21 Stable Matchings

Assume that there are two disjoint sets of agents, each having preferences over agents in the other set, and the goal is to find an assignment between agents in one set and agents in the other. Real-world examples for a setting like this include the assignment of students to schools or of medical residents to hospitals. We restrict our attention to the special case in which each agent is assigned to at most one agent in the other set. This case has originally been described in terms of marriages and is therefore sometimes referred to as the stable marriage problem.

We describe the problem in terms of a set S of *students* and a set A of potential *advisors*. Each student $i \in S$ has a strict linear order $\succ_i \subseteq A \times A$, each advisor $j \in A$ a strict linear order $\succ_j \subseteq S \times S$. A *matching* is a function $\mu : (S \cup A) \to (S \cup A)$ such that $\mu(\mu(i)) = i$ for all $i \in S \cup A$, $\mu(i) \in A$ for all $i \in S$, and $\mu(j) \in S$ for all $j \in A$. We henceforth assume that $|S| = |A|$, but note that the results can be extended to the case where some agents remain unmatched or prefer remaining unmatched to being matched to certain other agents.

A pair $(i, j) \in S \times A$ is a *blocking pair* for matching μ if i and j would rather be matched to each other than to their respective partners in μ, i.e., if $j \succ_i \mu(i)$ and $i \succ_j \mu(j)$. A matching is called *stable* if it does not have any blocking pairs. Stability is desirable in practice because matchings that are not stable tend to unravel: if two agents find out that they form a blocking pair, they have an incentive to leave the matching; this can introduce additional blocking pairs and eventually cause the whole matching to break down.

It is not obvious, but nevertheless true, that there always exists a stable matching. This can be shown by the so-called *deferred acceptance* procedure. There are two symmetric variants of this procedure, one in which students propose and one in which advisors propose. We concentrate on the former, but note that all results hold symmetrically for the latter. Deferred acceptance proceeds as follows:

1. Let each student $i \in S$ propose to the advisor it ranks highest.
2. Match advisor $j \in A$ tentatively to the highest-ranked among the students that have proposed to it, if any, and let it reject the others for good.
3. Let each student that has just been rejected but has not been rejected by all advisors propose to its next preferred advisor. If there are no such students, then stop and return the tentatively matched pairs. Otherwise go to Step 2.

THEOREM 21.1 (Gale and Shapley, 1962). *There always exists a stable matching.*

Proof. The deferred acceptance procedure clearly terminates after a finite number of rounds, and yields a matching when it does. Let μ be this matching, and assume for contradiction that it is not stable. Then there must exist a blocking pair $(i, j) \in S \times A$,

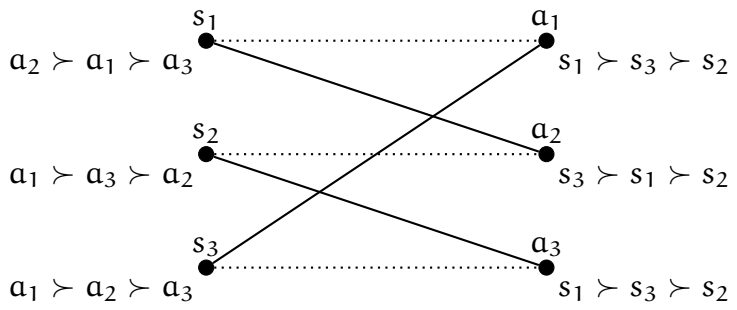

Figure 21.1: Matching problem and two matchings. In the first round of students-propose deferred acceptance, s_1 proposes to a_2 and both s_2 and s_3 propose to a_1, who in turn rejects s_2. Since s_2 was rejected it proposes to a_3 in the second round, and the procedure termines in the stable matching indicated by solid lines. The matching indicated by dotted lines is not stable, because (s_1, a_2) is a blocking pair.

and in particular $j \succ_i \mu(i)$. In the deferred acceptance procedure i must therefore have proposed to j before proposing to $\mu(i)$, and since the two were not matched, j must have rejected i. This means that j must have received a proposal from a student it ranks higher than i but not higher than the student $\mu(j)$ it was eventually matched with. Thus $\mu(j) \succ_j i$, contradicting the fact that (i, j) is a blocking pair. □

21.1 Optimality

There generally exist multiple stable matchings, so one might ask how they compare to each other in quality. Call $j \in A$ *achievable* for $i \in S$ if there exists a stable matching μ such that $\mu(i) = j$. A stable matching μ is then called *student-optimal* if for all $i \in S$, $\mu(i)$ is most preferred among the advisors achievable for i. It is not clear that there should be a stable matching that is student-optimal, i.e., simultaneously gives every student the most preferred advisor it is assigned to in any stable matching. Intriguingly, students-propose deferred acceptance yields such a matching.

