men move sequentially down their preference list during the algorithm, and the end result of the algorithm is a stable match, this means that every man must get his favorite woman among those who are possible for him.

To establish that no man is rejected by a woman who is possible for him during the DA, we argue by contradiction. Suppose that during the DA algorithm, the first time a man is rejected by a woman who is possible for him occurs when w rejects m in favor of some other man m'. Note that this presumption implies that (a) because w is possible for m, there exists a stable matching exists where m and w are paired; and (b) w likes m' better than m.

Now, for statement (a) to be true, then the stable matching that pairs (m, w) must pair m' with some woman w' who he likes better than w. Otherwise (m', w) would form a blocking pair. Hence w' is possible for m' and he likes her better than w. But then for m' to reach the point in the DA where he is proposing to w, he already must have proposed to w' and been rejected. That is, *before* w rejected m in the DA, w' rejected m' — contradicting our presumption that m was the first man to be rejected by a woman who was possible for him. It follows that there can be no first time when a man is rejected by a possible woman during the DA algorithm, and hence it never happens. *Q.E.D.*

It is also possible to show (try it as an exercise) that the man-proposing deferred acceptance algorithm results in the stable matching that is *woman-pessimal*, that is, in which every woman gets her worst possible partner across

all stable matchings. Similarly, the woman-proposing deferred acceptance algorithm results in the ma

5 The Set of Stable Matchings

We already have seen an example in which there are multiple stable matchings. It is also possible to have situations where there is a a unique stable matching. How do we know when there is a unique stable matching. In that case, the man-proposing and woman-proposing DA algorithm lead to the same stable matching.

Theorem 7 *There is a unique stable matching if and only if the man-proposing and woman-proposing deferred acceptance algorithms lead to the same (stable) matching.*

Proof. ($\Rightarrow$) Suppose there is a unique stable matching. Since both the man-proposing and woman-proposing DA algorithm lead to a stable matching, they must both find the (same) unique one.

($\Leftarrow$) Suppose the two versions of the DA lead to the same stable matching. Then from the above result, every man and also every woman have their best possible partner across all stable matchings. So if there were some other stable matching, all the men and women would be weakly worse off, and some strictly worse off than under the DA stable matching, meaning partners could be rearranged to make everyone better off. But then the non-DA candidate stable matching must be blocked (think about exactly why!). *Q.E.D.*

Here is an example of a situation with a unique stable matching.

Example 8 *Suppose* $M = \{m_1, m_2, m_3\}$ *and* $W = \{w_1, w_2, w_3\}$*, and preferences are*

$$\begin{array}{ll} m_1 : w_1, w_2, w_3 & w_1 : m_1, m_2, m_3 \\ m_2 : w_1, w_2, w_3 & w_2 : m_1, m_3, m_1 \\ m_3 : w_1, w_2, w_3 & w_3 : m_3, m_1, m_2 \end{array}$$

Again, no one wants to be unmatched, so we ignore preferences over not matching.

The unique stable matching is

$$(m_1, w_1), (m_2, w_3), (m_3, w_2)$$

You can show this by running through the two versions of the DA algorithm. Alternatively, you can note that all men have the same preference order. So to have stability, woman w_1 *must get her first choice man creating the pair* (m_1, w_1)*. Then woman* w_2 *must get her first choice of the remaining men, creating the pair* (m_3, w_2)*, which leaves* (m_2, w_3)*.*

An interesting additional result is that if there are multiple stable matchings, these matchings can involve a re-arrangement of partners among the matched men and women, but they cannot involve different men or women being left un-matched. For reasons discussed in class, this is called the *rural hospitals theorem.*

Theorem 9 *Fix a set of participants and preferences. If there are multiple stable matchings, the set of men and women who are unmatched is the same in all these matchings.*

Proof. Let M, W be the sets of men and women matched in the man-optimal stable matching (which is also woman-pessimal), and M',W' be the sets of men and women matched in some other stable matching. Then:

- Any man in M' must also be matched in the man-optimal stable matching, so $M' \subseteq M$, and $|M'| \leq |M|$.
- Any woman in W must be matched in any other stable matching, so $W \subseteq W'$ and $|W| \leq |W'|$.
- In any stable matching, the number of matched men equals the number of matched women, so $|W| = |M|$ and $|W'| = |M'|$.
- Therefore $|M'| = |M| = |W'| = |W|$.

