Mechanism Design and Social Choice
Part I: Matching

Cédric Wasser

Wintersemester 2017/18

1 Introduction

1.1 Overview

- Recently, there have been numerous applications of **matching theory**, e.g.,
  - entry level labor markets
  - school choice programs
  - kidney exchange
- Interplay between theory and practical design of matching mechanisms
- Belongs to the field of research called *market design*. The idea is to design markets (with or without money) using tools and insights from
  - economic theory, in particular game theory and mechanism design
  - computer science
  - operational research and optimization
  - behavioral sciences
- Nobel Memorial Prize 2012: Alvin E. Roth and Lloyd S. Shapley
  "for the theory of stable allocations and the practice of market design"

Matching theory

Foundations for the theoretical framework were laid in:

- Gale, D., and Shapley, L.S. (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1), 9–15.

  (highly recommended reading, on eCampus)

In Part I, we will look at

1. Two-sided matching
   - The marriage model
   - The college admissions model
   - Applications
2. One-sided matching

We will throughout focus on matching without transfers/money.

Literature

Textbook on two-sided matching:

- Roth, A. E., and Sotomayor, M. A. O. (1990). *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge University Press. $\rightarrow$ Chapters 2, 4, and 5 on eCampus

Background reading (written for general audience):

- Roth, A. E. (2015). *Who Gets What – and Why: The New Economics of Matchmaking and Market Design*. Houghton Mifflin Harcourt.

2 Two-sided matching: The marriage model

2.1–2.3 → Roth and Sotomayor (1990): Sections 2.1, 2.2, and 2.3

2.4 → Roth and Sotomayor (1990): Sections 4.1, 4.2, and 4.3

2.1 The marriage model

- Two finite and disjoint sets of agents:

  men $M = \{m_1, m_2, \ldots, m_n\}$ **women** $W = \{w_1, w_2, \ldots, w_p\}$

- **One-to-one** matching:
  each agent can either marry one agent of opposite sex or stay single
- No monetary transfers (no dowries)

Preferences: rational and **strict**

- each man $m \in M$ has an ordered list of preferences $P(m)$ on $W \cup \{m\}$
- each woman $w \in W$ has an ordered list of preference $P(w)$ on $M \cup \{w\}$
- Example: $P(m_1) = w_2, w_1, m_1, w_3, \ldots, w_p$
  - man m_1 strictly prefers woman w_2 over w_1, i.e., $w_2 \succ_{m_1} w_1$
  - he strictly prefers woman w_1 over staying single, i.e., $w_1 \succ_{m_1} m_1$
  - he strictly prefers staying single over all other potential partners

Matching

Definition

A **matching** is a function $\mu : M \cup W \to M \cup W$
such that for all $w \in W$ and $m \in M$

(i) $\mu(m) \in W \cup \{m\}$,

(ii) $\mu(w) \in M \cup \{w\}$,

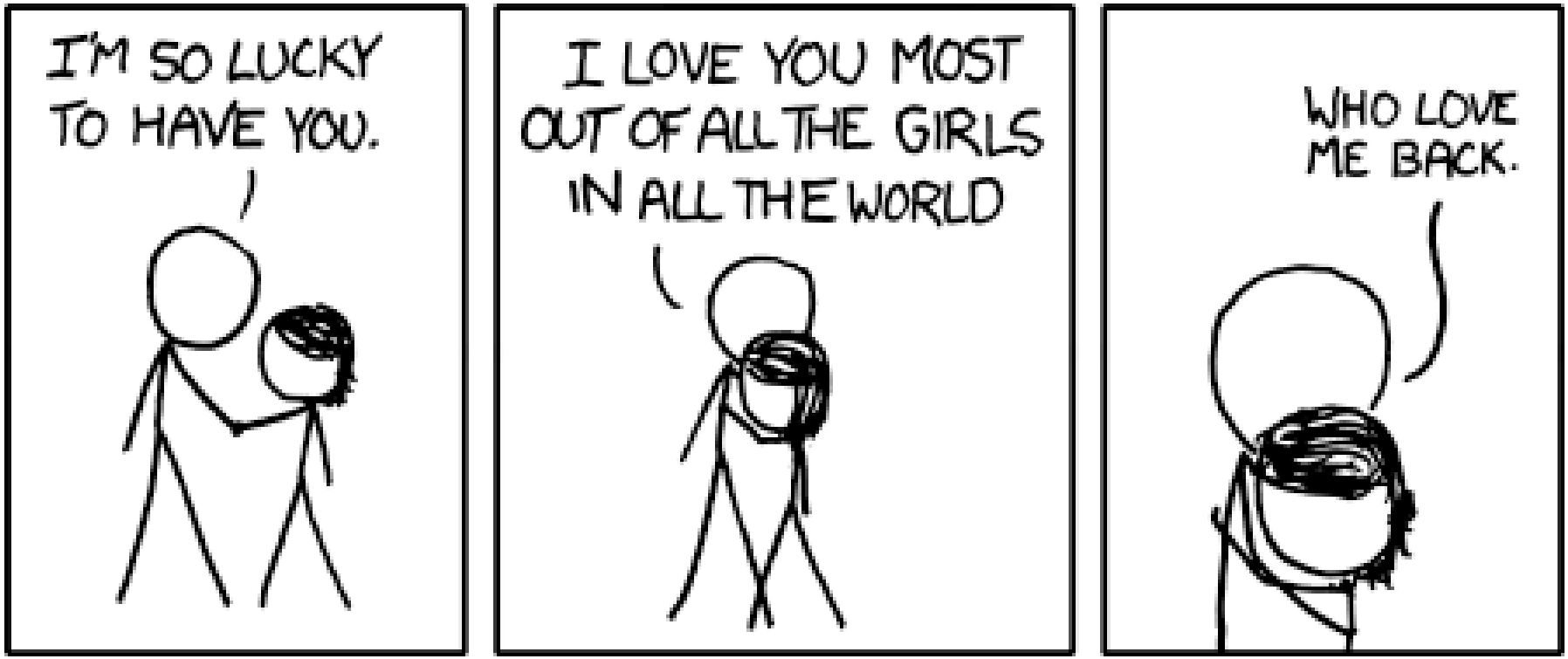
(iii) $\mu(m) = w$ if and only if $\mu(w) = m$.

