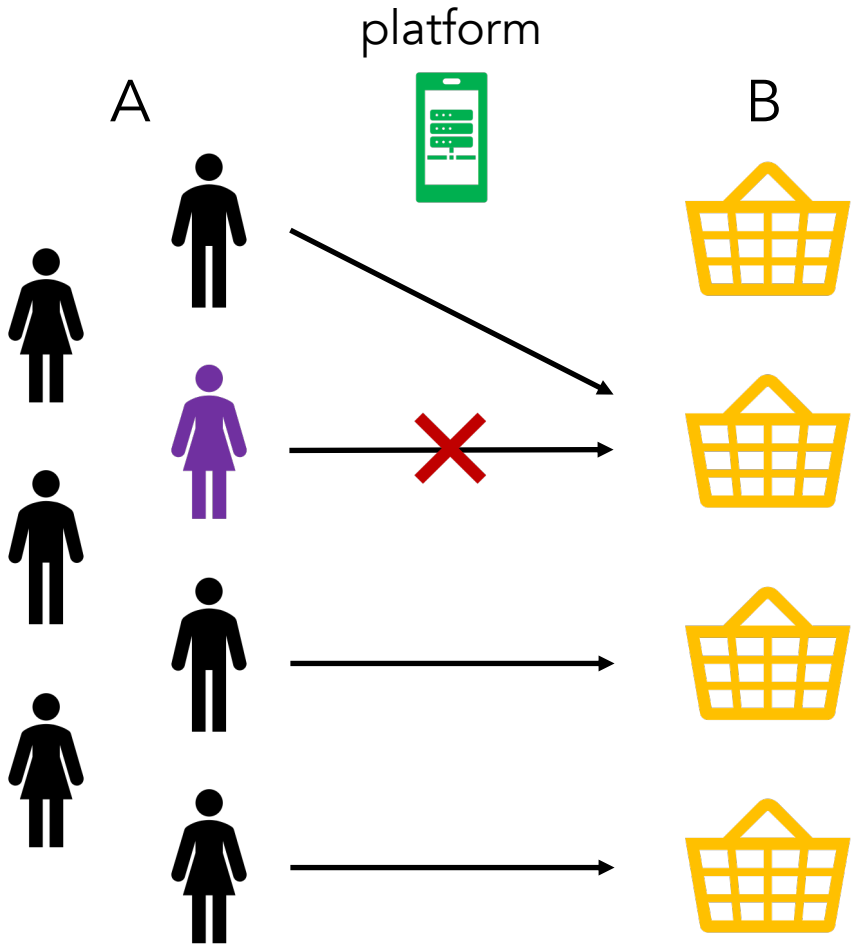
# Regret, stability, and fairness in matching markets with bandit learners

**Sarah Cen** and Devavrat Shah

LIDS Stats & Tea Talk
May 5, 2021
arXiv link here

# Motivation




Agents wish to be **matched**.

Agents in A have **preferences** over B.

Agents in B have **preferences** over A.

**Compete** for finite number of matches.

Agents **learn preferences** over time.

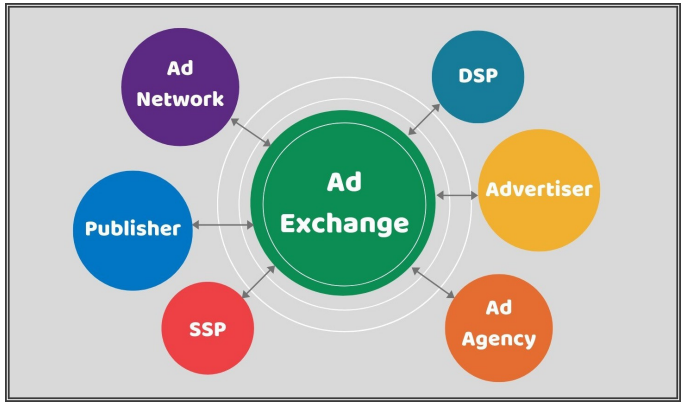
**Challenge: How does competition affect the agents' ability to learn and their regret?**

# Examples

Dating

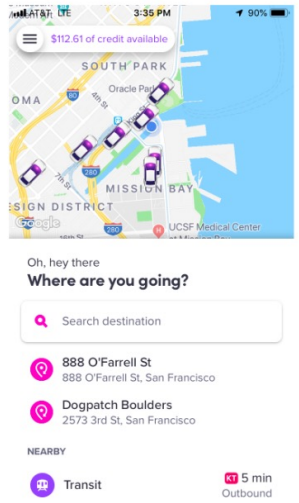

marketwatch.com

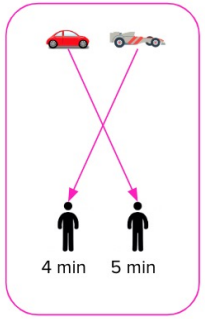
Ad exchanges




adpushup.com

Ride sharing




eng.lyft.com

Individuals compete while learning their preferences from a finite number of interactions.

