上海交通大学
约翰·霍普克罗夫特
计算机科学中心
John Hopcroft Center for Computer Science

Bandit Learning in Mechanism Design: Matching Markets and Beyond

Shuai Li
Associate professor at
Shanghai Jiao Tong University

Fang Kong
Ph.D. candidate student at
Shanghai Jiao Tong University

Outline

- Part 1: Two-sided matching markets 8:30-9:15
- Part 2: Multi-armed bandits 9:15-10:00
- Break: 10:00-10:30
- Part 3: Bandit algorithms in matching markets 10:30-11:30
- Part 4: Beyond matching markets 11:30-12:30

Part 1: Two-sided Matching Markets

Shuai Li, Fang Kong

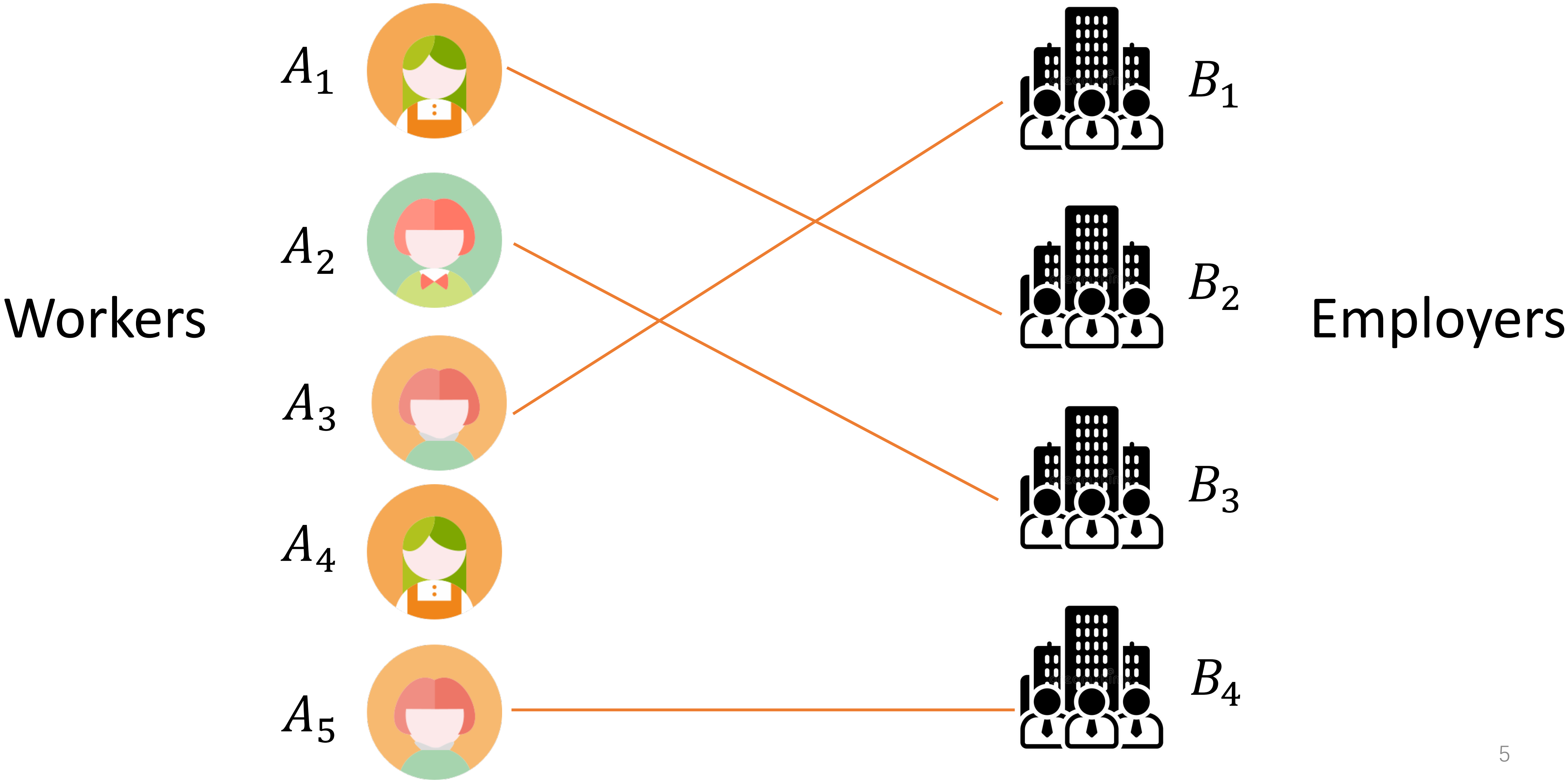
Shanghai Jiao Tong University

Matching markets

- Talent cultivation (school admissions, student internships)
- Task allocation (crowdsourcing assignments, domestic services)
- Resource distribution (housing allocation, organ donation allocation)