Under matching μ,

- if $\mu(i) = i$, agent i stays unmatched (single),
- otherwise, agent i is matched (married) to agent $\mu(i)$.

Which matchings are likely to occur? $\to$ concept of *stable* matchings

2.2 Stability



Source: `http://www.xkcd.com/770/`

Example

3 men and 3 women with preferences

$$P(m_1) = w_2, w_1, w_3, m_1 \qquad P(w_1) = m_1, m_3, m_2, w_1$$
$$P(m_2) = w_1, w_3, w_2, m_2 \qquad P(w_2) = m_3, m_1, m_2, w_2$$
$$P(m_3) = w_1, w_2, m_3, w_3 \qquad P(w_3) = m_1, m_3, m_2, w_3$$

- What if $\mu(m_3) = w_3$?
  Not stable, since $m_3 \succ_{m_3} w_3$.
- What if $\mu(w_3) = m_1$ and $\mu(w_2) = m_2$?
  Not stable, since m_1 and w_2 will both divorce and marry each other.

Stable matchings

Definition

A matching μ is **stable**, if the following conditions are satisfied:

(i) For each $i \in M \cup W$, $\mu(i) \succeq_i i$. **(individual rationality)**

(ii) There is no pair $(m, w) \in M \times W$ such that $m \succ_w \mu(w)$ and $w \succ_m \mu(m)$. **(no blocking pairs)**

A stable matching is Pareto-efficient.

Example of a stable matching for preferences on preceding slide:

$$\mu = \begin{matrix} w_2 & w_3 & w_1 \\ m_1 & m_2 & m_3 \end{matrix}$$

Existence of stable matchings

Theorem (Gale and Shapley, 1962)

A stable matching always exists in the marriage model.

The proof is constructive, using an algorithm that always produces a stable matching:

The **deferred acceptance algorithm** (DA).

Two versions: men-proposing and women-proposing

The men-proposing deferred acceptance algorithm

Definition: j is **acceptable** to agent i if $j \succ_i i$.

Round 1

a. Each man proposes to his favorite woman.

b. Each woman rejects all proposals except the one from her most preferred acceptable man.
Any man who is not rejected is temporarily accepted.

Round k

a. Each man who was rejected in the previous round proposes to his next choice as long as there remains an acceptable woman to whom he has not yet proposed.

b. Each woman rejects all but the most preferred acceptable man out of the new proposals and the temporarily accepted man from the previous round.

The algorithm stops when there are no new proposals.
Each temporary acceptance becomes final.

The women-proposing deferred acceptance algorithm

Definition: j is **acceptable** to agent i if $j \succ_i i$.

Round 1

a Each woman proposes to her favorite man.

b Each man rejects all proposals except the one from his most preferred acceptable woman.
Any woman who is not rejected is temporarily accepted.

Round k

a Each woman who was rejected in the previous round proposes to her next choice as long as there remains an acceptable man to whom she has not yet proposed.

b Each man rejects all but the most preferred acceptable woman out of the new proposals and the temporarily accepted woman from the previous round.

The algorithm stops when there are no new proposals.
Each temporary acceptance becomes final.

Deferred acceptance algorithm: Example

$$P(m_1) = w_2, w_1, w_3, m_1 \qquad P(w_1) = m_1, m_3, m_2, w_1$$
$$P(m_2) = w_1, w_3, w_2, m_2 \qquad P(w_2) = m_3, m_1, m_2, w_2$$
$$P(m_3) = w_1, w_2, m_3, w_3 \qquad P(w_3) = m_1, m_3, m_2, w_3$$

Outcome of men-proposing DA:

$$\mu^M = \begin{matrix} w_1 & w_2 & w_3 \\ m_3 & m_1 & m_2 \end{matrix}$$

Outcome of women-proposing DA:

$$\mu^W = \begin{matrix} w_1 & w_2 & w_3 \\ m_1 & m_3 & m_2 \end{matrix}$$

Properties of the DA

DA is well-defined

- Algorithm eventually stops as there is a finite number of women (men) and no man (woman) proposes more than once to any woman (man).
- Feasibility: In every round, each man (woman) proposes to at most one woman (man) and each woman (man) keeps at most one man (woman).

Outcome of DA is stable

- *Individual rationality*: No man or woman proposes to or accepts an unacceptable partner.
- *No blocking pairs*: Suppose man m and woman w are not matched under the outcome μ of the men-proposing DA but $w \succ_m \mu(m)$. Then m must have proposed to w but was subsequently rejected in favor of someone w liked better. Hence, $\mu(w) \succ_w m$ and (m, w) is not a blocking pair.

$\Longrightarrow$ Proof of existence of a stable matching: DA always produces one.

Men-optimal and women-optimal stable matchings

- A matching μ is a **men-optimal** stable matching if μ is stable and for each other stable matching μ' and each $m \in M$ we have $\mu(m) \succeq_m \mu'(m)$.
- A matching μ is a **women-optimal** stable matching if μ is stable and for each other stable matching μ' and each $w \in W$ we have $\mu(w) \succeq_w \mu'(w)$.

Theorem

The men-proposing DA produces the unique men-optimal stable matching.
The women-proposing DA produces the unique women-optimal stable matching.

Proof.

- Definition: w and m are **achievable** for each other if there exists a stable matching μ where $\mu(m) = w$.
- Consider the women-proposing DA.
  We will show: no woman is ever rejected by an achievable man.
  $\Longrightarrow$ every woman is matched to her most preferred achievable man
  $\Longrightarrow$ women-proposing DA yields unique women-optimal stable matching
- Proof by induction: Suppose we are in round k of the algorithm and in all preceding rounds no woman was rejected by an achievable man.
  - Suppose man m rejects woman w.
  - If $m \succ_m w$, then m is clearly not achievable.
  - Otherwise, some woman w' must have proposed to m in round k, where $w' \succ_m w$. We prove by contradiction that m is not achievable for w.
  - We know w' prefers m over all other men except for the unachievable ones that rejected her in previous rounds. Suppose there is a stable matching μ that matches w to m and everybody else to an achievable partner. But then $m \succ_{w'} \mu(w')$ and $w' \succ_m w$. $\Rightarrow \mu$ is unstable. $\Rightarrow m$ not achievable for w

□

2.3 The set of stable matchings

- Stability is related to a solution concept from *cooperative game theory*:

  The core requires that no coalition of players can break away and take a joint action that makes them better of.