Model sequential decision-making of learning agents competing for resources under uncertainty.

# Game theory + reinforcement learning

Game theory:

- Competing for limited resources.
- Preferences.
- Desirable to reach equilibrium.

Reinforcement learning:

- Learn through interactions.
- Maximize long-term reward.
- Exploitation vs. exploration.

# What do we want from our system?

Suppose we're helping the **platform** to design or evaluate their **matching**.

Stability: agents are not incentivized to leave the platform.

Low regret: learning under competition does not come at a high cost.

Fairness: good performance for some is not at the expense of others.

Social welfare: utilitarian performance measure (sum of all agents' utilities).

We consider these four but not others (e.g., strategy-proofness).

# Today (high-level)

Modeling this problem: Matching & MAB problems.

Stable matching at every time step → efficient learning for free*!

Pessimal regret grows $O(\log T)$ [Liu, Mania & Jordan '20].

*Optimal regret can be $\Omega(T)$ + fairness & high SW not guaranteed.

Costs and transfers between agents [Cen & Shah '21].

1. Aspect of competition + exogeneous effects.
2. With structure, can simultaneously guarantee all four criteria.

# Matching Markets with Bandit Learners

# Matching & MAB Problems

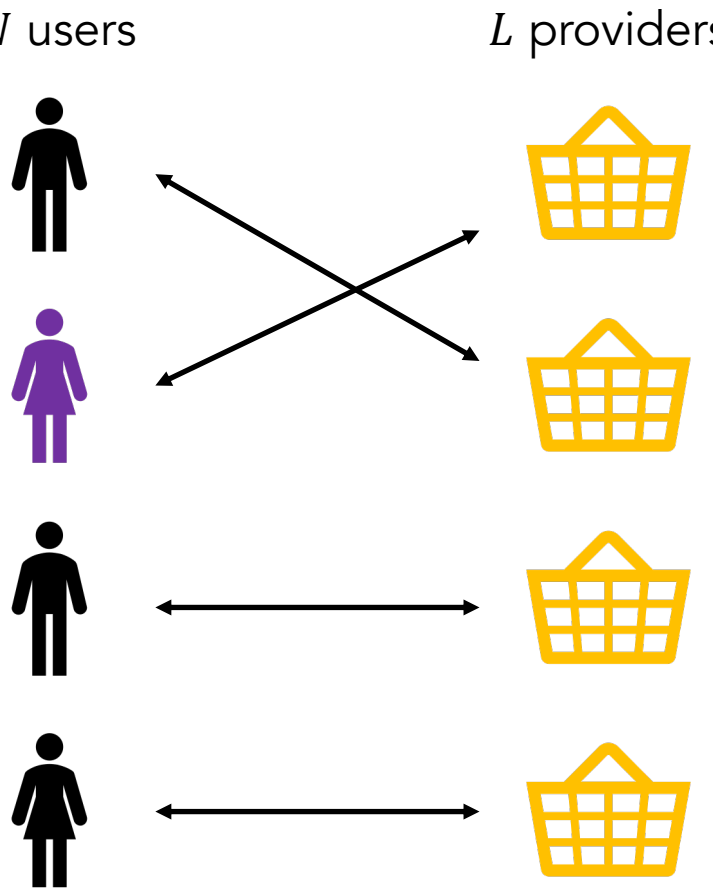
How to combine game theory and RL?

**Matching + MAB.**

"Matching market with bandit learners"

Introduced by Das & Kamenica (2005).
Important results by Liu, Mania & Jordan (2020).