https://www.freepik.com; https://twitter.com/IslingtonBC/status/1623340900725272578

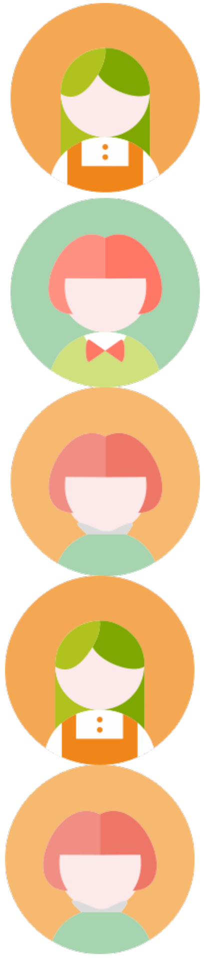
Matching market has two sides



Both sides have preferences over the other side

Worker side

Based on payment or prior familiarity of the task



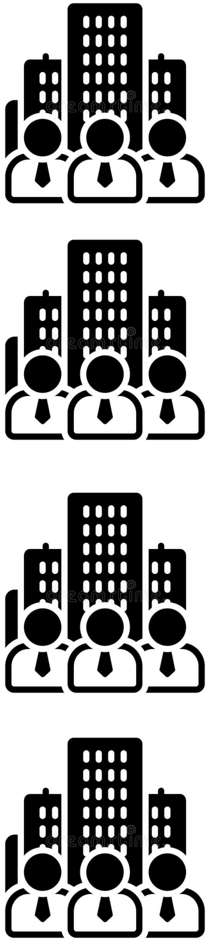
A_1 : $B_2 > B_3 > B_1 > B_4$

A_2 : $B_1 > B_2 > B_3 > B_4$

A_3 : $B_3 > B_1 > B_2 > B_4$

A_4 : $B_1 > B_2 > B_3 > B_4$

A_5 : $B_1 > B_2 > B_3 > B_4$



B_1

B_2

B_3

B_4

Both sides have preferences over the other side

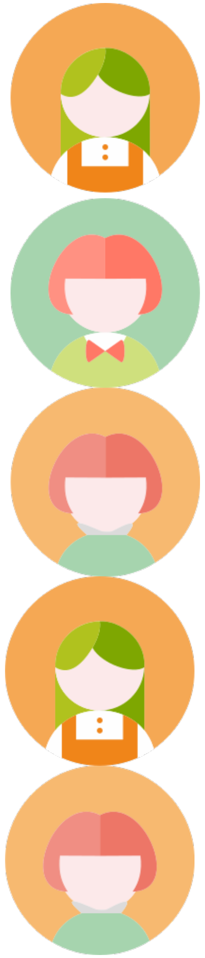
A_1

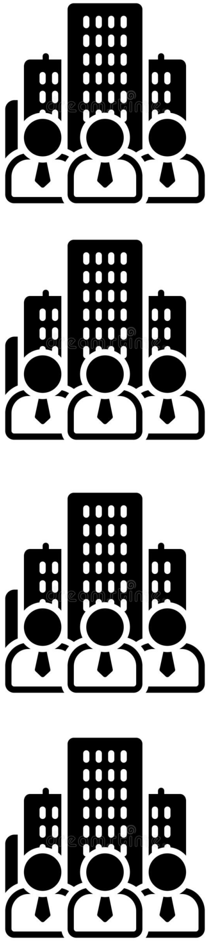
A_2

A_3

A_4

A_5





$B_1 \ : A_1 > A_2 > A_3 > A_4 > A_5$

$B_2 \ : A_2 > A_1 > A_4 > A_3 > A_5$

$B_3 \ : A_3 > A_1 > A_2 > A_5 > A_4$

$B_4 \ : A_4 > A_5 > A_1 > A_2 > A_3$

Employer side

Based on the skill levels of workers

A case study: Medical interns [Roth (1984)]