Definition

In the marriage model, **the core** is the set of matchings μ for which there is no coalition $S \subseteq M \cup W$ and matching μ' such that

(i) $\mu'(i) \in S$ for each $i \in S$,

(ii) $\mu'(i) \succeq_i \mu(i)$ for each $i \in S$ and $\mu'(j) \succ_j \mu(j)$ for some $j \in S$.

Stable matchings and the core

Theorem

In the marriage model, the core is equal to the set of stable matchings.

Proof.

- In the core $\Rightarrow$ stable:
  If μ is in the core, no coalition consisting of one agent or of two agents can do better, implying individual rationality and no blocking pairs.
- Not in the core $\Rightarrow$ not stable:
  If μ is not in the core, there are S and μ' satisfying (i) and (ii). For at least one $i \in S$, $\mu'(i) \succ_i \mu(i)$. If $\mu'(i) = i$, μ is not individually rational for i. If $\mu'(i) = j \neq i$, we also have $i \succ_j \mu(j)$, i.e., (i, j) is a blocking pair. $\square$

Opposition of interests

Define **common preferences of men** by

- $\mu \succeq_M \mu'$ if and only if $\mu(m) \succeq_m \mu'(m)$ for all $m \in M$,
- $\mu \succ_M \mu'$ if and only if $\mu \succeq_M \mu'$ and $\mu \neq \mu'$.

Define **common preferences of women** $\succeq_W$, $\succ_W$ analogously.

Theorem (Opposition of interests)

If μ and μ' are two stable matchings, then $\mu \succ_M \mu'$ if and only if $\mu' \succ_W \mu$.

$\Longrightarrow$ The women-optimal (men-optimal) stable matching is the worst stable matching for men (women).

Proof.

Let μ, μ' be stable and $\mu \succ_M \mu'$. We will show that $\mu' \succ_W \mu$ by contradiction.

- Suppose not. Then, there is a woman w such that $\mu(w) \succ_w \mu'(w)$.
- As man $m = \mu(w)$ also has different partners under μ and μ', we must have $w = \mu(m) \succ_m \mu'(m)$.
- But then μ' is not stable, because m and w are not matched under μ' but prefer each other over their partners.

□

The lonely wolf theorem

Theorem

The set of singles is the same for any stable matching.

Proof.

- Let μ^M be the men-optimal stable matching. Define for all matchings μ,
  - $\mathcal{M}(\mu)$: the set of married men
  - $\mathcal{W}(\mu)$: the set of married women
- As matchings are one-to-one, $|\mathcal{M}(\mu)| = |\mathcal{W}(\mu)|$ for all μ.
- As $\mu^M \succ_M \mu$ and $\mu^M \prec_W \mu$ for all stable matchings $\mu \neq \mu^M$,
  - $\mathcal{M}(\mu) \subseteq \mathcal{M}(\mu^M)$ and hence $|\mathcal{M}(\mu)| \leq |\mathcal{M}(\mu^M)|$,
  - $\mathcal{W}(\mu) \supseteq \mathcal{W}(\mu^M)$ and hence $|\mathcal{W}(\mu)| \geq |\mathcal{W}(\mu^M)|$.
- $\implies |\mathcal{M}(\mu)| = |\mathcal{M}(\mu^M)| = |\mathcal{W}(\mu)| = |\mathcal{W}(\mu^M)|$
- $\implies \mathcal{M}(\mu) = \mathcal{M}(\mu^M)$ and $\mathcal{W}(\mu) = \mathcal{W}(\mu^M)$ for all stable μ □

2.4 Incentives

- So far: assumed that agents' preferences commonly known
- But preferences are typically *private information*
  $\rightarrow$ need to be elicited during matching process

- Do agents have an incentive to behave *straightforwardly* when facing a market organized, e.g., via the deferred acceptance algorithm?

  Or can strategically misrepresenting one's true preferences be profitable?

- will use game theory to study equilibrium outcomes,
  focusing on *dominant strategies*

Marriage market as a preference revelation game

- Marriage market: men M and women W with true *profile of preferences*

$$\mathbf{P} = \{P(m_1), \ldots, P(m_n), P(w_1), \ldots, P(w_p)\}$$

- Let $\mathcal{P}_i$ denote the set of all possible preferences for agent i.
- Let $\mathcal{P} := \mathcal{P}_{m_1} \times \cdots \times \mathcal{P}_{w_p}$ denote the set of possible preference profiles.
- Let Ψ denote the set of all possible matchings μ.

- We will focus on *direct mechanisms*:
  Agents report preferences to a matchmaker/clearinghouse that chooses a matching based on the reports according to commonly known rules.
- *revelation principle*: restriction on direct mechanisms is without loss

($\rightarrow$ *formal treatment of these concepts in* **Part II**)

Matching mechanisms

- Each agent i reports preferences $Q(i) \in \mathcal{P}_i$.
- Reported preference profile $\mathbf{Q} := \{Q(m_1), \ldots, Q(w_p)\}$

Definition

A direct **matching mechanism** is a function $h \colon \mathcal{P} \to \Psi$ that chooses a matching $\mu = h(\mathbf{Q})$ for every reported profile of preferences $\mathbf{Q}$.

A matching mechanism h is **stable** if, for each $\mathbf{Q} \in \mathcal{P}$, $h(\mathbf{Q})$ is a stable matching under preferences $\mathbf{Q}$.

- Example: matchmaker runs men-proposing DA using reported preferences.

Strategy-proof matching mechanisms

Each agent i evaluates outcomes according to true preferences:

$\mu = h(\mathbf{Q}) \succeq_i h(\mathbf{Q}') = \mu'$ if and only if $\mu(i) \succeq_i \mu'(i)$ according to $P(i)$.

Reporting $Q^*(i)$ is a **dominant strategy** for agent i if

$$h(Q^*(i), Q_{-i}) \succeq_i h(Q(i), Q_{-i}) \quad \text{for all } Q(i) \in \mathcal{P}_i \text{ and } Q_{-i} \in \mathcal{P}_{-i}.$$

Definition

A matching mechanism h is **strategy-proof** if reporting the true preferences $P(i)$ is a dominant strategy for each agent $i \in M \cup W$.

