This note starts by looking at side-optimal stable matchings and provides a way to check whether there is a unique stable matching. It then discusses participants’ incentives in reporting their preferences to deferred acceptance: the proposing side wants to report their preferences truthfully, but the accepting side can sometimes profitably misreport their preferences. In the process we cover the lemma on rural hospitals. We finish by looking at Pareto efficiency of matchings.

1 The Best and Worst Stable Matching for a Market Side

We have seen in Lecture 3 that DA generates a stable matching. We have also seen that in the following example:

\[
\begin{array}{c|c|c|c}
  m_1 & m_2 & w_1 & w_2 \\
  w_1 & w_2 & m_2 & m_1 \\
  w_2 & w_1 & m_1 & m_2 \\
\end{array}
\]
DA selects the stable matching which is preferred by the proposing side. This holds true in general:

**Theorem 1.** *In deferred acceptance, an agent on the proposing side obtains an outcome that is at least as good as this agent’s outcome in any other stable matching.*

In other words, in no stable matching a man is married to a woman whom he prefers over his outcome in the men-proposing deferred acceptance. An optional proof is left as a hard (and optional) exercise; I will post a proof in a separate file.

In line with this result, let us call a stable matching *men-optimal* if each man is matched with the best woman the man can be matched with in any stable matching. Let us call a stable matching *women-optimal* if each woman is matched with the best man the woman can be matched with in any stable matching.

The above result is an example of a conflict of interests between the two sides of the market. Note that if a man and a woman are matched under one stable matching but not another, then one of them prefers the matching in which they are together and one of them prefers the matching in which they are not. Indeed, if they both preferred one of the matchings over the other, then they would be a blocking pair in the matching that is worse for them.

An analogous result is true for the worst among stable matchings for the accepting side.

**Theorem 2.** *In deferred acceptance, an agent on the accepting side obtains an outcome that is the same or worse than this agent’s outcome in any other stable matching.*

In other words, in no stable matching a woman is married to a man who is worse for her than her outcome in the men-proposing deferred acceptance. In no stable matching a man is married to a woman who is worse for him than his outcome in the women-proposing deferred acceptance.
2 How to Check Whether the Stable Matching is Unique?

As an important corollary of the above theorems is

**Corollary 1.** *If the men-proposing and the women-proposing deferred acceptance leads to the same stable matching for a given profile of agents’ preferences, then this stable matching is the unique stable matching under this profile of preferences.*

Let us look at an optional proof of this corollary. Assume that the Men-Proposing and the Women-Proposing DA give the same stable matching, and consider another matching. Let \( m \) be a man who has a different outcome under DA than in this other matching. If \( m \) is single under DA then by Theorem 1 he prefers being single to any other stable outcome, and hence the other matching cannot be stable. Consider thus the case \( m \) is matched to a woman \( w \) under the DA stable matching, and has another outcome in the other matching. If the other matching is stable, then Theorem 1 implies that man \( m \) would rather be with woman \( w \). Furthermore, a symmetric version of Theorem 1 for the Women-Proposing DA tells us that woman \( w \) would rather be with \( m \) than whatever befallen her in the other matching. Thus \( m \) and \( w \) would form a blocking pair, contradicting the stability of the other matching.

3 Where the Preferences Come From: Agents’ Incentives in Reporting Their Preferences

Do participants have incentives to report their true preferences to DA? Let us start with some definitions

A **mechanism** is a function that maps agents reports to outcomes. In this class we study matching mechanisms that map reported preference profiles
(reports) to matchings (outcomes). For example, the deferred acceptance algorithm determines a matching mechanism.

We say that a given report is a dominant strategy for a particular agent (given his or her preferences) if it is optimal report, regardless of the reports of the other players. A mechanism is strategyproof if truth-telling is a dominant strategy for all agents.

Most of you have met the concept of dominant strategies in Econ 101 in the context of 2x2 normal-form games. Reporting strategies to a mechanism such as deferred acceptance is a game, even if somewhat more complex than 2x2 normal form games.

3.1 Deferred acceptance is not strategy-proof

Example 1. [Deferred Acceptance is not strategy proof] Consider 2 men and 2 women with the following preferences:

<table>
<thead>
<tr>
<th></th>
<th>m₁</th>
<th>m₂</th>
<th>w₁</th>
<th>w₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>w₁</td>
<td>m₁</td>
<td>m₂</td>
<td>w₂</td>
<td>w₁</td>
</tr>
<tr>
<td>w₂</td>
<td>w₁</td>
<td>m₁</td>
<td>m₂</td>
<td>m₁</td>
</tr>
</tbody>
</table>

Under man-proposing DA algorithm, if everyone reports truthfully, we get (m₁, w₁), (m₂, w₂).

If w₁ reports that m₁ is unacceptable, the outcome is instead (m₁, w₂), (m₂, w₁) – better for w under her true preferences!