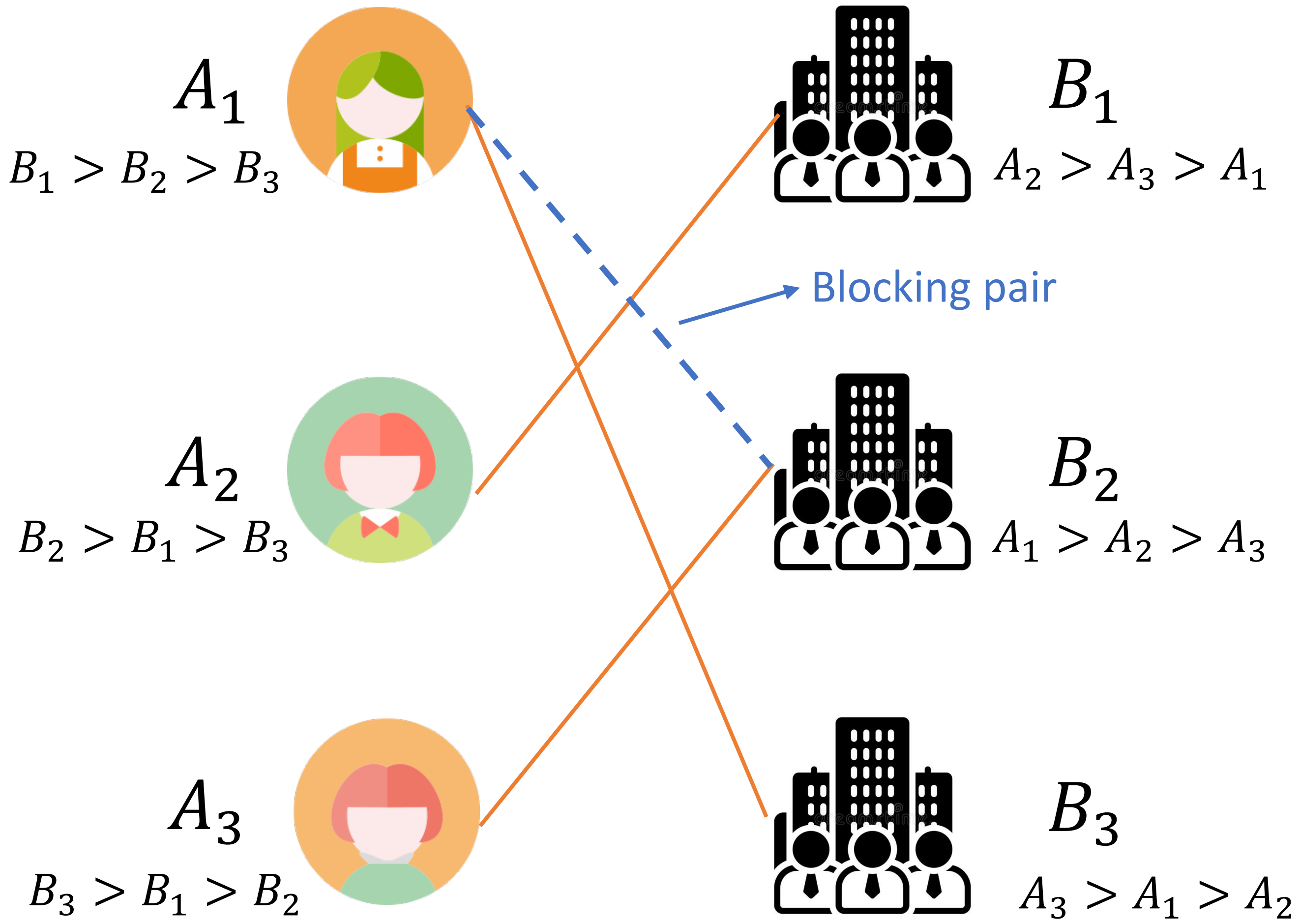
- Hospital side
  - Internship has relatively low cost
- Student side
  - closely engage with clinical medicine through internships

- Historical practice
  - Medical schools first publish students’ grade ranking
  - Then hospitals start signing internship agreements with students
- How to match?