Example

2 men and 2 women with true preferences $\mathbf{P}$:

$$P(m_1) = w_1, w_2, m_1 \qquad P(w_1) = m_2, m_1, w_1$$
$$P(m_2) = w_2, w_1, m_2 \qquad P(w_2) = m_1, m_2, w_2$$

Two stable matchings: $\mu_a = \begin{matrix} w_1 & w_2 \\ m_1 & m_2 \end{matrix}$ $\mu_b = \begin{matrix} w_1 & w_2 \\ m_2 & m_1 \end{matrix}$

For any stable matching mechanism h, $h(\mathbf{P}) \in \{\mu_a, \mu_b\}$.

- Suppose $h(\mathbf{P}) = \mu_a$.
  - w_1 could report $Q(w_1) = m_2, w_1, m_1$ instead of $P(w_1)$.
  - μ_b is the unique stable matching for $\mathbf{P}' = (P(m_1), P(m_2), Q(w_1), P(w_2))$.
  - $h(\mathbf{P}') = \mu_b$ for any stable mechanism h
  - w_1 is strictly better off since $\mu_b(w_1) \succ_{w_1} \mu_a(w_1)$.
- Suppose $h(\mathbf{P}) = \mu_b$. Similarly, m_1 can profitably misstate his preferences.

$\Longrightarrow$ In all stable matching mechanisms, some agent profits from lying.

Impossibility result

- The example shows that for $|M| = |W| = 2$ there is no stable matching mechanism that is strategy-proof.
- Example extends to larger marriage markets: Let preferences of m_1, m_2, w_1, and w_2 be as before (all other agents are unacceptable). In all stable matching mechanisms, some $i \in \{m_1, m_2, w_1, w_2\}$ profits from lying.

Theorem (Roth, 1982)

There is no stable matching mechanism that is strategy-proof.

Strategy-proof for one side

- The **men-optimal (women-optimal)** stable matching mechanism is the mechanism that yields the men-optimal (women-optimal) stable matching for the reported preferences.

Theorem (Dubins and Freedman, 1981; Roth, 1982)

In the men-optimal stable matching mechanism,
reporting the true preferences is a dominant strategy for all men.

In the women-optimal stable matching mechanism,
reporting the true preferences is a dominant strategy for all women.

- The DA is strategy-proof for the proposing side.

Proof.

We only prove the result for the men-optimal stable mechanism h^M.

By way of *contradiction*, suppose there is a man m with true preferences $P(m)$ who profits from report $\overline{P}(m)$: $\overline{\mu} = h^M(\overline{P}(m), Q_{-m}) \succ_m h^M(P(m), Q_{-m})$.

Stability implies $\overline{\mu}(m) \succ_m m$.

Successive weak improvements for m, culminating in truthful reporting:

1. If m reports $Q^1(m) = \overline{\mu}(m), m, \ldots$ instead of $\overline{P}(m)$, $\overline{\mu}$ is still stable (no new blocking pairs possible).
   Hence, $\mu^1(m) = \overline{\mu}(m)$ under matching $\mu^1 = h^M(Q^1(m), Q_{-m})$.
   Lonely wolf theorem $\Rightarrow$ For each matching under $(Q^1(m), Q_{-m})$ where m is single, there is a blocking pair $(\tilde{m}, \tilde{w})$.
2. Let m report $Q^2(m)$ such that it is equal to $P(m)$ up to woman $\overline{\mu}(m)$ while all remaining women are claimed unacceptable. Matchings where m is single are still blocked by $(\tilde{m}, \tilde{w})$. $\implies$ $h^M(Q^2(m), Q_{-m}) \succeq_m \overline{\mu}$.
3. But then matching $h^M(Q^2(m), Q_{-m})$ is also stable under $(P(m), Q_{-m})$ (no new blocking pairs possible). $\implies$ $h^M(P(m), Q_{-m}) \succeq_m \overline{\mu}$. □

3 Two-sided matching: The college admissions model

$\rightarrow$ Roth and Sotomayor (1990): Sections 5.1 to 5.3

3.1 The college admissions model

In many two-sided markets of interest (workers and firms, school choice,...) matching is not one-to-one. $\rightarrow$ extension of marriage model:

The college admissions model

- Two finite and disjoint sets of agents:

  students $S = \{s_1, s_2, \ldots, s_n\}$ **colleges** $C = \{c_1, c_2, \ldots, c_p\}$

- **Many-to-one** matching: Each student may attend only one college but several students may attend the same college.
- Each college c has a **quota** $q_c \in \mathbb{N}$, i.e., it can admit up to q_c students.

As in marriage model, **strict** preferences of

- each student $s \in S$ over colleges: ordered list $P(s)$ on $C \cup \{s\}$
- each college $c \in C$ over *individual students*: ordered list $P(c)$ on $S \cup \{c\}$

Matchings

Definition

A **matching** is a function $\mu : S \cup C \to S \cup C \cup 2^S$ such that for all $s \in S$, $c \in C$

(i) $\mu(s) \in C \cup \{s\}$,

(ii) $\mu(c) \subseteq S$ and $|\mu(c)| \leq q_c$,

(iii) $\mu(s) = c$ if and only if $s \in \mu(c)$.

Under matching μ,

- college c admits set of students $\mu(c)$
- student s is admitted by college $\mu(s)$
- if $\mu(s) = s$, s does not attend any college
- if $\mu(c) = \varnothing$, c does not admit any students

Preferences over matchings

- Students only care about which college they are admitted to:

  As in marriage model, preferences over matchings correspond to preferences over colleges $P(s)$.

- Each colleges c cares about which **group** of students it admits.
  - Preferences over subsets of students could be very complex.
  - We will assume that preferences over groups of students are related in a natural way to preferences over *individual* students $P(c)$.

    $\rightarrow$ **responsiveness**

Responsive preferences

Definition

Preferences over sets of students of a college $c \in C$ are **responsive** to $P(c)$ if for each $T \subseteq S$ and

(i) for all $s, s' \in S \setminus T$,

$$T \cup \{s\} \succ_c T \cup \{s'\} \quad \text{if and only if} \quad s \succ_c s' \text{ under } P(c),$$

(ii) for all $s \in S \setminus T$,

$$T \cup \{s\} \succ_c T \quad \text{if and only if} \quad s \succ_c c \text{ under } P(c),$$

(iii) for all $s \in T$,

$$T \setminus \{s\} \succ_c T \quad \text{if and only if} \quad c \succ_c s \text{ under } P(c).$$

Assumption: Each college $c \in C$ has preferences that are responsive to $P(c)$.

