Matching and Stable Allocations

Game Theory · Module 6 — Mechanism Design

Node 6.6

Position in Knowledge Graph

System: Game Theory **Structural Domain:** Mechanism Design **Node:** 6.6

Matching theory is mechanism design without money. When the designer cannot use transfers — because the setting is ethically, legally, or practically incompatible with prices (medical residencies, school choice, kidney exchange, marriage) — the IC and efficiency questions take a different form. The central concept shifts from "incentive-compatible allocation + payment" to **stability**: no pair of agents should prefer to abandon their assigned match and form a new pair. The foundational algorithm is **Gale–Shapley deferred acceptance** (1962), which produces a stable matching in any two-sided market with strict preferences. The foundational impossibility is that no stable mechanism can be strategy-proof for both sides simultaneously (Roth 1982). The applied revolution is **Alvin Roth's market design** program, which redesigned the National Resident Matching Program (NRMP), school choice in New York and Boston, and kidney exchange — all using the structural insights of matching theory. This node develops the mathematical foundations, the key impossibility results, and the connection between stability and incentive compatibility.

Formal Definition

Two-sided matching market. Two finite disjoint sets: $M = \{m_1, \ldots, m_p\}$ (e.g., men, firms, hospitals) and $W = \{w_1, \ldots, w_q\}$ (e.g., women, workers, residents). Each $m \in M$ has a strict preference ordering $\succ_m$ over $W \cup \{m\}$ (where m represents "being unmatched"). Each $w \in W$ has a strict preference ordering $\succ_w$ over $M \cup \{w\}$. An agent a finds partner b **acceptable** if $b \succ_a a$.

A **matching** $\mu : M \cup W \to M \cup W$ satisfies: 1. $\mu(m) \in W \cup \{m\}$ for all $m \in M$. 2. $\mu(w) \in M \cup \{w\}$ for all $w \in W$. 3. $\mu(m) = w \iff \mu(w) = m$.

A pair (m, w) **blocks** matching μ if both $w \succ_m \mu(m)$ and $m \succ_w \mu(w)$: both prefer each other to their current assignment. A matching is **stable** if: - No agent is matched to an unacceptable partner (**individual rationality**). - No blocking pair exists (**pairwise stability**).

Gale–Shapley deferred acceptance (DA). The M-proposing DA algorithm:

1. Each unmatched m proposes to the highest-ranked acceptable w not yet rejected.
2. Each w receiving proposals tentatively accepts the best acceptable proposer and rejects the rest.
3. Repeat until no unmatched m has acceptable proposals left.

Strategy-proofness. A matching mechanism ϕ is **strategy-proof** for side M if no $m \in M$ can obtain a strictly preferred match by misreporting their preference ordering, regardless of others' reports.

Intuition

Stability is the matching-theoretic analogue of individual rationality plus incentive compatibility. In a market with money, IC is enforced by transfers: if you lie about your value, the payment rule makes you worse off. In a market without money, stability is enforced by the threat of **blocking pairs**: if the mechanism produces a matching where Alice and Bob would both prefer to be matched with each other, they will deviate from the mechanism and match privately. A stable matching is one where no such deviation is profitable for any pair.

Gale–Shapley's deferred acceptance is remarkable for three reasons. First, it always produces a stable matching — this is not obvious, since "stable matching exists" is a non-trivial theorem (Gale and Shapley 1962). Second, it produces the **M-optimal stable matching**: among all stable matchings, each $m \in M$ gets their best possible partner. Third, the algorithm is computationally efficient (polynomial time), which matters for practical market design.

The M-optimal stable matching is simultaneously the **W-pessimal** stable matching: each $w \in W$ gets their *worst* stable partner. This is a fundamental structural asymmetry: the proposing side benefits from deferred acceptance, the receiving side does not. This asymmetry extends to strategic properties: Roth (1982) showed that the M-proposing DA is **strategy-proof for** M but *not* for W. No stable mechanism can be strategy-proof for both sides.

The practical lesson: who proposes matters. In the NRMP, the algorithm was redesigned (at Roth's recommendation) from hospital-proposing to resident-proposing, shifting the strategic advantage to the residents. In school choice, the mechanism was redesigned from a manipulable Boston mechanism to a strategy-proof deferred acceptance, at the cost of some efficiency.

