Econ 501B Lecture 12 10/2/18

Designing Matching Mechanism

In both a centralized and decentralized market, stated preferences of agents are needed to produce stable matchings. In fact, a matchmaker or algorithm depends on it. But what if agents don't report true preferences, but instead something else so as to get themselves a better match? Is it possible to to incentivize all agents to report preferences truthfully all of the time?

I Matching mechanisms in One-to-one MARKET

Define $\mathcal{E} = (T, B; (\succsim)_{i \in T \cup B})$. Assume a designer (match maker) knows agents in T and B but does not know their preferences.

Let $\mathcal{P}_i$ be the set of all complete and transitive preference relations.

Think of designer as asking each agent i to report a preference relation.

Let $\hat{\succsim}_i$ be the reported preference which may or may not be true (i.e. the true pref $\succsim_i$ for agent i)

Denote set of matches as $\mathcal{M}$ and so $M \in \mathcal{M}$.
M will depend on T, B.

Then $\mathcal{P}$ is set of all preferences and $\mathcal{M}$ set of all matches

Definition MATCHING MECHANISM

A function, that, for each input, describes the matching output that the algorithm produces

$$m : \prod_{i \in T \cup B} \mathcal{P}_i \to M$$

Formally, a matching mechanism m is a function m whose range is the set of all possible inputs $(T, B; (\succsim)_{i \in T \cup B})$, or, rather set of all preferences for each agent, and whose output is a matching of T and B.

Note: designer does not know true preference relations of the agents, he only knows what they report.

Example: T-proposing Deferred Acceptance Mechanism.

Notice difference in title. We say Mechanism and not algorithm because we apply the algorithm to reported preferences which may or may not be true.

We will get out as output a match. The match will be stable relative to the reported preferences. Key Question: When is such a matching, then, stable relative to true preferences in spite of reported preferences?

DEFINITION: Call mechanism $m: \prod_{i \in T \cup B} P_i \to m$ a stable mechanism if it produces:

> a stable matching relative to each reported preference.

Example: Let $T = \{t_1, t_2\}$ $B = \{b_1, b_2\}$

True preferences:	T-prop DAA would give:
$t_1 : b_1 > b_2 > t_1$	$M_{T0}(t_1) = b_1$
$t_2 : b_2 > b_1 > t_2$	$M_{T0}(t_2) = b_2$
$b_1 : t_2 > t_1 > b_1$	
$b_2 : t_1 > t_2 > b_2$	

Would there be an incentive to misreport preferences?

- Of course not for t agents who get their best match — that is if all agents are reporting truthfully, then no t agent has incentive to misrepresent preferences.

Suppose t_1, t_2, b_2 each tell the truth. Then b_1 has incentive to misrepresent.

If b_1 reports $\hat{\succsim}_{b_1}$ such that:

$$t_2 \mathrel{\hat{\succ}_{b_1}} b_1 \mathrel{\hat{\succ}_{b_1}} t_1$$

Then, T-proposing DA Algorithm will do the following:

Round 1: "$\rightarrow$" proposes

$t_1 \rightarrow b_1$ $\quad$ b_1 unmatched

$t_2 \rightarrow b_2$ $\quad$ b_2 accepts t_2

Round 2

$t_1 \rightarrow b_2$ $\quad$ b_2 accepts t_1

$t_2 \rightarrow b_2$ $\quad$ b_1 unmatched

Round 3:

$t_1 \rightarrow b_2$ $\quad$ b_1 matches with t_2

$t_2 \rightarrow b_1$ $\quad$ b_2 matches with t_1

When one b misreports, the matching becomes B-optimal.

Question: Can we design mechanism such that no agent has incentive to misreport?

Need to be more precise...

- Want no incentive before the match at the reporting stage.
- Want ex post that no one has incentive to deviate from a match (need stability)

DEFINITION: A strategy for agent i: $S_i = \mathcal{P}_i \to \mathcal{P}_i$
For every true preference relation (input) the output is a reported preference relation.