Medical interns (cont.)

- Bad case
  - Student s_1
    - Receives offer from h_2 but knows he is on the waiting list of h_1
    - Wishes to wait for h_1
    - If s_1 is forced to accept h_2 and then h_1 sends an invitation?
  - Hospital h_2
    - Rejected by s_1 at the last moment
    - Students on the waiting list have already accepted other offers
- Important to guarantee stability

Stable matching



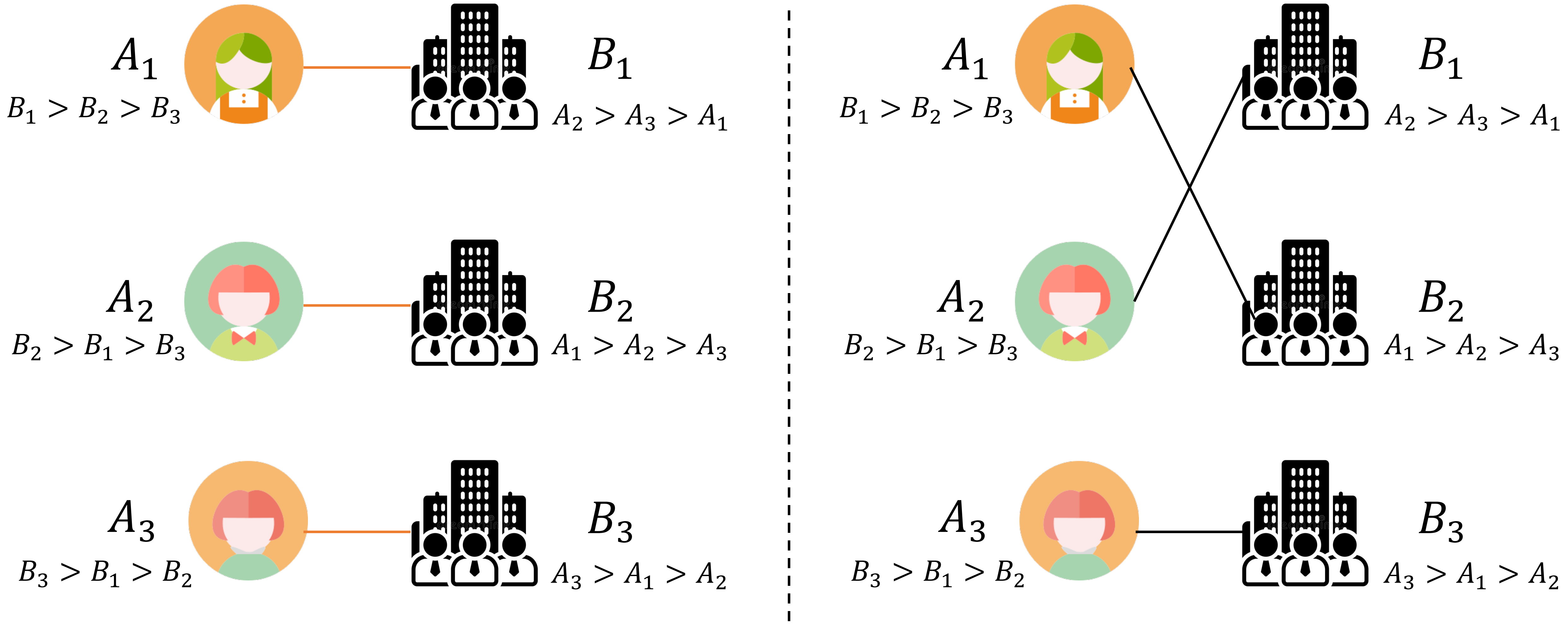
Participants have no incentive to abandon their current partner,

i.e.,

no blocking pair such that they both preferred to be matched with each other than their current partner