Are there strategy-proof matching mechanisms? Yes, for instance, we could ignore agents’ preferences and always match them in the same way. Such a mechanism does not necessarily generate stable outcomes. As an optional exercise you may want to think whether there are mechanisms which are both strategy-proof and produce stable outcomes.
3.2 Deferred acceptance is strategy-proof for the proposing side

While DA can be manipulated by the accepting side, Al Roth showed that DA induces the proposing side of the market to report truthfully (see Roth “The economics of matching: Stability and incentives,” *Mathematics of Operations Research* 7:617-628, 1982).

**Theorem 3.** In DA, it is a dominant strategy for the proposing side to report their true preferences.

The optional proof relies on the following

**Lemma 1. (Rural Hospitals)** Given a profile of preferences, an agent is unmatched in a stable matching if and only if this agent is unmatched in all stable matchings.

While the proof of the theorem and of the lemma are optional; the lemma itself might be useful (though not necessary) in solving some of the exam problems.

This lemma has an interesting back-story. Back in the 1950s and 1960s rural hospitals had problems filling their residency slots. By proving this lemma, Al Roth showed that this problem cannot be remedied given that the NRMP wants to stick to stable matchings.

Proof of the theorem (the proof is optional). To prove the theorem, suppose that if man $m$ submits some preference ranking $\succ$ then Men Proposing Deferred Acceptance matches $m$ with some woman $w$. Let us call the outcome of the deferred acceptance $M$.

Suppose now that $m$ reports that only $w$ and women $m$ truly prefers over $w$ are acceptable, and if he ranks these women truthfully. Then $m$ must be matched with $w$ or one of the women he prefers. Indeed, the only alternative is that when $m$ submits this new profile then DA leaves him unmatched. The outcome of DA would remain stable if $m$ reported that $w$ is the only acceptable woman. But, with this latter report, matching $M$ above is also
stable. And, by the rural hospital lemma, these two matchings cannot both be stable for the same set of reports.

Finally, notice that the outcome of DA does not depend on how \( m \) ranks women below the woman DA is matching him with. Thus, by reporting his true preferences, \( m \) is matched at least with \( w \). This ends the proof.

### 3.3 What strategies should the agents on the accepting side follow?

What strategies should the agents on the accepting side follow?

**Exercise 1.** [Optimal Strategies for the Accepting Side] Consider 4 men and 4 women with the following preferences:

<table>
<thead>
<tr>
<th></th>
<th>( m_1 )</th>
<th>( m_2 )</th>
<th>( m_3 )</th>
<th>( m_4 )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>( w_3 )</th>
<th>( w_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_1 )</td>
<td>( w_1 )</td>
<td>( w_3 )</td>
<td>( w_3 )</td>
<td>( m_3 )</td>
<td>( m_2 )</td>
<td>( m_1 )</td>
<td>( m_1 )</td>
<td></td>
</tr>
<tr>
<td>( w_2 )</td>
<td>( w_3 )</td>
<td>( w_4 )</td>
<td>( w_1 )</td>
<td>( m_4 )</td>
<td>( m_3 )</td>
<td>( m_2 )</td>
<td>( m_2 )</td>
<td></td>
</tr>
<tr>
<td>( w_3 )</td>
<td>( w_2 )</td>
<td>( w_4 )</td>
<td>( m_1 )</td>
<td>( m_4 )</td>
<td>( m_3 )</td>
<td>( m_4 )</td>
<td>( m_4 )</td>
<td></td>
</tr>
<tr>
<td>( w_4 )</td>
<td>( w_4 )</td>
<td>( w_1 )</td>
<td>( w_2 )</td>
<td>( m_2 )</td>
<td>( m_1 )</td>
<td>( m_4 )</td>
<td>( m_3 )</td>
<td></td>
</tr>
</tbody>
</table>

Under man-proposing DA algorithm, if everyone reports truthfully, we get \( \{(m_1, w_2), (m_2, w_3), (m_3, w_4), (m_4, w_1)\} \). What is the optimal manipulation of woman \( w_1 \) when others submit the true preferences? What about other women?

### 4 Pareto Efficiency

We say that a matching is **Pareto efficient** if no other matching is at least as good for all agents and strictly better for at least one agent.
Exercise 2. [Pareto Efficiency] Consider 2 men and 2 women with the following preferences:

<table>
<thead>
<tr>
<th></th>
<th>$m_1$</th>
<th>$m_2$</th>
<th>$w_1$</th>
<th>$w_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>$w_2$</td>
<td>$m_2$</td>
<td>$m_1$</td>
<td></td>
</tr>
<tr>
<td>$w_2$</td>
<td>$w_1$</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td></td>
</tr>
</tbody>
</table>

We have seen in earlier discussion that there are two stable matchings in this market. Are these stable matchings Pareto efficient? Is the matching $\{(m_1, w_1), m_2, w_2\}$ Pareto efficient?

As an aside, let me note that efficiency and incentives are among the key organizing ideas of market design. We will continue to study them in subsequent parts of this course.