Structural Role

6.6 is the capstone of Module 6, showing what mechanism design looks like when the designer's most powerful tool — money — is removed. The structural progression:

- **6.1–6.4** develop mechanism design with transfers: VCG, Myerson's optimal auction, the full quasi-linear toolkit.
- **6.5** asks about robustness (full implementation).
- **6.6** asks: what if transfers are unavailable?

Without money, the envelope/payment formula from 6.3 is gone. The designer cannot use "pay more, get more" to elicit truthful reports. Instead, the designer relies on the structure of preferences and the threat of instability. This makes matching theory both more constrained (fewer tools) and more applied (the settings where money cannot be used are the settings where market design matters most).

The impossibility of two-sided strategy-proofness means the designer must make a choice: which side gets the strategic advantage? This is a distributional question, not an efficiency question (all stable matchings are Pareto-efficient for the proposing side), and it connects matching theory to questions of fairness and power in market design.

The connections to other modules:

- **6.1 Gibbard–Satterthwaite** returns: the impossibility of two-sided strategy-proofness in stable matching is a consequence of the general impossibility of non-dictatorial strategy-proof rules over rich preference domains.
- **6.5 Implementation** connects: stability can be viewed as a Nash implementation requirement (a matching is stable iff it is a core allocation, and the core is Nash-implementable under certain conditions).
- **Module 3 extensive form** connects through the sequential structure of DA, which can be modeled as an extensive-form game.

Mathematical Development

Existence of stable matchings

Theorem (Gale–Shapley 1962). For any two-sided matching market with strict preferences, a stable matching exists. The M-proposing DA algorithm terminates in at most $|M| \cdot |W|$ steps and produces a stable matching.

Proof of stability. Let μ be the matching produced by M-proposing DA. Suppose (m, w) is a blocking pair: $w \succ_m \mu(m)$ and $m \succ_w \mu(w)$. Since $w \succ_m \mu(m)$, agent m proposed to w before proposing to $\mu(m)$ (agents propose in preference order). Since m is not matched to w, agent w must have rejected m at some point — either immediately or by replacing m with a better proposer. But DA only replaces with a strictly better proposer, so w's final match $\mu(w) \succ_w m$. This contradicts $m \succ_w \mu(w)$. ■

M-optimality and W-pessimality

Theorem. The M-proposing DA produces the M-optimal stable matching: for every m, $\mu^M(m) \succeq_m \mu(m)$ for any other stable matching μ. Simultaneously, it produces the W-pessimal stable matching: for every w, $\mu(w) \succeq_w \mu^M(w)$ for any other stable matching μ.

Proof sketch. Suppose, for contradiction, that some m is rejected by a partner w whom they get in some stable matching μ. Consider the first such rejection in the execution of DA. At this moment, w holds a proposal from some m' with $m' \succ_w m$. Since this is the *first* rejection of a "stable partner," m' has not yet been rejected by anyone they are matched to in any stable matching — in particular, $w \succeq_{m'} \mu(m')$. But then (m', w) blocks μ: $w \succ_{m'} \mu(m')$ (since m' proposed to w before $\mu(m')$) and $m' \succ_w m \succeq_w \mu(w)$ (since $\mu(w) = m$ or worse). Contradiction with stability of μ. ■

Lattice structure of stable matchings

Theorem (Conway; Knuth 1976). The set of stable matchings forms a **distributive lattice** under the partial order $\mu \succeq_M \mu'$ iff every m weakly prefers $\mu(m)$ to $\mu'(m)$. The M-optimal matching is the lattice's top element; the W-optimal (= M-pessimal) matching is the bottom element.

This means: given any two stable matchings μ, μ', there exist stable matchings $\mu \vee \mu'$ and $\mu \wedge \mu'$ (join and meet), where the join assigns each m their preferred partner from the two matchings, and the meet assigns each m their less-preferred partner. Both are stable.

Strategy-proofness of DA

Theorem (Roth 1982). The M-proposing DA is strategy-proof for M: no m can obtain a strictly better partner by misreporting preferences, regardless of others' reports.