DEFINITION: A strategy is truthful if $S_i(\succsim_i) = \succsim_i$ for each $\succsim_i \in \mathcal{P}_i$ (or truthful reporting)

DEFINITION A strategy S_i is dominant for mechanism $m : \prod_{i \in T \cup B} \mathcal{P}_i \to m$ if for each preference relation $\succsim_i \in \mathcal{P}_i$ and for each $\hat{\succsim}_{-i} \in \mathcal{P}_j$ $_{j \in T \cup B \setminus \{i\}}$ such that

$$\underbrace{m(\underbrace{S_i(\succsim_i)}_{\text{my report}}, \underbrace{\hat{\succsim}_{-i}}_{\text{others report}})(i)}_{\substack{\text{who } i \text{ is matched with} \\ \text{when } i \text{ reports } S_i(\succsim_i) \text{ and} \\ \text{everyone else reports } \hat{\succsim}_{-i}}} \underbrace{\succsim_i}_{\substack{i\text{'s pref} \\ \text{relation}}} \underbrace{m(\underbrace{\hat{\succsim}_i}_{\text{my report}}, \hat{\succsim}_{-i})(i)}_{\substack{\text{who } i \text{ is matched with when } i \text{ reports } \hat{\succsim}_i \\ \text{and everyone else reports } \hat{\succsim}_{-i}}}$$

Definition: A mechanism is strategy proof if for each agent $i \in T \cup B$ the truthful strategy is dominant.

Question: Does there exist a stable matching that is strategy proof?

NO

Theorem 1 If $\min\{|T|, |B|\} \geq 2$ There is no stable strategy proof mechanism (Impossibility Thm) by E7

In other words, no stable matching mechanisms exists for which stating true preferences is a dominant strategy for every agent.

Proof:

Let $T = \{t_1, \dots, t_{|T|}\}$ $|T| \geq 2$
$B = \{b_1, \dots, b_{|B|}\}$ $|B| \geq 2$

Group Agents

Suppose:

$A(t_1) = A(t_2) = \{b_1, b_2\}$
$A(b_1) = A(b_2) = \{t_1, t_2\}$

$A(t) = \emptyset$ $t \in T \setminus \{t_1, t_2\}$
$A(b) = \emptyset$ $b \in B \setminus \{b_1, b_2\}$

Recall $A(i) = \{j \in J, j \succsim_i i\}$ or acceptable set. This proof says there is no stable strategy (for $\forall$ all strategy proof m's) proof matching mech., So, contradiction based proof only needs one deviation ($\exists$)

Preferences Over Agents $\succsim^*$

$b_1 \succ^*_{t_1} b_2 \quad t_2 \succ^*_{b_1} t_1$

$b_2 \succ^*_{t_2} b_1 \quad t_1 \succ^*_{b_2} t_2$

if μ is stable for $\succsim^*$

$\mu(t) = t$ if $t \in T \setminus \{t_1, t_2\}$

$\mu(b) = b$ if $b \in B \setminus \{b_1, b_2\}$

Stable for $\succ^*$ $\begin{cases} \mu(t_1), \mu(t_2) \in \{b_1, b_2\} \\ \mu(b_1), \mu(b_2) \in \{t_1, t_2\} \end{cases}$

So, if there were a stable strategy proof mechanism:

(1) For each $\succsim \in \prod_{i \in T \cup B} \mathcal{P}_i$: $m(\succsim)$ is stable relative to $\succsim$.

(2) For every $i \in T \cup B$ and every $\hat{\succsim}_{-i} \in \prod_{j \neq i} \mathcal{P}_j$

$m(\succsim_i, \hat{\succsim}_{-i})(i) \succsim_i m(\hat{\succsim}_i, \hat{\succsim}_{-i})(i)$

for all $\succsim_i \in \mathcal{P}_i$

To show this is false: show there exists $\succsim^* \in \prod \mathcal{P}_i$ such that we cannot have a stable match with (2) satisfied with $\succsim^*$.

T-Proposing DA:

$\mu^*_{TD}(t_1) = b_1$

$\mu^*_{TD}(t_2) = b_2$

$\mu^*_{TD}(b_1) = t_1$

$\mu^*_{TD}(b_2) = t_2$

B-proposing:

$\mu^*_{BD}(t_1) = b_2$

$\mu^*_{BD}(t_2) = b_1$

$\mu^*_{BD}(b_1) = t_2$

$\mu^*_{BD}(b_2) = t_1$

Since $m(\succsim^*) \in \{M^*_{TD}, M^*_{BD}\}$ is stable, we will need to go through both cases and show someone always has incentive to misrepresent.