Example: responsive preferences

Suppose college c has preferences that are **responsive** to

$$P(c) = s_1, s_2, s_3, s_4, c, s_5.$$

Some examples of what we can infer from this assumption:

- $\{s_1, s_3\} \succ_c \{s_2, s_4\}$
- $\{s_1\} \succ_c \{s_1, s_5\}$
- $\varnothing \succ_c \{s_5\}$
- $\{s_1, s_4\} \succ_c \varnothing$

Preference between $\{s_1, s_4\}$ and $\{s_2, s_3\}$ is not pinned down by responsiveness to $P(c)$.

Stability

Definition

A matching μ is **stable** if the following conditions are satisfied:

(i) No student is matched to an unacceptable college, i.e., for each $s \in S$, $\mu(s) \succeq_s s$. **(IR for students)**

(ii) No college wants to drop *one* of the students matched to it, i.e., for each $c \in C$ and each $s \in \mu(c)$, $s \succ_c c$. **(IR for colleges)**

(iii) There is no student-college pair (s, c) such that $c \succ_s \mu(s)$, $s \succ_c c$, and either $|\mu(c)| < q_c$ or $s \succ_c s'$ for some $s' \in \mu(c)$. **(no blocking pairs)**

- Because preferences are responsive, only colleges' preferences over individual students are relevant for stability.
- For responsive preferences, one can also show that if matching μ is stable, there is no blocking coalition of colleges and students, i.e., μ is *in the core*.

3.2 A related marriage market

Many results from the marriage model carry over to the present setting (given preferences are responsive).

Key insight: can construct a **related marriage market** as follows.

1. Replace each college $c \in C$ by q_c "women" $c^1, c^2, \ldots, c^{q_c}$ and equip each woman c^k with preferences $P(c^k) = P(c)$.
2. The preferences $\tilde{P}(s)$ of "man" $s \in S$ are derived from the preferences $P(s)$ of student s by replacing each $c \in C$ with the string $c^1, c^2, \ldots, c^{q_c}$.

- For each matching μ in the college admissions model, define a **corresponding matching** $\tilde{\mu}$ in the related marriage market such that

$$\tilde{\mu}(s) = c^k \text{ and } \tilde{\mu}(c^k) = s \quad \Longleftrightarrow \quad \begin{array}{l} s \text{ is the } k\text{th most preferred student} \\ \text{in } \mu(c) \text{ under preferences } P(c). \end{array}$$

Stable matchings

Lemma

A matching μ in the college admissions model is stable if and only if the corresponding matching $\tilde{\mu}$ in the related marriage market is stable.

Theorem

Consider a college admissions problem with responsive preferences.

(i) *A stable matching always exists.*

(ii) *The student-proposing deferred acceptance algorithm (SDA) produces the unique student-optimal stable matching.*

(iii) *The sets of unmatched students and unfilled positions are the same for all stable matchings.*

Incentives

- The marriage model is a college admissions model with $q_c = 1$ for all c.

Theorem

There is no stable matching mechanism that is strategy-proof.

- Strategically, students in the college admissions model are in the same situation as in the related marriage market.

Theorem

In the student-optimal stable matching mechanism, reporting the true preferences is a dominant strategy for all students.

- In contrast, one can show that there is no stable matching mechanism that makes truthful reporting a dominant strategy for colleges (unless $q_c = 1 \forall c$).
- Intuition: A college is like a coalition of agents that can jointly misreport.

3.3 Importance of stability

- In our theoretical matching market, unstable matchings *cannot* persist.
- But is stability relevant for "real-life" markets?

- Fact 1: Stable mechanisms often successful while unstable ones are not
  - U.S. market for new doctors: National Intern Matching Program (Roth, 1984)
  - Regional markets for new physicians and surgeons in the U.K. (Roth, 1991)
  - Experimental evidence (Kagel and Roth, 2000)

- Fact 2: Has proven useful guideline for design of institutions
  - Redesign of the residency matching program (Roth and Peranson, 1999)
  - School choice

4 School Choice

4.1 The school choice problem

- Traditionally students assigned to public school closest to their home.
- Today many school choice districts (in particular in the US) take student preferences into account in assigning students to public schools.
- In most applications, **schools** are **not strategic** agents but rather *objects to be consumed by students* (Abdulkadiroglu and Sönmez, 2003).
- Admission at over-demanded schools regulated by assigning **priorities** to students (based on, e.g., distance to school, entry exams, siblings,...).
- What is a *good* way of matching students to public schools?

Setup

Adapt and reinterpret *college admissions model* (CAM) to study school choice:

- **students** $S = \{s_1, \ldots, s_n\}$ and **schools** $C = \{c_1, \ldots, c_p\}$
- Each school $c \in C$ can admit at most $q_c \in \mathbb{N}$ students.

- Student $s \in S$ has an ordered list of strict preferences $P(s)$ on $C \cup \{s\}$.
- School $c \in C$ has a strict **priority** ranking $R(c)$ of individual students S.
  - corresponds to colleges' preferences in CAM but priority ranking is determined by fixed and publicly known legal criteria
  - each student is acceptable to all schools ($s \succ_c c$ for all $s \in S$ and $c \in C$)
- School c's ranking of groups of students implied by priorities $R(c)$ corresponds to preferences in CAM that are **responsive** to $R(c)$.

- Matching μ is defined as in CAM.

Important: Welfare evaluations are based only on the preferences of students.

Reinterpreting stability

Definition

A matching μ is

- **individually rational** if $\mu(s) \succeq_s s$ for each student $s \in S$,
- **non-wasteful** if $c \succ_s \mu(s)$ implies $|\mu(c)| \geq q_c$,
- **fair** if there is no pair (s, c) such that $c \succ_s \mu(s)$ and $s \succ_c \hat{s}$ for some $\hat{s} \in \mu(c)$.

μ is **stable** if and only if it is individually rational, non-wasteful, and fair.