It follows that $M' = M$ and also that $W' = W$. *Q.E.D.*

6 The DA Algorithm and Incentives

So far, we have taken preferences as given and considered how to find stable matchings given these preferences, and also what the possible stable matchings might look like. To actually run a market, however, we most likely need to ask people what their preferences are, and then there is a question of whether they will report them truthfully, or instead try to manipulate or "game" the matching process. To study the incentives for participants in such a setting, we instroduce the idea of a *matching mechanism.*

A *matching mechanism* asks all participants to report their preference orderings, and maps reported preferences into a matching. For instance, one possible mechanism is to ask men and women to report their preferences, run the man-proposing deferred acceptance algorithm given these reported preferences, and assign people according to the algorithm's output. We say that a matching mechanism is *strategy-proof* if each participant finds it optimal to be truthful regardless of the preferences reported by the other participants.

Theorem 10 *The man-proposing deferred acceptance algorithm is strategy-proof for the men.*

Proof. To prove this, we fix the reports of all the women and all the men except for m_1. We then consider that whatever preferences m_1 considers making, there is a sequence of modifications to this report that leave him no worse off, and that culminate in the truthful report. To see this, suppose that man m_1 is considering making some preference report that, given everyone else's report, will yield a particular matching — call it x — in which m_1 gets w_1. The following changes can only improve m_1's outcome.

- Reporting that w_1 is his only acceptable woman. If before w_1 was not first on m_1's list, then m_1 will effectively be telling the algorithm to "skip" some rounds of asking and getting rejected. But the DA will still result in m_1 getting w_1.
- Reporting honestly, but only the part of m_1's preference list up until w_1. In the DA, it can't hurt to ask women m_1 likes better than w_1. One of them might end up accepting him, or alternatively, he will end up asking w_1 and getting her.
- Reporting honestly with no truncation. This won't affect the DA relative to the above strategy because w_1 won't reject m_1 and so he will never get farther down his list.

It follows that m_1 does at least as well with an honest report as with any other report. *Q.E.D.*

Sadly, the deferred acceptance algorithm does not give perfect truth-telling incentives to all parties.

Theorem 11 *The man-proposing deferred acceptance algorithm is not strategy-proof for the women.*

Proof. We can prove this using the example from class. There are two men and two women with preferences:

$$\begin{array}{ll} m_1 : w_1, w_2 & \quad w_1 : m_2, m_1 \\ m_2 : w_2, w_1 & \quad w_2 : m_1, m_2 \end{array}$$

If everyone reports truthfully, then under the man-proposing DA we end up with (m_1, w_1) and (m_2, w_2). However, if w_1 reports that her prefences are instead: $m_2 \succ \varnothing \succ m_1$, then in the man-proposing DA, she will reject m_1's offer in the first round, he will propose to w_2, who will accept him and reject m_2, who will then propose to w_1 and the final result will be (m_1, w_2) and (m_2, w_1). This outcome is preferred for w_1, making the misreport beneficial. *Q.E.D.*

One of the interesting results in matching theory is that the opportunities for this type of misreporting are most important when there are only a small number of men and women. In a "large" market, it can be shown that the opportunities for participants to gain by misreporting under the DA algorithm are vanishingly small.

7 Discussion and Extensions

1. (Stability and Pareto Efficiency) The notion of stability should be distinguished from the notion of Pareto Efficiency that we will study in the house allocation problem. In the marriage model, we can say that a matching is *Pareto efficient for the men* if no two men would like to swap wives. The notion of Pareto efficiency for the men pays no mind to the preferences of the women. When we look at school choice, we will set that a matching can be stable, but not Pareto efficient for the men, or conversely Pareto efficient for the men, but not stable.

2. (Many-to-one matching) We have looked at a model of one-to-one matching. Later when we consider school choice, and discuss medical residents, we will consider models of many-to-one matching, where a school might accept several students or a hospital might hire several residents. Many of the results shown here will carry over to that setting, provided that the schools or hospitals have preferences that satisfy a certain type of "substitutability".

3. (Harder Math) For the serious math types in the class, you may be intrigued to learn that many of the results here can be obtained in a very concise and elegant fashion using ideas from lattice theory. This approach also allows one to connect matching theory to auction theory (we'll touch on this later). If you take the honors micro theory class from Fuhito Kojima, you're likely to see some of these ideas, as well as other extensions of matching theory.

8 References

Gale, David and Lloyd Shapley, "College Admissions and the Stability of Marriage," American Mathematical Monthly, 1962.

Roth, Alvin E. and Marilda Sotomayor, Two-Sided Matching: A Study in Game-Theoretical Modeling and Analysis, Cambridge University Press, 1990.