Proof sketch. Suppose m misreports $\succ'_m \neq \succ_m$ and gets partner $w' \succ_m w = \mu^M(m)$ under the M-proposing DA with the misreport. Consider the "true" M-proposing DA. Since m was rejected by w' in the true DA, there was some step where w' held a proposal from $m'' \succ_{w'} m$. The misreport changes m's proposal sequence but cannot change w''s comparison between m and m''. One shows that the misreport creates a blocking pair in the resulting matching under true preferences — contradicting stability. The formal proof uses the "rural hospitals theorem" and the lattice structure. ■

Theorem (Roth 1982). No stable mechanism is strategy-proof for both sides simultaneously.

Roth's market design: the NRMP

The National Resident Matching Program matches medical graduates to residencies. Before 1998, the algorithm was hospital-proposing DA, which gave hospitals their optimal stable match. Roth analyzed the strategic incentives and recommended switching to applicant-proposing DA, giving applicants strategy-proofness. The redesigned NRMP has been in use since 1998 and handles ~35,000 matches annually.

The key design insight: in a market where one side (residents) has weaker bargaining power, strategy-proofness for that side is more valuable, because weak-side agents are less likely to have the information or sophistication to manipulate successfully.

Many-to-one matching and the Kelso–Crawford extension

In hospital-resident matching, each hospital has multiple slots. The model extends to **many-to-one matching**: each hospital h has a quota q_h and preferences over *groups* of residents. When hospital preferences satisfy **substitutability** (dropping a resident from a group cannot make a previously rejected resident acceptable), a stable matching exists and the hospital-proposing DA finds the hospital-optimal stable matching. Kelso and Crawford (1982) proved this by connecting matching to a "salary adjustment" process in a model with transfers — matching theory's deepest link to general equilibrium.

The rural hospitals theorem

Theorem (Roth 1986). In any stable matching of a many-to-one market, the set of unmatched hospitals (those with unfilled positions) and the set of unmatched residents are the same across *all* stable matchings. Any hospital that has unfilled positions in one stable matching has the same number of unfilled positions in every stable matching, and is matched to exactly the same set of residents.

This theorem was empirically important: rural hospitals, which often fail to fill positions, cannot blame the matching algorithm. The algorithm's choice among stable matchings does not affect which hospitals go unfilled — that is a structural feature of the preferences, not an artifact of the mechanism.

Top Trading Cycles and one-sided assignment

In the **housing market** (Shapley–Scarf 1974), each agent owns one house and has preferences over all houses. The **Top Trading Cycles** (TTC) algorithm finds the unique core allocation: each agent points to the owner of their most-preferred house, forming directed cycles; agents in each cycle trade; repeat on the reduced market. TTC is DSIC (Ma 1994) and Pareto-efficient. Unlike DA, TTC works in a one-sided market (no "proposing" and "receiving" sides), but it requires initial endowments.

Worked Example

A 3x3 matching market

Men: m_1, m_2, m_3. **Women:** w_1, w_2, w_3.

Preferences: - $m_1 : w_1 \succ w_2 \succ w_3$ - $m_2 : w_1 \succ w_3 \succ w_2$ - $m_3 : w_1 \succ w_2 \succ w_3$ - $w_1 : m_2 \succ m_1 \succ m_3$ - $w_2 : m_1 \succ m_3 \succ m_2$ - $w_3 : m_1 \succ m_2 \succ m_3$

M-proposing DA execution.

Round 1: m_1, m_2, m_3 all propose to w_1. w_1 holds m_2 (top choice), rejects m_1, m_3.

Round 2: m_1 proposes to w_2 (next on list). m_3 proposes to w_2. w_2 holds m_1 (prefers m_1 to m_3), rejects m_3.

Round 3: m_3 proposes to w_3. w_3 holds m_3.

Final matching: $\mu^M = \{(m_1, w_2), (m_2, w_1), (m_3, w_3)\}$.

