UPenn NETS 412: Algorithmic Game Theory February 22, 2018

Lecture 11

Lecturer: Bo Waggoner *Scribe: Aaron Roth, Bo Waggoner*

Stable Matchings

In this lecture, we'll consider a model of 1950's dating. Although this is the metaphor we will use, *stable matchings* are an extremely useful object, and are used in practice to among other things assign graduating medical students to residencies. In general, the setting we describe is important in *two sided markets*, in which both sides have preferences over the other, and money cannot be used as (the primary) medium of exchange.

Let M and W denote sets of *men* and *women* respectively. Assume $|M| = |W| = n$.

Definition 1 *A **matching** $\mu : M \cup W \to M \cup W$ is an assignment of men to women so that each man is assigned to exactly one woman and vice versa. For each $m \in M$ and $w \in W$, $\mu(m) = w$ if and only if $\mu(w) = m$.*

As in last lecture, we model each agent on one side of the market as having a strict preference ordering $\succ$ over the other side of the market. Specifically, each $m \in M$ has a strict preference ordering $\succ_m$ over the set W, and each $w \in W$ has a strict preference ordering $\succ_w$ over the set M. Recall that we write $m \succ_w m'$ if w strictly prefers m over m'.

Just as in the exchange problem we considered last time, we have two desiderata when coming up with a matching algorithm:

1. We would like the matching that we compute to be *good* in some sense, and
2. We would like to incentivize participants to reveal their true preferences to the mechanism.

Just as in the allocation problem, we will look for a solution here that does not use monetary payments (dating markets with payments being outside the scope of this class). We will be able to achieve 1, and to a limited extent 2.

First of all, we need to define what a "reasonable" matching is. In particular, we will ask that a matching (once suggested) is *stable* – i.e. that it is somehow robust to unilateral deviations among couples. If our matchings are not stable, there isn't much reason to suspect that people will follow our suggestions.

Definition 2 *Given a matching μ, a pair (m, w) with $m \in M$ and $w \in W$ is a **blocking pair** if $\mu(m) \neq w$ and:*

$$w \succ_m \mu(m) \quad \text{and} \quad m \succ_w \mu(w)$$

If there is a blocking pair (m, w), then they would both be happier by abandoning μ and just matching to each other.

Definition 3 *A matching μ is **stable** if it has no blocking pairs.*

Stability is a minimal requirement on a "reasonable" matching we might suggest – it is an equilibrium like property. We might later ask to compute the "best" stable matching in some sense, but it is not even clear at the moment that *any* stable matching need exist!

But one always does:

Theorem 4 (Gale and Shapley) *For any set of preferences $(\succ_{m_1}, \ldots, \succ_{m_n}, \succ_{w_1}, \ldots, \succ_{w_n})$, a stable matching μ exists.*

Furthermore, the deferred acceptance algorithm finds it (in at most n^2 rounds).

Algorithm 1 The Deferred Acceptance Algorithm (Male Proposing Version)

DeferredAcceptance($\succ$):

Initially, $\mu(m) = \emptyset$ for all $m \in M$. (i.e. nobody is yet matched).

Each man $m \in M$ *proposes* to his most preferred $w \in W$. For each woman $w \in W$, let m' be her most preferred man among the set that proposed to her, and set $\mu(m) \leftarrow w'$, $\mu(w') \leftarrow m$. All other men are *rejected* (and hence unmatched).

while There exists any unmatched man $m \in M$: **do**

m **proposes** to his most preferred $w \in W$ that he has not yet proposed to.

If $m \succ_w \mu(w)$, then $\mu(\mu(w)) \leftarrow \emptyset$ and $\mu(w) \leftarrow m$, $\mu(m) \leftarrow w$ (i.e. w rejects her current match and instead matches to m).