Case A: $m(\succsim^*) = M^*_{TD}$

Let b_1 report $\hat{\succsim}_{b_1}$: $t_2 \hat{\succ}_{b_1} b_1$ and $b_1 \hat{\succ}_b t$ for all $t \in T \setminus \{t_2\}$

That is only t_2 is acceptable to b_1 and no other t is.

Assume all other agents report truthfully. We want to know: Does b_1 prefer $m(\hat{\succsim})$ to M^*_{TD}

Now, we do not know what $m(\hat{\succsim})$ is. We do know $M(\hat{\succsim})$ is stable relative to $\hat{\succsim}$.

- We will show that there is only one stable match for preference profile $\hat{\succsim}$, and it corresponds to T-proposing DA applied to $\hat{\succsim}$.

To show that T-prop DA applied to $\hat{\succsim}$ is only stable match, it suffices to show that when we apply T-prop DA to $\hat{\succsim}$ and B-prop DA to $\hat{\succsim}$ we get the same match.

T-prop DA on $\hat{\succsim}$

Round 1

$t_1 \to b_1$ $\quad$ b_1 rejects t_1

$t_2 \to b_2$ $\quad$ b_2 keeps t_2

Round 2

$t_1 \to b_2$ $\quad$ b_1 unmatched

$t_2 \to b_2$ $\quad$ b_2 keeps t_1

Round 3

$t_1 \to b_2$ $\quad$ b_2 accepts t_1

$t_2 \to b_1$ $\quad$ b_1 accepts t_2

B-prop DA on $\hat{\succsim}$

$b_1 \to t_2$ $\quad$ t_1 accepts b_2 $\quad \Rightarrow$ Gives exact same match as above

$b_2 \to t_1$ $\quad$ t_2 accepts b_1

Conclude: $m(\hat{\succsim}) = \hat{\mu}_{TD}$

$$m(\hat{\succsim}_{b_1}, \succsim^*_{-b_1})(b_1) = t_2 \succsim_{b_1} m(\succsim^*_{b_1}, \succsim^*_{-b_1}) = t_1$$

Then our assumption of truth telling being dominant is contradicted. That is, we assumed mechanism gave us something stable for all preferences. Then, we assumed that a truth telling preference generated matching would be preferred by all agents. Here we have our "counterexample"

But we are not done.

Case B: $m(\succsim^*) = \mu^*_{BD}$

- let t_1 report $\hat{\succsim}_{t_1}$ such that $b_1 \hat{\succsim}_{t_1} t_1$ and $t_1 \succsim_{t_1} b$ for all $b \neq 1$
- Let all other agents report truthfully

$$\hat{\succsim}_i = \succsim^*_i$$

Need to know what is $m(\hat{\succsim})$. We know that matchings generated from its preferences will be stable. Will show it is T-prop DAA applied to $\hat{\succsim}$, and that is the same as B-prop DA algorithm applied to $\hat{\succsim}$.

<u>T-prop DA Algo applied to $\hat{\succsim}$:</u>

$t_1 \to b_1$ b_1 accepts t_1
$t_2 \to b_2$ b_2 accepts t_2 DONE

<u>B-prop DA Algo</u>

$R_1 \left[\begin{array}{l} b_1 \to t_2 \\ b_2 \to t_1 \end{array} \right.$ t_2 accepts b_1, t_1 rejects b_2 | $R_2 \left[\begin{array}{l} b_1 \to t_2 \\ b_2 \to t_2 \end{array} \right.$ t_2 keeps b_2, b_1 no offer

$R_3 \left[\begin{array}{l} b_1 \to t_1 \\ b_2 \to t_2 \end{array} \right.$ t_1 accepts b_1, t_2 retains b_2

Then, both are the same! Hence: $\hat{\mu}_{TB} = m(\hat{\succsim})$

$$m(\hat{\succsim}_{t_1}, \succsim^*_{-t_1}) = b_1 \succ_{t_1} m(\succsim^*_{t_1}, \succsim^*_{-t_1}) = b_2$$

Again, we have a contradiction, for the same reason as case A.