Definition of **stability** coincides with definition in CAM:

- *IR for students*
- *IR for schools* automatically satisfied
- *no blocking pairs* if and only if non-wasteful and fair

Example

3 students and 3 schools with $q_{c_1} = q_{c_2} = q_{c_3} = 1$.

$$P(s_1) = c_2, c_1, c_3, s_1 \qquad R(c_1) = s_1, s_3, s_2$$
$$P(s_2) = c_1, c_2, c_3, s_2 \qquad R(c_2) = s_2, s_1, s_3$$
$$P(s_3) = c_1, c_2, c_3, s_3 \qquad R(c_3) = s_2, s_1, s_3$$

- Unique stable matching: $\mu = \begin{matrix} s_1 & s_2 & s_3 \\ c_1 & c_2 & c_3 \end{matrix}$.

- The matching $\tilde{\mu} = \begin{matrix} s_1 & s_2 & s_3 \\ c_2 & c_1 & c_3 \end{matrix}$ is not stable

  but Pareto dominates μ, i.e., $\tilde{\mu} \succ_S \mu$.

Stability and Efficiency

- Example shows: there are student preferences $\mathbf{P}$ such that for any fair matching μ there is another matching $\tilde{\mu}$ that all students weakly prefer over μ, with at least one strict preference.

  $\Longrightarrow$ **Fairness** may not be compatible with Pareto **efficiency**.

- Rationale for fairness:
  - justification of rejections
  - priorities as society's preferences regarding access to public schools

- A **fair matching mechanism** h chooses a matching $h(\mathbf{Q})$ that is stable under the reported preferences $\mathbf{Q} = \{Q(s_1), \ldots, Q(s_n)\}$ of students and the known priorities of schools.

The student-optimal stable matching mechanism

Let h^S denote the student-optimal stable matching mechanism.

- h^S operates as in the CAM, e.g., it runs the student-proposing DA using priorities of schools and *reported* preferences of students.

Our results for the CAM imply the following.

Theorem

(i) *h^S is strategy-proof.*

(ii) *h^S Pareto dominates any other fair matching mechanism: if h is a fair matching mechanism, then $h^S(\mathbf{P}) \succeq_S h(\mathbf{P})$ for all $\mathbf{P} \in \mathcal{P}$.*

If fairness is the prime objective, h^S is a compelling mechanism.

4.2 The Boston mechanism

- How are actual school choice programs organized?
- The following algorithm has been a popular choice among policymakers.

Boston school choice algorithm:

Round 1
- a Each student applies to her (reported) top choice.
- b Each school c admits applicants one at a time according to $R(c)$ until either the school's capacity is exhausted or there are no more students who ranked it first.

Round k
- a Each unmatched student applies to her kth choice.
- b Each school c with *remaining capacity* admits applicants one at a time according to $R(c)$ until either the school's capacity is exhausted or there are no more students who ranked it kth.

Example

- 3 student and 3 schools with $q_{c_1} = q_{c_2} = q_{c_3} = 1$
- Preferences:

$$P(s_1) = c_1, c_3, c_2, s_1 \quad P(s_2) = c_1, c_2, c_3, s_2 \quad P(s_3) = c_2, c_1, c_3, s_3$$

- Priorities: $R(c_1) = R(c_2) = R(c_3) = s_1, s_2, s_3$
- Under truthful reporting, the Boston mechanism h^B yields the matching

$$h^B(\mathbf{P}) = \begin{matrix} s_1 & s_2 & s_3 \\ c_1 & c_3 & c_2 \end{matrix}.$$

- But student s_2 would be better off by applying to c_2 first...

Remarks

- The outcome is efficient with respect to the *reported* preferences.
- However, strong incentives to manipulate $\implies$ h^B is not strategy-proof
- Empirical evidence in Abdulkadiroglu et al. (2006) that
  1. students/their parents act upon incentives to manipulate and
  2. manipulation is particularly harmful to parents who strategize suboptimally (a strategy-proof mechanism would *level the playing field*).

- In 2005, the Boston Public Schools decided to replace the existing Boston mechanism by the student-proposing DA mechanism (h^S) in their school choice program (Abdulkadiroglu et al., 2006).
- In New York City a version of the student-proposing DA mechanism has been in use since 2003.

Equilibrium characterization

Ergin and Sönmez (2006) study Nash equilibria of the Boston mechanism h^B assuming the following.

- Students have **complete information** about all true preferences $\mathbf{P}$. (strong assumption!)
- Each students $s \in S$ simultaneously reports $Q(s) \in \mathcal{P}_s$.
- Given reports $\mathbf{Q}$, the Boston mechanism determines the matching $h^B(\mathbf{Q})$.

Definition

$\mathbf{Q}^* = (Q^*(s), Q^*_{-s})$ is a **Nash equilibrium** of h^B under true preferences $\mathbf{P}$ if

$$h^B(Q^*(s), Q^*_{-s}) \succeq_s h^B(Q(s), Q^*_{-s}) \quad \text{for each } Q(s) \in \mathcal{P}_s \text{ and each } s \in S.$$

Equilibrium characterization: Result

Theorem (Ergin and Sönmez, 2006)

A report profile $\mathbf{Q}^*$ is a Nash equilibrium of the Boston mechanism h^B if and only if the matching $h^B(\mathbf{Q}^*)$ is stable with respect to the true preferences $\mathbf{P}$.

Corollary

For all Nash equilibria $\mathbf{Q}^*$ of the Boston mechanism h^B under preferences $\mathbf{P}$,

$$h^S(\mathbf{P}) \succeq_S h^B(\mathbf{Q}^*).$$

- can be seen as a theoretical justification for moving to the student-proposing DA mechanism in Boston in 2005
- **Caveat:** the theorem does not hold when
  - there is uncertainty about students' preferences,
  - some students do not act strategically,
  - priorities are not strict.

University Admission in Germany

Centralized allocation procedure for places in human medicine, dentistry, veterinary medicine, and pharmacy at German public universities, organized by the *Stiftung für Hochschulzulassung* (`http://www.hochschulstart.de`).

Sequential procedure:

- **Stage 1**: *Boston mechanism* is used to allocate
  - up to 20% of the places to the applicants with the best school grades
  - up to 20% of the places to the applicants with the longest waiting times
- **Stage 2**: All remaining places are allocated among remaining applicants using the *university-proposing DA mechanism*.

See Westkamp (2013) for a theoretical analysis and an alternative procedure.