Verify stability. Check all non-matched pairs: - (m_1, w_1): w_1 has $m_2 \succ m_1$, so w_1 does not prefer m_1. Not blocking. - (m_1, w_3): m_1 has $w_2 \succ w_3$. Not blocking (m_1 does not prefer w_3). - (m_2, w_2): m_2 has $w_1 \succ w_2$. Not blocking. - (m_2, w_3): m_2 has $w_1 \succ w_3$. Not blocking. - (m_3, w_1): w_1 has $m_2 \succ m_3$. Not blocking. - (m_3, w_2): w_2 has $m_1 \succ m_3$. Not blocking.

Stable. No blocking pairs.

W-proposing DA would produce a different matching (the W-optimal = M-pessimal stable matching). The lattice structure guarantees both exist.

Poker Anchor

Concept mapping

Formal object	Poker instantiation
Two-sided market	Players seeking tables; tables seeking players (or: players seeking staking deals; backers seeking horses)
Preferences over partners	Player prefers softer tables; table ecosystem prefers balanced lineups
Stable matching	A table assignment where no player-table pair would mutually prefer to switch
Blocking pair	A player on a tough table and a soft table with an open seat that would welcome them
Deferred acceptance	Waitlist systems at casinos: players request tables in preference order; floor assigns based on game health
Strategy-proofness	Whether reporting your true table preference is optimal under the floor's assignment rule

Concrete situation: staking as a matching market

The poker staking market is a two-sided matching problem. **Horses** (players seeking backing) have preferences over backers (stake percentage, markup, freedom, reputation). **Backers** have preferences over horses (skill level, volume, game selection, reliability). A staking deal is a match. A deal is "unstable" if there exists a horse-backer pair who would both prefer to be matched with each other over their current deal — they would break their current contracts and form a new one.

In practice, the staking market is unstructured: deals are formed through personal networks, forums, and negotiation. There is no centralized matching mechanism. The result is frequent instability: horses switch backers mid-deal, backers poach horses from other stables, and information asymmetry is rampant (horses misrepresent their results, backers misrepresent their terms).

A Gale–Shapley-style centralized mechanism for staking — where horses submit preference rankings over backers and vice versa, and a DA algorithm produces a stable matching — would eliminate the instability but would require standardized terms and truthful preference reporting. Strategy-proofness for one side (say, horses under horse-proposing DA) would protect unsophisticated players from manipulation.

Concrete situation: table changes and seat assignments

In a casino with multiple tables of the same game, the floor manager's seat-assignment problem is a matching problem. Players have preferences (softer games, specific opponents to avoid, comfort). Tables have implicit "preferences" (game health requires balanced skill levels, balanced stack sizes). The floor manager is the mechanism designer.

Current practice is ad-hoc: first-come-first-served or floor discretion. This is neither stable (players constantly request changes) nor strategy-proof (players misrepresent their preferences to get better seats). A formal DA mechanism — players rank available seats, floor ranks players for each seat based on game-health criteria — would produce a stable assignment and be strategy-proof for players.

What this understanding buys you

The staking market is inefficient because it lacks a centralized stable mechanism. Every time a horse leaves a backer or a backer poaches a horse, that is a blocking pair being resolved informally. Understanding matching theory tells you that a centralized mechanism would produce a stable outcome, eliminating the churn. It also tells you the fundamental impossibility: you cannot make the mechanism strategy-proof for both horses and backers simultaneously.

Table selection is the matching problem you solve every session. When you choose which table to sit at, you are solving a one-sided matching problem: you have preferences over tables, and you want to find your best available match. Understanding that the set of stable matchings forms a lattice tells you that there is a "best table for you" and a "best table for the casino" and they may not be the same — the casino's seating algorithm, if designed well, balances these interests.

Tournament seating draws are random matchings, and randomness is a design choice. In tournaments, players are randomly assigned to tables. Randomness ensures no player can manipulate their table draw — a brute-force strategy-proofness guarantee. The cost is that randomness does not optimize for game quality, balance, or entertainment. A DA-based mechanism could do better on those dimensions but would sacrifice the simplicity and perceived fairness of random assignment.