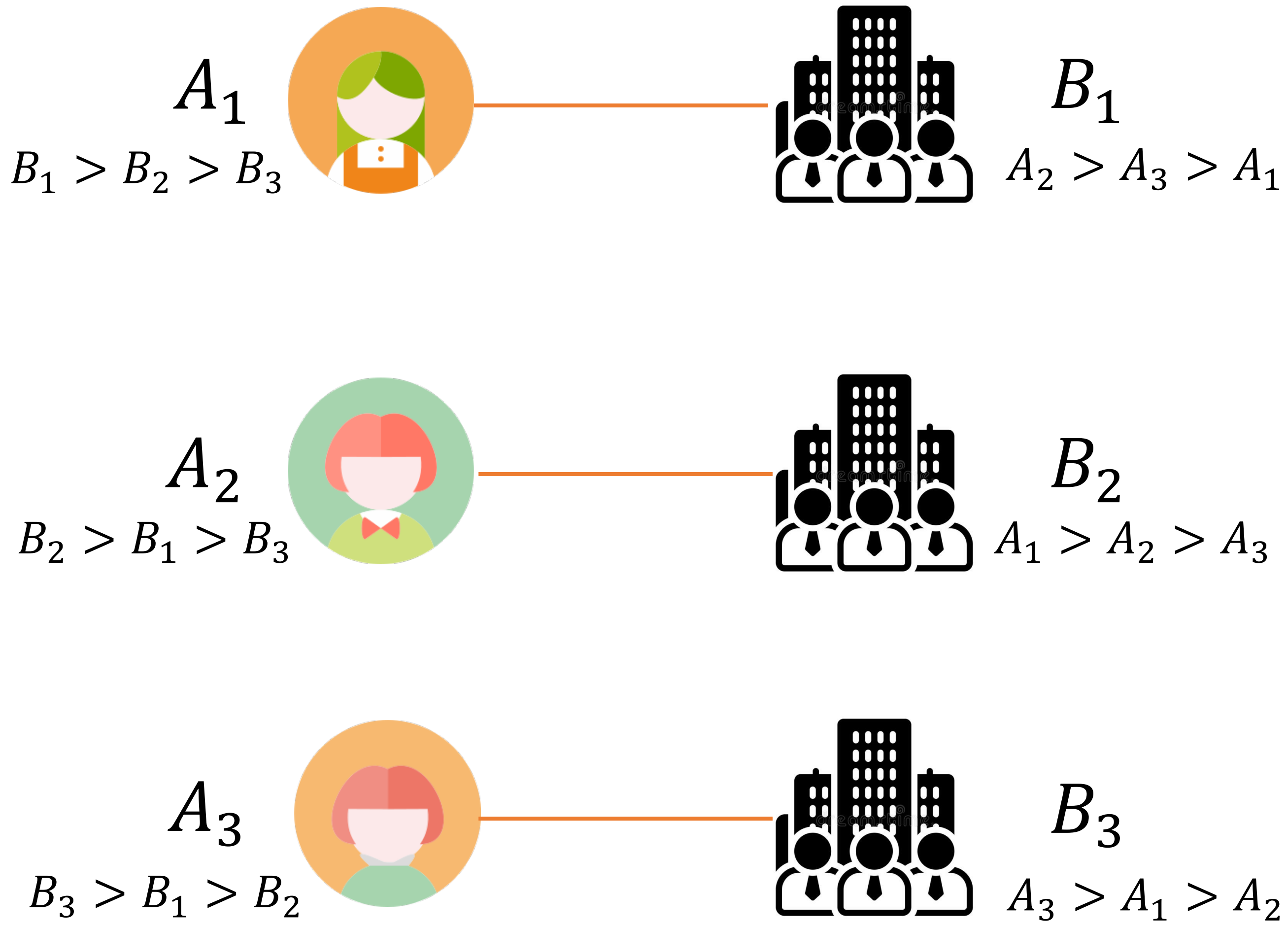
Alvin E. Roth and Lloyd S. Shapley jointly won the Nobel Prize in 2012 for their contributions to stable matching theory.