Recently, a centralized procedure ("Dialogorientiertes Serviceverfahren") has also been introduced for locally admission-restricted subjects.
It uses the *university-proposing DA mechanism*.

5 One-sided matching

5.1 The house allocation model

- A finite set of **agents** $N = \{1, \ldots, n\}$
- A finite set of **houses** (objects) $A = \{a_1, \ldots, a_n\}$
- Initial **endowments**: agent i owns house a_i.
- No monetary transfers possible (rents are exogenously given).
- Each agent $i \in N$ has an ordered list of **strict** preferences $P(i)$ on A.
  Profile of preferences: $\mathbf{P} := (P(1), \ldots, P(n))$.

Applications

- subsidized housing: on-campus dormitories, public housing
- work related: offices, parking lots
- human organs
- to some extent also school choice

Matchings

Definition

A **matching** is a function $\mu : N \to A$ such that, for all $i, j \in N$, if $\mu(i) = \mu(j)$, then $i = j$ (i.e., μ is *injective*).

- A matching assigns one house $\mu(i)$ to each agent i.
- No house can be assigned to several agents.
- Preferences over matchings: $\mu \succeq_i \mu'$ if and only if $\mu(i) \succeq_i \mu'(i)$.

Definition

A matching μ is

- Pareto **efficient** if there is no other matching $\hat{\mu}$ such that

$$\hat{\mu}(i) \succeq_i \mu(i) \text{ for each } i \in N \text{ and } \hat{\mu}(j) \succ_j \mu(j) \text{ for some } j \in N.$$

- **individually rational** if $\mu(i) \succeq_i a_i$ for each $i \in N$.

Stability and the core

As in the marriage model, we use **the core** from cooperative game theory to get a notion of *stable* matchings. $\rightarrow$ Stability requires that no coalition of agents can improve by breaking away and trading their houses among themselves.

Definition

In the house allocation model, **the core** is the set of matchings μ for which there is no coalition $S \subseteq N$ and matching μ' such that,

(i) for each $i \in S$ there is some $j \in S$ such that $\mu'(i) = a_j$,

(ii) for each $i \in S$, $\mu'(i) \succeq_i \mu(i)$ and for some $j \in S$, $\mu'(j) \succ_j \mu(j)$.

If a matching is in the core, then it is

- **individually rational** (no one-agent coalition $S = \{i\}$ satisfies (i) & (ii)),
- Pareto **efficient** (the coalition $S = N$ does not satisfy (i) & (ii)).

5.2 Top trading cycles

Theorem (Shapley and Scarf, 1974)

In the house allocation model, there exists a unique core matching.

The theorem can be proved using an algorithm that always produces the unique core matching:

The **top trading cycles algorithm** (TTC), due to David Gale.

The TTC is an iterative process:

- Each agent points to the agent that owns her most preferred house.
- Agents that form a cycle trade their houses accordingly and are removed.
- The procedure is repeated until no agents remain.

The top trading cycles algorithm

Set $N^1 = N$ and $A^1 = A$. Start round $k = 1$ of the algorithm.

Round k:

- Construct a directed graph G^k with nodes N^k. There is a directed edge from agent $i \in N^k$ to agent $j \in N^k$ if a_j is agent i's top house in A^k.
- Allocate houses along every cycle of the graph G^k.
  Formally, if $(i^1, i^2, \ldots, i^m, i^1)$ is a cycle in G^k,
  then set $\mu(i^1) = a_{i^2}$, $\mu(i^2) = a_{i^3}$, ..., $\mu(i^m) = a_{i^1}$.
- Let $\hat{N}^k$ be the set of agents that are not part of any cycle in G^k.
  Let $\hat{A}^k$ be the set of houses initially owned by $\hat{N}^k$.
- If $\hat{N}^k$ is empty, **stop**: μ is the final matching.
  Otherwise, set $N^{k+1} = \hat{N}^k$ and $A^{k+1} = \hat{A}^k$ and start round $k + 1$.

TTC: Example

six agents and six houses (agent i owns house a_i)

Preferences:

$$P(1) = a_3, a_1, a_2, a_4, a_5, a_6 \qquad P(4) = a_2, a_1, a_5, a_4, a_3, a_6$$
$$P(2) = a_3, a_2, a_1, a_5, a_4, a_6 \qquad P(5) = a_2, a_1, a_6, a_4, a_5, a_3$$
$$P(3) = a_1, a_4, a_3, a_2, a_6, a_5 \qquad P(6) = a_1, a_3, a_2, a_4, a_5, a_6$$

Outcome of the TTC:

$$\mu = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \\ a_3 & a_2 & a_1 & a_5 & a_6 & a_4 \end{matrix}$$

Properties of the TTC

The TTC is well-defined.

- The algorithm stops after at most $|N|$ rounds: In each round, there is at least one cycle. Hence, at least one agent is removed in each round.
- Feasibility: the TTC produces a matching.
- Individual rationality: by pointing at herself, an agent can always form a cycle and keep her initially owned house.

In the following, we will prove the theorem using the TTC:

1. The matching produced by the TTC is in the core.
2. The matching produced by the TTC is the only core matching.

Lemma

The matching produced by the TTC is in the core.

Proof.

Let μ be the outcome of the TTC. Suppose, by contradiction, that μ is not in the core, i.e., there is a coalition $S \subseteq N$ and a matching μ' satisfying (i) & (ii) in the definition of the core. Let $S^* := \{i \in S : \mu'(i) \succ_i \mu(i)\}$ denote the members of the coalition that strictly prefer μ' over the TTC. (ii) implies $|S^*| \geq 1$.

- Let round k be the earliest round of the TTC where some agent in S^* is matched to a house. As all agents in S who are matched before round k are indifferent between μ' and μ, they get the same house under μ' and μ.
- Consider an agent $i \in S^*$ who is matched in round k of the TTC. As $\mu(i)$ is i's most preferred house out of those still available in round k, all the houses that i prefers over $\mu(i)$ are matched before round k.
- But also under μ' agent i cannot get a better house because all the better houses owned by agents in S are taken by the same agents as in the TTC. $\Longrightarrow$ $S^* = \varnothing$, contradiction. □

Lemma

The matching produced by the TTC is the only core matching.

Proof.

Let μ be the outcome of the TTC and consider any matching $\hat{\mu} \neq \mu$.