Else, m is rejected (the matching is not changed).

end while

Return μ

We will prove this theorem algorithmically, by analyzing the (male proposing) deferred acceptance algorithm. This algorithm iteratively builds up a matching μ, in which initially everyone is unmatched. We write $\mu(m) = \emptyset$ to denote that m is unmatched. It is called the "deferred" acceptance algorithm because women (who are proposed to) can *tentatively* agree to be matched to men (who propose), but can later revoke the agreement, in which case the man reverts to being unmatched.

Proof We begin by showing that the DA algorithm halts in at most n^2 rounds of the while loop, and returns a full matching. Then, we'll show that it's stable.

First, note that all men are matched if and only if all women are matched, as there are the same number of each. So if the while loop halts, it returns a full matching. Second, note that each man never proposes to the same woman twice, so after being chosen at most n times in the while loop, he has proposed to every woman. So after at most n^2 loops (n loops per man), every woman has received at least one proposal. This implies every woman is matched, since each becomes matched after receiving her first proposal and remains matched forever after.

Now, to show stability, we show that the final matching μ cannot have any blocking pairs. Consider any m, w who are not matched in μ, where $w \succ_m \mu(m)$, i.e. m would rather match to w than his partner in the matching μ. Then during the DA algorithm, he proposed to w at some point (as he ultimately proposed to $\mu(m)$, whom he preferred less). Yet w did not accept him, so she must have rejected him in favor of some m' such that $m' \succ_w m$. Similarly, if she did not ultimately match to m', then she rejected him in favor of someone even more preferred, and so on until $\mu(w)$, so we must have $\mu(w) \succ_w m$. So (m, w) cannot be a blocking pair. ■

We now turn our attention to the quality of the matching produced. What can we say about it, other than stability? To do so, we have to define who is *achievable* for whom in a stable matching. Clearly, not everybody can always be matched to their first choice partner!

Definition 5 *Fix a set of preferences. For $m \in M$ and $w \in W$, we say that w is **achievable** for m if there exists a stable matching μ such that $\mu(m) = w$.*

*Say w is the **best achievable** for m if w is achievable and, for all achievable w', $w \succeq_m w'$.*

*Say w is the **worst achievable** for m if w is achievable and, for all achievable w', $w \preceq_m w'$.*

(The definitions are exactly analogous with M and W reversed.)

Definition 6 *A matching μ is **male optimal** if every $m \in M$ is matched to his best achievable partner. It is **male pessimal** if every $m \in M$ is matched to his worst achievable partner. **Female optimal** and **pessimal** are defined analogously.*

We will show that it is good to be on the proposing side of the market, and bad to be on the "proposed to" side.

Theorem 7 *The stable matching μ output by the male-proposing deferred acceptance algorithm is* male optimal.

Proof In the first round, if w rejects m, then w is not achievable for m, because w prefers some m' who proposed to w in the first round because his first choice is w.

Now by induction, in the while loop, if w rejects m, then w is not achievable for m: w prefers some m' who has proposed to w, and by induction, none of the previous choices of m' were achievable. So in any stable matching, we could not have $\mu(m) = w$ because (m', w) would be a blocking pair. That is, w prefers m' to m, and m' prefers w over any other achievable partner of his.

So we've shown that, whenever m is rejected by some w, she is not achievable for him. The woman he's matched to in DA is his favorite woman who doesn't reject him, hence his favorite achievable woman. ■

Of course, the theorem applies equally well to the (symmetrically defined) *female proposing* deferred acceptance algorithm and *female optimal* matchings.

We now turn to the incentive properties of the deferred acceptance algorithm. We show that in the male-proposing deferred acceptance algorithm, reporting their true preferences is a dominant strategy for the men.