May be more than one stable matchings



$m_1 = \{(A_1, B_1), (A_2, B_2), (A_3, B_3)\}$ $m_2 = \{(A_1, B_2), (A_2, B_1), (A_3, B_3)\}$

A-side optimal stable matching[1]

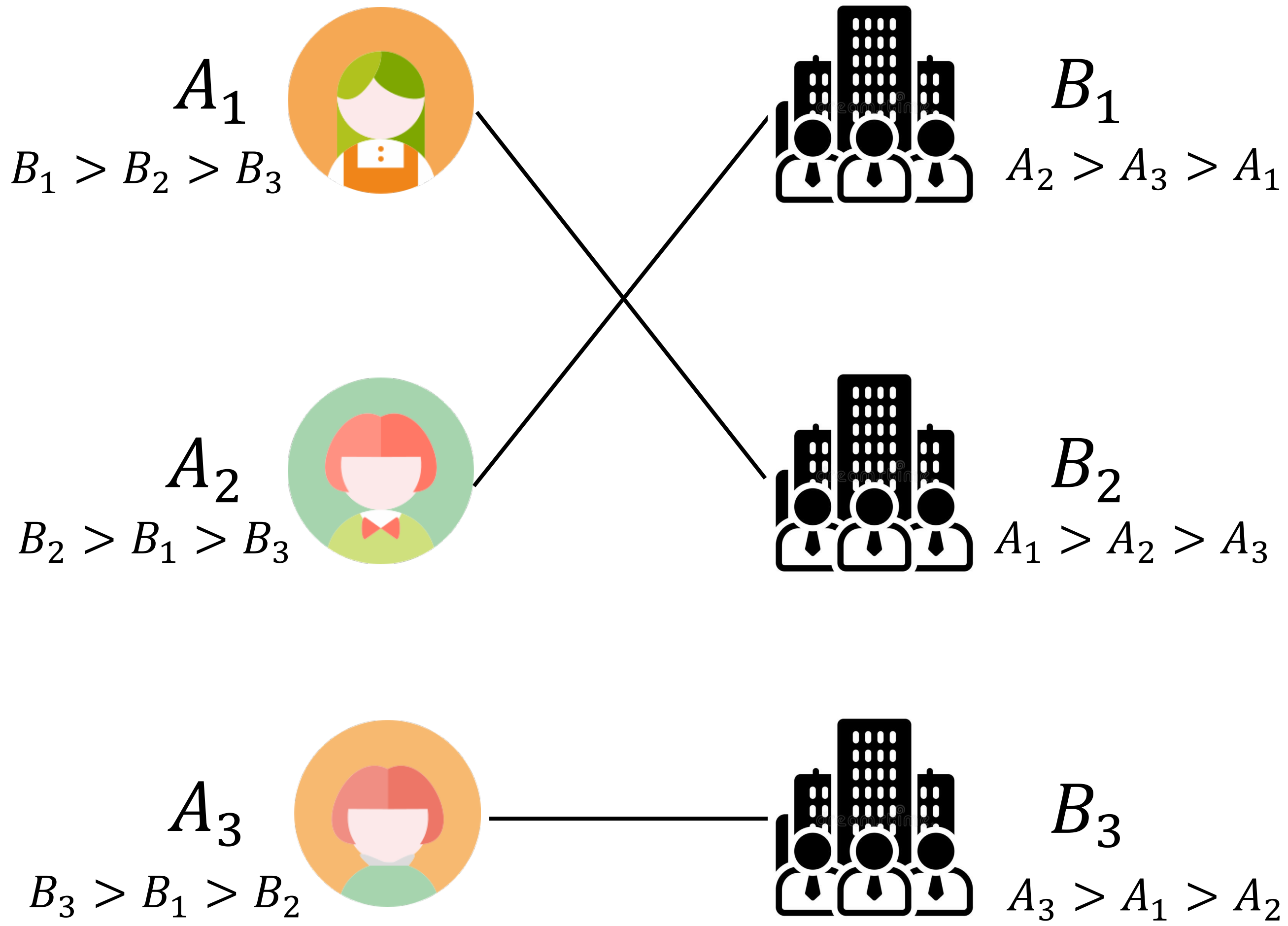


Each agent on A-side is matched with the most preferred partner among all stable matchings

$$m_1 = \{(A_1, B_1), (A_2, B_2), (A_3, B_3)\}$$

[1] The existence is proved by Gale and Shapley (1962).