# Two-sided matching problem




$N \geq L$

Known preferences (two-sided)

$u_2 : p_1 \succ p_2 \succ p_3 \succ p_4$

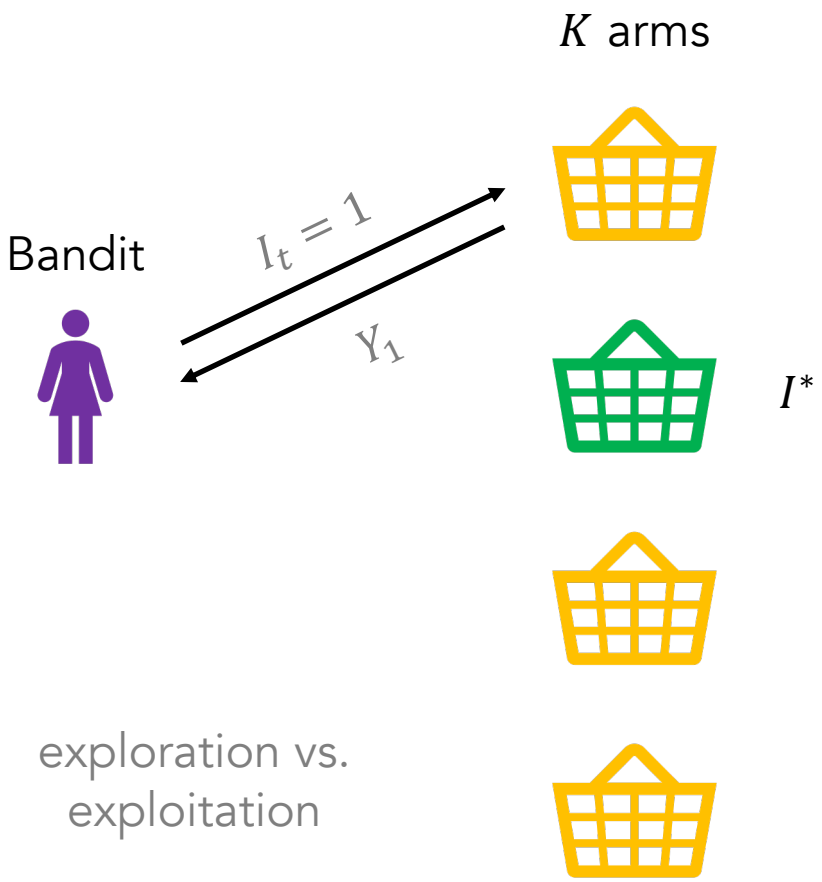
One-to-one matching $\mathcal{M}$

**Stability**: No user-provider pair is incentivized to defect from $\mathcal{M}$ and pair off together.

$u_3 : p_4 \succ p_3$ but $p_4 : u_4 \succ u_3$

**But do agents know their preferences a priori?**

# Multi-armed bandit problem




Agent do not know their own preferences a priori.
Learn from repeated interactions.

Maximize reward over $T$ time steps.

At each $t \in [T]$:

(1) Choose arm $I_t$

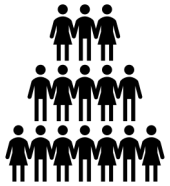
(2) Receive noisy reward $Y_{I_t}$.

Minimize regret: $R = \mathrm{E}\left[\sum_{t=1}^{T} X_{I*} - X_{I_t}\right]$

Goal: $R = O(\log T)$.

# Matching market with bandit learners

$N$ users



$\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$

$N \geq L$
$\mathcal{A} = \mathcal{U} \cup \mathcal{P}$
$\mathcal{A}^+ = \mathcal{A} \cup \{\emptyset\}$

$L$ providers

$\mathcal{P} = \{p_1, p_2, \ldots, p_L\}$

---

True (unknown) preferences $\mu(a_1, a_2) \in \mathbb{R}$.
Two-sided preferences. No ties → $\succ$.

Bandit learners: $\nu_t(a_1, a_2) = \hat{\mu}_t(a_1, a_2) + \sqrt{\dfrac{2\,\sigma^2 \alpha \log t}{T_{t-1}(a_1, a_2)}}$

## Centralized matching

At each time step $t \in [T]$:

$\nu_t$ is information state.

Platform's matching $\mathcal{M}(\cdot\,; \nu_t)$.

SubG reward $X_t\big(a, \mathcal{M}(a; \nu_t)\big)$.

Cost $\mathcal{C}(a, \mathcal{M}(a; \nu_t)\,; \nu_t)$.

Transfer $\mathcal{T}(a, \mathcal{M}(a; \nu_t)\,; \nu_t)$.

