NETS 412: Algorithmic Game Theory February 23, 2017

Lecture 12

Lecturer: Aaron Roth *Scribe: Aaron Roth*

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, and assign sorority pledges to sororities. 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 $\mu : M \cup W \rightarrow M \cup W$ *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.

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 money (payments to your partner are frowned upon in dating markets). 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 at the very least 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 *A matching* μ *is* unstable *if there exists an* $m \in M$ *and* $w \in W$ *such that* $\mu(m) \neq w$*, but:*

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

We call such an (m, w) *pair a* blocking pair *for* μ*. (A blocking pair witnesses instability because* m *and* w *could mutually benefit by leaving their proposed partners and pairing with one another).*

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 3 (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.*

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

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 $\mu(m) \leftarrow w'$. 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 $\mu(w) \leftarrow m$ (i.e. $w$ rejects her current match and instead matches to $m$). Else, $m$ is rejected.
  end while
  Return $\mu$
```

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 consider the behavior of the deferred acceptance algorithm. First, observe that it always halts and outputs some matching μ. This follows because every woman receives at least one proposal over the course of the algorithm. (If there is a woman without a proposal, there is an unmatched man, and the algorithm has not halted unless he has proposed to all women). Once a woman has received a proposal, she becomes matched, and stays matched for the rest of the algorithm (she always accepts her first proposal, and only subsequently rejects her match if she receives a better proposal). But since $|W| = |M|$, once all women are matched, all men are matched. Note also that the algorithm therefore halts after at most n^2 proposals, since no man ever proposes to the same woman twice.

We next observe that the final matching μ cannot have any blocking pairs. Suppose otherwise – i.e. there is a blocking pair (m_1, w_1) with $\mu(m_1) \neq w_1$, but $w_1 \succ_{m_1} \mu(m_1)$ and $m_1 \succ_{w_1} \mu(w_1)$. Since $w_1 \succ_{m_1} \mu(m_1)$, m_1 must have proposed to w_1 before he proposed to $\mu(m_1)$. Since $\mu(m_1) \neq w_1$, m_1 must have been *rejected* by w_1 in favor of some other man m'. Since women only ever change who they are matched to in favor of more preferred men, we must have:

$$\mu(w_1) \succeq_{w_1} m' \succ_{w_1} m_1$$

which contradicts $m_1 \succ_{w_1} \mu(w_1)$. This completes the proof. ■

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

Definition 4 *For $m \in M$ and $w \in W$, we say that w is* achievable *for m (and vice versa) if there exists a stable matching μ such that $\mu(m) = w$.*

Definition 5 *A matching μ is* male optimal *if for every achievable pair (m, w), $\mu(m) \succeq_m w$ – i.e. simultaneously for every $m \in M$, he he matched to his most preferred achievable $w \in W$. Similarly, we can define* female optimal *matchings, and male and female* pessimal *matchings. (A matching μ is female pessimal if for every achievable pair (m, w), $m \succeq_w \mu(w)$ – i.e. simultaneously for every $w \in W$, she is matched to her* least preferred *feasible $m \in M$.)*

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

Proof Suppose otherwise. In this case, there must be some first round k at which a man m is rejected by his most preferred achievable woman w, in favor of m'. Therefore, it must be that:

$$m' \succ_w m \tag{1}$$

Since w is achievable for m, there must be some stable matching μ such that $\mu(m) = w$ and $\mu(m') = w'$ (and hence w' is achievable for m').

We must also have:

$$w \succ_{m'} w' \tag{2}$$

since m' proposed to w, and can't have been rejected by any achievable woman (in particular w'), since by assumption, k was the first round at which a man was rejected by an achievable woman. But combining 1 and 2, we have:

$$m' \succ_w m \quad w \succ_{m'} w'$$

which means that (m', w) form a blocking pair for μ, contradicting the fact that μ is stable. This completes the proof. ■

Finally, we show:

Theorem 7 *The stable matching produced by the male-proposing deferred acceptance algorithm is female pessimal.*

Proof We will show that *every* male-optimal stable matching μ is female pessimal. Suppose otherwise.

Then, there exists some w with $\mu(w) = m$, and $m \succ_w m'$ for some other achievable man m'. In this case, there must exist a different stable matching μ' with $\mu'(m') = w$, and $\mu'(m) = w'$. But note that we must have:

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

because μ is male-optimal and w' is achievable for m. So (m, w) are a blocking pair for μ', which contradicts its stability. ■

Of course, the theorem is reversed if we were to use the (symmetricly defined) *female proposing* deferred acceptance algorithm.

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:

$$m_r \succ_{w_\ell} \mu'(w_\ell) \tag{3}$$

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 3, 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 that in fact it is not in general. Specifically, *no* algorithm can make it a dominant strategy for *both* the men and the women to report their true preferences, for every set of preferences reported by their opponents.