Again, to prove this (that there is no stable strategy proof match) we assumed (i) all preferences in $\mathcal{P}_i$ give stable match in m. And, it is better <u>for</u> <u>every</u> agent to tell truth. Then, we presented case where no matter what, there was incentive to lie.

Theorem 2 Fix an environment $\mathcal{E} = (T; B, (\succsim_i)_{i \in T \cup B})$ with strict preferences that has at least two stable matches. Then, for any stable matching mechanism, there exists some agent $j \in T \cup B$ and some report $\hat{\succsim}_j$:

$$m(\hat{\succsim}_j, \succsim^*_{-j})(j) \succ^*_j m(\succsim^*_j, \succsim^*_{j-1})(j)$$

Moreover, $\hat{\succsim}_j$ can be chosen such that

$m(\hat{\succsim}_j, \succsim^*_j)(j)$ is j's most preferred achievable outcome.

$\{k : \mu(i) = k$ for stable $\mu\}$

In English (RS 88): When any stable mechanism is applied to a marriage market in which preferences are strict and there is more than one stable matching, then at least one agent can profitably misrepresent his preferences, assuming others tell the truth. This agent can misrepresent in such a way as to be matched to his or her most preferred achievable mate under the true preferences at every stable matching under the misstated preferences.

<u>Proof</u>: Fix environment $\mathcal{E} = (T, B; (\succsim))$ and assume preferences are strict and that there are at least two stable matches

- Then, b/c preferences are strict: $M^*_{TD} \neq M^*_{BD}$, as applied to $\succsim^*$

- $m(\succsim^*)$ → don't know what it is but it's stable
there exists some $I \in \{T \cup B\}$ such that
$$M^*_{ID} = m(\succsim^*)$$
WLOG, take $I = T$, choose $t \in T$ such that:
$$M^*_{TD}(t) \neq m(\succsim^*)(t).$$
We know such a t exists. If there were no such t, we could conclude $M^*_{TD} = m(\succsim^*)$

- Because $m(\succsim^*)$ is stable and M^*_{TD} is t-optimal (since prefs are strict):
$$M^*_{TD}(t) \succsim_t m(\succsim^*)(t)$$

- For this $t \in T$: $M^*_{TD}(t) \succ_t m(\succsim^*)(t)$

Want to show: If everyone reports truthfully, there is some misreport t can make that would get him M^*_{TD}

Let $\hat{\succsim}_t$: $\underbrace{\mu_{TD}(t) \succ_t t}_{\mu^*_{TD}(t) \neq t}$ and $t \succ_t b \;\forall\, b \neq \mu^*_{TD}(t)$

To show: for any match $\hat{\mu}$ that is stable on $(\hat{\succsim}_t, \succsim^*_{-t})$

$$\hat{\mu}(t) = \mu^*_{TD}(t)$$

(a) Note: μ_{TD} is still stable for $(\hat{\succsim}_t, \succsim^*_{-t})$

(b) Give match $\hat{\mu}$ that is stable for misreported preferences profile when everyone is telling the truth.

By IR, $\hat{\mu}(t) \in \{\mu^*_{TD}(t), t\}$ if $\hat{\mu}(t) \neq \mu_{TD}(t)$ then $\hat{\mu}(t) = t$

Stable + strict pref.

NB on LHT

$\mu^*_{TD}(t) \in B$

By rural hospital theorem, (When preferences are strict and any hospital does not fill its quota at some stable matching is assigned precisely the same set of students at every stable matching) there cannot be a situation where t is matched under DAA and not matched;

$$\Rightarrow \hat{\mu}(t) = \mu^*_{TD}(t).$$

Step 1 there is some t s.t. $\mu^*_{TD}(t) >_t m(\underset{\sim}{z}^*)(t)$

Step 2 there is report of t $\underset{\sim}{\hat{z}}_t$ s.t. $m(\underset{\sim}{\hat{z}}_t, \underset{\sim}{z}^*_{-t})(t) = \mu_{TD}(t)$