Observed payoffs

$$\mathrm{U}(\cdot,\cdot\,; \nu_t) = X_t\,(\cdot,\cdot) - \mathcal{C}(\cdot,\cdot\,; \nu_t) + \mathcal{T}(\cdot,\cdot\,; \nu_t)$$

# Matching market with bandit learners

Matching process:

At $t = 0$, platform decides on $(\mathcal{M}, \mathcal{C}, \mathcal{T})$.These rules are made known to all agents.

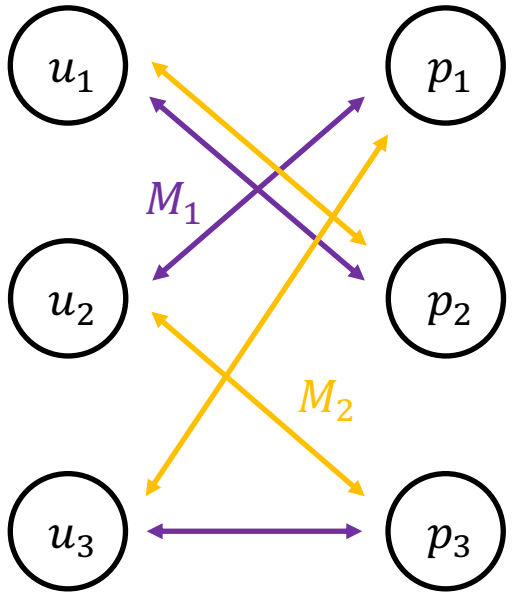
At each $t \in [T]$:

1) Update. Agents update estimates $\hat{\mu}_t(\cdot,\cdot)$.

2) Report. Agents report UCB preferences $\nu_t(\cdot,\cdot)$.

3) Match. Platform implements $\mathcal{M}$ + agents observe $X_t$.

4) Pay and transfer. Agents pay $\mathcal{C}$ and $\mathcal{T}$.




Agents observe own information only.

# Evaluating performance

Stability: No pair of agents is incentivized to defect.

$$\nexists a, a' : V\big(a, M(a)\big) < V(a, a') \cap V\big(a', M(a')\big) < V(a', a)$$

Low (optimal) regret: Competition does not prevent learning.

$$\bar{R}(a; \mathcal{M}) = O(\log T) \quad \forall a \in \mathcal{A}$$

Fairness: Regret is distributed evenly across agents.

$$\nexists a, a' : \bar{R}(a; \mathcal{M}) = O(\log T) \cap \bar{R}(a'; \mathcal{M}) = \omega(\log T)$$

High social welfare: Utilitarian measure of global performance.

$$W_t(\mathcal{M}) \geq \max_{M \in \mathcal{W}} W_t(M) \,/\, 2$$

# Main results

# Stability: Gale-Shapley algorithm

Stability: no user-provider pair incentivized to leave platform.

**Lemma**. Running the GS algorithm over UCB payoffs $V(\cdot, \cdot; \nu_t)$ at every time step $t \in [T]$ ensures stability is satisfied.

For remainder, assume $\mathcal{M}(\cdot; \nu_t)$ is product of GS algorithm.

# No cost, no transfer

$$\mathcal{C}(a, a'; \psi) = 0$$
$$\mathcal{T}(a, a'; \psi) = 0$$

**Proposition 3**. Under no costs or transfers, applying the GS algorithm at every time step gives:

$$\underline{R}(a; \mathcal{M}) \leq 2N^2 L \Delta_{\max}(a) \left( \frac{8\sigma^2 \alpha \log T}{\Delta_{\min}^2} + \frac{\alpha}{\alpha - 2} \right).$$

However, under stability, there exist $(\mathcal{A}, \mu)$ and $a$ such that $\bar{R}(a; \mathcal{M}) = \Omega(\mathrm{T})$.

# Proportional cost

$\gamma \in [0,1]$

$$\mathcal{C}(a, a'; \psi) = \gamma\, \psi(a, a')$$
$$\mathcal{T}(a, a'; \psi) = 0$$

Ex. advertisers pay bids, users pay in time.

**Proposition 4**. Under proportional costs and $\gamma \in [0,1)$, applying the GS algorithm at every time step gives:

$$\underline{R}(a; \mathcal{M}) \leq 2N^2 L(1-\gamma)\Delta_{\max}(a)\left(\frac{8\sigma^2 \alpha \log T}{(1-\gamma)^2 \Delta_{\min}^2} + \frac{\alpha}{\alpha - 2}\right).$$

However, under stability, there exist $(\mathcal{A}, \mu)$ and $a$ such that $\bar{R}(a; \mathcal{M}) = \Omega(\mathrm{T})$.