THEOREM 21.2. *Students-propose deferred acceptance yields a student-optimal stable matching.*

Proof. Assume for contradiction that the statement of the theorem is false. Then there has to exist a student $i \in S$ and an advisor $j \in A$ such that j is achievable for i but rejects i in students-propose deferred acceptance. Let k be the round in which this happens, and assume without loss of generality that no student is rejected by an achievable advisor in any earlier round. Let $i' \in S$ be the student tentatively matched to j at the end of round k, so that $i' \succ_j i$. Since j is achievable for i, there must exist a stable matching μ with $\mu(i) = j$. Observe that also $\mu(j) = i$, and thus $i' \succ_j \mu(j)$.

Now let $j' = \mu(i')$, which means in particular that j' is achievable for i'. We claim that $j \succ_{i'} j'$. If this was not the case, i.e., if instead $j' \succ_{i'} j$, then i' would have had

to propose to, and be rejected by, j' before being tentatively matched to j. This is impossible because the latter happened in round k, and we assumed that no student was rejected by an achievable advisor before round k. Therefore, $j \succ_{i'} \mu(i')$.

We conclude that (i', j) is a blocking pair for μ, contradicting its stability. □

Curiously, any student-optimal stable matching μ is necessarily *advisor-pessimal*, meaning that for all $j \in A$, $\mu(j)$ is least preferred among the students achievable for j.

THEOREM 21.3. *Every student-optimal stable matching is advisor-pessimal.*

Proof. Consider a student-optimal stable matching μ and a stable matching ν, and assume that $\mu(j) \succ_j \nu(j)$ for some $j \in A$. Let $i = \mu(j)$ and observe that $\nu(j) \neq i$. Since μ is student-optimal, $\mu(i) \succ_i \nu(i)$. Thus (i, j) is a blocking pair for ν, a contradiction. □

21.2 The Stable Matching Polytope

The deferred acceptance procedure necessarily leads to a stable matching that is as good for one side and bad for the other. One might wonder whether it is possible to find a stable matching that is fair for both sides, e.g., one that minimizes $\sum_{i \in S} \sum_{j \in A} r(i, j) + r(j, i)$ or $\max_{(i,j) \in (S \times A) \cup (A \times S)} r(i, j)$, where $r(i, j)$ denotes the rank of agent j in the preference order of agent i. It turns out that this can be done for any linear objective, using a characterization of stable matchings as the extreme points of a convex polytope. Fix a stable matching μ and let $x_{ij} = 1$ for $i \in S$ and $j \in A$ if $\mu(i) = j$, and $x_{ij} = 0$ otherwise. Then,

$$\begin{aligned}
\sum_{j \in A} x_{ij} &= 1 \quad \text{for all } i \in S, \\
\sum_{i \in S} x_{ij} &= 1 \quad \text{for all } j \in A, \\
x_{ij} + \sum_{k : j \succ_i k} x_{ik} + \sum_{k : i \succ_j k} x_{kj} &\leqslant 1 \quad \text{for all } i \in S,\ j \in A, \\
x_{ij} &\geqslant 0 \quad \text{for all } i \in S,\ j \in A.
\end{aligned}$$

The first two constraints are satisfied since every agent is matched with exactly one agent of the respective other type. If the third constraint was not satisfied, then $\sum_{k : j \succ_i k} x_{ik} = 1$ and $\sum_{k : i \succ_j k} x_{kj} = 1$, which would mean that (i, j) is a blocking pair. The last constraint obviously holds as well.

Let P be the polytope described by the above constraints, and observe that by Theorem 21.1, $P \neq \emptyset$. We moreover have the following.

THEOREM 21.4. *A vector is an extreme point of* P *if and only if it is a stable matching.*