The lattice structure appears in range construction. When building a strategy, you often face a choice between two "stable" range constructions (each internally consistent, no blocking pair of "hand-that-should-be-in-range and hand-that-should-be-out"). The lattice structure of stable matchings predicts that there is a "most aggressive" stable range (analogous to the proposer-optimal matching) and a "most passive" stable range (the receiver-optimal), and that both are valid equilibrium constructions. The choice between them is a design decision driven by your strategic goals, just as the NRMP's choice between hospital-proposing and resident-proposing DA is a design decision driven by policy goals.

Kidney exchange is the purest mechanism design without money. Roth, Sonmez, and Unver (2004) applied matching theory to kidney exchange: patients with incompatible donors are matched in cycles so that each patient receives a compatible kidney. No money changes hands (illegal for organ sales). The TTC algorithm finds the efficient allocation; strategy-proofness ensures patients report their true compatibility information. This is mechanism design at its most consequential — lives saved by the correct application of matching theory.

Cross-Links

- **6.1 Social choice and IC** — stability is the matching-theoretic analogue of IC + IR.
- **6.2 Revelation principle** — DA can be viewed as a direct revelation mechanism for one side.
- **6.5 Implementation theory** — stability as a Nash implementation requirement (core implementation).
- **7.x Evolutionary game theory** — matching in biological markets (mate choice, mutualism).
- **Economics (labor markets)** — Roth's market design for medical residencies, school choice.
- **Computer science (algorithms)** — computational complexity of finding stable matchings and extensions.

Structural Compression

Invariant: In any two-sided matching market with strict preferences, a stable matching exists. The set of stable matchings forms a distributive lattice. The proposing-side-optimal stable matching is produced by deferred acceptance and is strategy-proof for the proposing side. No stable mechanism is strategy-proof for both sides.

Mechanism: Deferred acceptance: agents on one side propose in preference order; agents on the other side hold their best acceptable proposal and reject the rest; rejected agents propose to their next choice;

the process terminates at a stable matching. Stability is enforced by the structure of tentative holding and rejection: any blocking pair would have been formed during the algorithm's execution.

Cross-System Echo: Stable matching is the discrete-combinatorial analogue of **competitive equilibrium** in general equilibrium theory. In both, the solution concept is a fixed point where no agent or pair wants to deviate. The lattice structure of stable matchings mirrors the lattice of Walrasian equilibria in economies with gross substitutes (Kelso–Crawford 1982). The fundamental impossibility of two-sided strategy-proofness mirrors the impossibility of Walrasian equilibrium being robust to all forms of manipulation. Gale–Shapley is the matching-theoretic analogue of tatonnement: an iterative adjustment process that converges to equilibrium.

Exercises

1. Run the W-proposing DA on the 3x3 market from the Worked Example. Verify that the resulting matching is stable, that it is W-optimal, and that it differs from the M-proposing result.
2. Show that in a market where all agents on both sides have identical preferences (all men rank women the same way, all women rank men the same way), there is a unique stable matching. What is it?
3. Construct a 3x3 matching market with at least three distinct stable matchings. Verify the lattice structure by computing the join and meet of two of them.
4. Prove that the M-proposing DA terminates in at most $|M| \cdot |W|$ steps. (Hint: each man proposes to each woman at most once.)
5. In the NRMP, couples submit joint preference lists (pairs of positions). Show by example that stable matchings may not exist when couples are present. (Hint: construct a small market with one couple and show that every matching has a blocking pair or a blocking coalition involving the couple.)
6. Consider a one-sided matching problem (roommate problem): n agents, each with preferences over all others, matched into pairs. Show by example (with $n = 4$) that a stable matching may not exist. (Hint: construct cyclic preferences.) Relate this to the two-sided existence theorem.
7. **Argue, structurally,** why strategy-proofness for both sides is impossible in any stable matching mechanism with at least two agents per side. Identify the structural feature of two-sided preferences that forces the tradeoff — specifically, why the proposing side's optimality necessarily comes at the receiving side's expense, and why this asymmetry prevents simultaneous strategy-proofness.

Game Theory · Module 6 — Mechanism Design · Node 6.6
Concepts: mechanism-design, incentive-compatibility, social-choice
Prerequisites: 6.1, 6.2, 6.5

hbar.university — own.your.cognition