# Proportional cost

$\gamma \in [0,1]$

$$\mathcal{C}(a, a'; \psi) = \gamma\, \psi(a, a')$$
$$\mathcal{T}(a, a'; \psi) = 0$$

platform can optimize any function while preserving agent participation

**Corollary 6**. Under objective $F_t: \mathcal{W} \to \mathbb{R}$, $\arg\max_{M \in \mathcal{W}} F_t(M) \in S(V(\cdot,\cdot; \nu_t))$.

**Proposition 7**. Under $\gamma = 1$, applying the GS algorithm at every time step gives $\underline{R}(a; \mathcal{M}) \leq 0$. However, under stability, no guarantee on fairness or SW and there exist $(\mathcal{A}, \mu)$ and $a$ such that $\bar{R}(a; \mathcal{M}) = \Omega(\mathrm{T})$.

# Recap

GS algorithm at every $t$ → stability.

Have seen guarantee on low pessimal regret but not on optimal regret, fairness or SW.

Reassures pessimistic agents → at least reach worst-case performance under a true stable matching in log(T) time steps. But optimistic agents may be disappointed.

Implication? Agents continuing to use the platform ≠ agents participate happily.

Unhappy due to high regret or unfairness.

Suggests the platform may suffer if an alternate platform that offers higher agent payoffs arises.

*Is it possible to guarantee low regret, fairness, and high SW alongside stability?*

# Balanced transfer

$$\mathcal{C}(a, a'; \psi) = 0$$

$$\mathcal{T}(a, a'; \psi) = \frac{1}{2}(\psi(a', a) - \psi(a, a'))$$

Compensating for preference imbalance (i.e., bargaining)

**Theorem 9**. Under balanced transfers and pairwise-unique $\rho(a, a') = \frac{1}{2}\big(\mu(a, a') + \mu(a', a)\big)$, applying the GS algorithm at every time step.

$$\underline{R}(a; \mathcal{M}) = \bar{R}(a; \mathcal{M}) \leq \Delta^{*,\rho}_{\max}(a) N^2 L \left( \frac{8\sigma^2 \alpha \log T}{\left(\Delta^{\rho}_{\min}\right)^2} + \frac{\alpha}{\alpha - 2} \right).$$

Moreover, fairness and high social welfare are guaranteed.

# Balanced transfer

$$\mathcal{C}(a, a'; \psi) = 0$$

$$\mathcal{T}(a, a'; \psi) = \frac{1}{2}(\psi(a', a) - \psi(a, a'))$$

Implications?

Under balanced transfers, stability → low regret, fairness, and high social welfare for free.

Bargaining elegantly aligns local and global desiderata.

# Pricing

$$\mathcal{C}(p,u;\psi) = c_1 \qquad \mathcal{C}(p,u;\psi) = c_2$$

$$\mathcal{T}(p,u;\psi) = g(p;\psi) \qquad \mathcal{T}(u,p;\psi) = -g(p;\psi)$$

Ex. prices of goods

**Theorem 11**. If $|\mu(u,\cdot)| \leq B$ for all $u \in \mathcal{U}$, applying the GS algorithm at every time step gives:

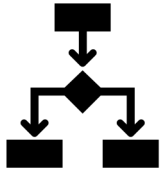
$$\underline{R}(a;\mathcal{M}) = \bar{R}(a;\mathcal{M}) \leq 2\Delta^{*,B}_{\max}(a)N^2L\left(\frac{8\sigma^2\alpha\log T}{(\Delta_{\min})^2} + \frac{\alpha}{\alpha-2}\right).$$

Moreover, fairness is guaranteed, but high social welfare is not.

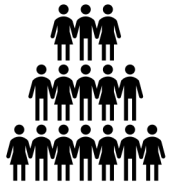
# Discussion

# Summary

Sequential decision-making



Learning agents



Uncertainty

Competition

Centralized matching market with bandit learners.

Liu, Mania & Jordan (2020) → stability + low pessimal regret.

Adding costs & transfers …

Allows us to model competition and exogeneous effects.

Makes it possible to simultaneously guarantee stability, low regret, fairness, and high social welfare.

# Key intuition

Four ingredients:

1. GS algorithm at every time step → stability.
2. Costs & transfers must give unique true stable matching.
3. Cost & transfer rules do not require knowledge of $\mu$.
4. Ensure costs & transfers do not interfere with learning.

# Thanks!