- Let round k be the earliest round of the TTC where some agent i is matched to a house $\mu(i) \neq \hat{\mu}(i)$. Let C_k denote the agents that form the cycle including i in round k.
- For all $j \in C_k$, $\mu(j) \succeq_j \hat{\mu}(j)$, as in the TTC each j obtains her best house still available in round k and all the unavailable houses are also unavailable under $\hat{\mu}$. Moreover, $\mu(i) \succ_i \hat{\mu}(i)$ because $\mu(i) \neq \hat{\mu}(i)$.

$\Longrightarrow$ The coalition $S = C_k$ using μ is better off than under $\hat{\mu}$ (in the sense of (i) & (ii) in the definition of the core).

$\Longrightarrow$ $\hat{\mu}$ is not in the core. $\square$

5.3 Incentives

- Let $\mathcal{P}_i$ denote set of all possible preferences for agent i and define $\mathcal{P} := \mathcal{P}_1 \times \mathcal{P}_2 \times \cdots \times \mathcal{P}_n$.
- Let Ψ denote the set of possible matchings.

Definition

A direct **matching mechanism** is a function $h : \mathcal{P} \to \Psi$ that chooses a matching $\mu = h(\mathbf{Q})$ for every reported profile of preferences $\mathbf{Q}$.

A mechanism h is **efficient** (**individually rational**) if, for all $\mathbf{Q} \in \mathcal{P}$, $h(\mathbf{Q})$ is an efficient (individually rational) matching under preferences $\mathbf{Q}$.

A mechanism h is **strategy-proof** if, for all $i \in N$ and true preferences $P(i) \in \mathcal{P}_i$,

$$h(P(i), Q_{-i}) \succeq_i h(Q(i), Q_{-i}) \quad \text{for all } Q(i) \in \mathcal{P}_i \text{ and } Q_{-i} \in \mathcal{P}_{-i}.$$

The TTC mechanism

- The **TTC mechanism** h^T is the mechanism that chooses the core matching for the reported preferences, i.e., it runs the TTC using the reports.
- As h^T produces the core matching, h^T is efficient and individually rational.

Theorem

The TTC mechanism is strategy-proof.

- 3 compelling properties: strategy-proof, efficient, and individually rational
- Remarkably, the 3 properties characterize the TTC mechanism:
  It can be shown that a mechanism is strategy-proof, efficient, and individually rational *if and only if* it is the TTC mechanism (Ma, 1994).

Proof.

Consider agent i and hold the reports Q_{-i} of the other agents fixed.
Let round k be the round in which agent i is matched if i reports truthfully.
Are there profitable deviations from truthful reporting?

Deviations that result in i being matched in round k or later are not profitable, as truthful reporting ensures i the most preferred house still available in round k.

Suppose agent i deviates such that she is already matched in round $r < k$.

- Hence, i is part of a subset of agents C_r that form a cycle in round r.
- If i reports truthfully, the agents in C_r stay unmatched in round r. Moreover, i stays unmatched and house a_i stays available until round k. Hence, the agent in C_r that points to i in round r will continue to point to i until round k and stay unmatched. Similarly, all agents in C_r will point to the same agents in rounds r to k and stay unmatched.
- Consequently, being matched to the house of an agent in C_r in round r cannot be profitable for i because the same house is also still available in round k if i reports truthfully. $\implies$ no profitable deviations $\square$

5.4 Applications

The TTC mechanism can be adapted to the **school choice** problem (Abdulkadiroglu and Sönmez, 2003):

- Each student points to her top school and each school points to the student with the highest priority. Students in cycles are assigned and removed. Schools are removed once their capacity is exhausted.

- allows students to trade schools for which they have the highest priority
- strategy-proof and Pareto efficient, but not *fair*

Kidney exchange

- In most countries, illegal to trade human organs for money.
  But organs, such as kidneys, can be donated.
- Patients may have a willing donor that is incompatible (blood type etc.)
  $\rightarrow$ possibility of lifesaving exchanges

- Kidney exchange similar to house allocation model:
  - each patient (agent) has an incompatible donor (initially owned house)
  - physician determines suitability of kidneys of other donors (preferences)

  $\Longrightarrow$ in principle, can use TTC

- TTC needs to be adapted to practical application:
  - exchange cycles with more than three patient-donor pairs usually impossible (simultaneous surgeries needed to prevent donors from retracting).
  - waiting lists and deceased donors
  - undirected donor (rare) may start chain

- Modified TTCs in use by various kidney exchange programs

6 References

- Abdulkadiroglu, A., and Sönmez, T. (2003). School Choice: A Mechanism Design Approach. *American Economic Review*, 93(3), 729–747.
- Abdulkadiroglu, A., Pathak, P.A., Roth, A.E., and Sönmez, T. (2006). Changing the Boston School Choice Mechanism. NBER Working Paper No. 11965.
- Dubins, L.E., and Freedman, D.A. (1981). Machiavelli and the Gale-Shapley Algorithm. *American Mathematical Monthly*, 88(7), 485–494.
- Ergin, H., and Sönmez, T. (2006). Games of school choice under the Boston mechanism. *Journal of Public Economics*, 90(1), 215–237.

- Gale, D., and Shapley, L.S. (1962). College admissions and the stability of marriage. *American Mathematical Monthly*, 69(1), 9–15.
- Kagel, J.H., and Roth, A.E. (2000). The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment. *Quarterly Journal of Economics*, 115(1), 201–235.
- Ma, J. (1994). Strategy-proofness and the strict core in a market with indivisibilities. *International Journal of Game Theory*, 23, 75–83.
- Roth, A.E. (1982). The Economics of Matching: Stability and Incentives. *Mathematics of Operations Research*, 7(4), 617–628.
- Roth, A.E. (1984). The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *Journal of Political Economy*, 92(6), 991–1016.
- Roth, A.E. (1991). A natural experiment in the organization of entry-level labor markets: regional markets for new physicians and surgeons in the United Kingdom. *American Economic Review*, 81(3), 415–440.
- Roth, A.E., and Peranson, E. (1999). The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design. *American Economic Review*, 89(4), 748–780.
- Shapley, L., and Scarf, H. (1974). On cores and indivisibility. *Journal of Mathematical Economics*, 1(1), 23–37.
- Westkamp, A. (2013). An analysis of the German university admissions system. *Economic Theory*, 53(3), 561–589.