Theorem 8 *The male proposing deferred acceptance algorithm is dominant strategy incentive compatible. (i.e. reporting their true preferences $\succ_m$ is a dominant strategy for each $m \in M$).*

Proof Suppose otherwise; i.e. there is a set of preferences $\succ = (\succ_{m_1}, \ldots, \succ_{m_n}, \succ_{w_1}, \ldots, \succ_{w_n})$ and (without loss of generality) a deviation $\succ'_{m_1}$ such that if $\mu = DE(\succ)$ and $\mu' = DE(\succ')$ (where $\succ' = (\succ'_{m_1}, \succ_{-m_1})$), then:

$$\mu'(m_1) \succ_{m_1} \mu(m_1).$$

Note that we must also have that μ is stable and male optimal with respect to preferences $\succ$, and μ' is stable and male optimal with respect to preferences $\succ'$. We define two sets. Let:

$$R = \{m : \mu'(m) \succ_m \mu(m)\}$$

i.e. the set of men who prefer their match in μ' to their match in μ. Note that $m_1 \in R$ by assumption. Let

$$T = \{w : \mu'(w) \in R\}$$

i.e. women whose partners in μ' are in R (and thus prefer them to their match in μ). We will show:

1. $w \in T \Leftrightarrow \mu(w) \in R$. (i.e. if a woman's partner in μ' prefers μ' to μ, so does her partner in μ), and from this derive that:
2. There exists a $w_\ell \in T$ and a $m_r \in R$ such that (w_ℓ, m_r) form a blocking pair in μ' with respect to $\succ'$, a contradiction.

We start with the first claim:

Claim 9

$$w \in T \Leftrightarrow \mu(w) \in R$$

Proof For any $m \in R$, let $w = \mu'(m) \in T$. Let $m' = \mu(w)$ be w's partner in μ. If $m' = m_1$, we are done. Hence, we can assume $m' \neq m_1$, and therefore that $\succ_{m'} = \succ'_{m'}$. Since $m \in R$, we know that:

$$w = \mu'(m) \succ_m \mu(m)$$

Since μ is stable with respect to $\succ$, it must be that:

$$\mu(w) = m' \succ_w m$$

But because μ' is stable with respect to $\succ'$, it must be that:

$$\mu'(m') \succ_{m'} \mu(m') = w$$

and hence $m' \in R$ as we wanted ■

Next, we show the second claim, which leads to our contradiction:

Claim 10 *There exists a $w_\ell \in T$ and a $m_r \in R$ such that (w_ℓ, m_r) form a blocking pair in μ' with respect to $\succ'$*

Proof Since for every $m \in R$, $\mu'(m) \succ_m \mu(m)$, by the stability guarantee, it must be that for all $w \in T$:

$$\mu(w) \succ_w \mu'(w).$$

Thus, when running DE($\succ$), it must be that every $m \in R$ proposes to $\mu'(m)$, and is rejected by $\mu'(m)$ at some round. Let m_ℓ be the *last* $m \in R$ who proposes during the DE algorithm.

This proposal must be to $\mu(m_\ell) \equiv w_\ell$. By the first claim, since $m_\ell \in R$, $w_\ell \in T$. It must be that w_ℓ rejected $\mu'(w_\ell)$ at a strictly earlier round (since m_ℓ is the last $m \in R$ to propose), and hence when m_ℓ proposes to w_ℓ, w_ℓ rejects some $m_r \notin R$ such that:

$$m_r \succ_{w_\ell} \mu'(w_\ell) \tag{1}$$

Since m_r had proposed to w_ℓ before $\mu(m_r)$, it must be that:

$$w_\ell \succ_{m_r} \mu(m_r)$$

Note that $m_r \neq m_1$ (since $m_1 \in R$), and so $\succ_{m_r} = \succ'_{m_r}$. Hence, since $m_r \notin R$, we also know:

$$\mu(m_r) \succeq_{m_r} \mu'(m_r)$$

and hence:

$$w_\ell \succ_{m_r} \mu'(m_r)$$

Together with 1, this means that (m_r, w_ℓ) form a blocking pair for μ', which is a contradiction. ■ ■

Note that we have shown that it is a dominant strategy for the *men* to report their true preferences, but we have not shown this for the women. On the homework, you will show it is not. (In fact, it turns out to be the case that no algorithm can be truthful for all participants.)