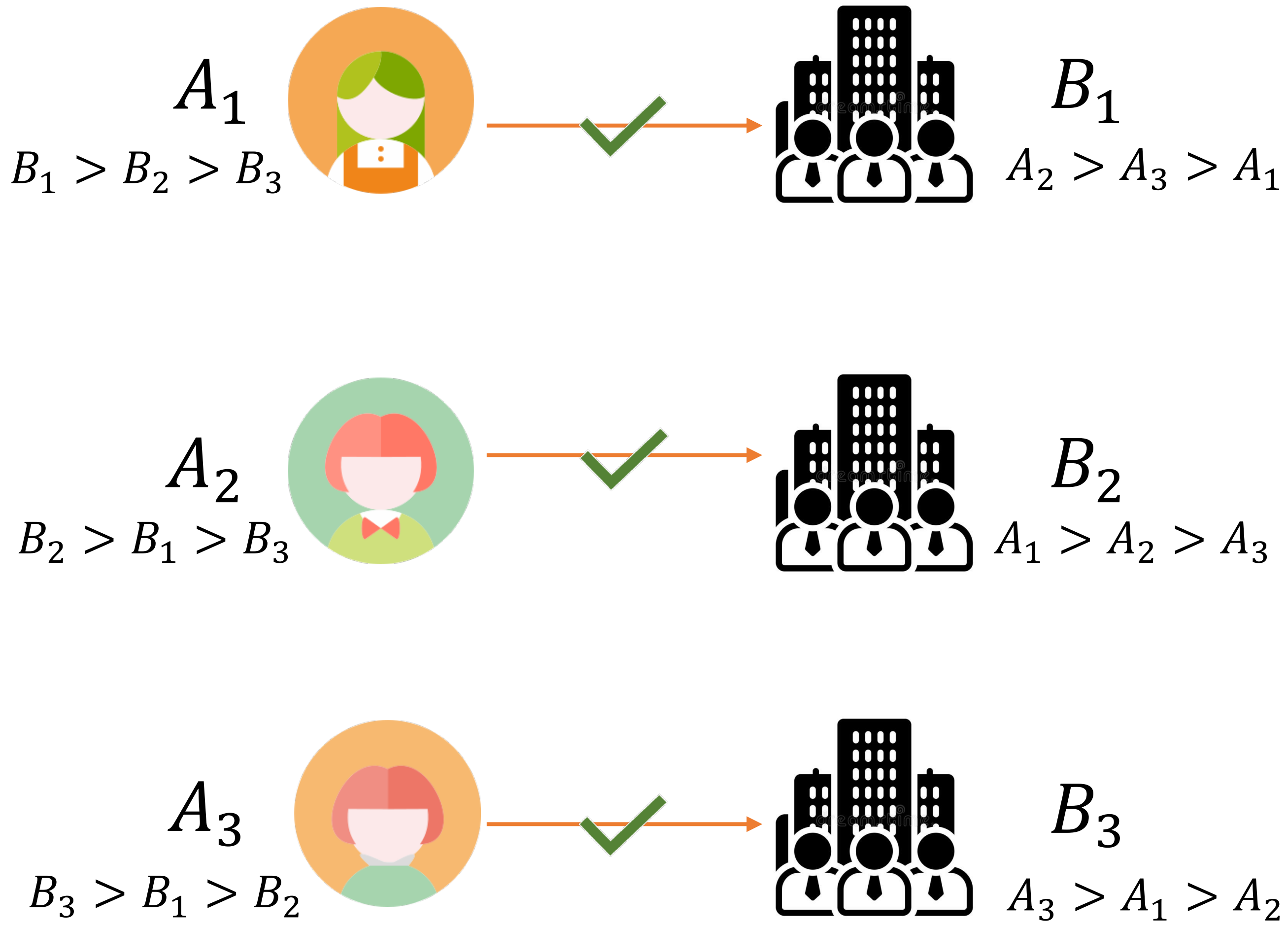
A-side pessimal stable matching



Each agent on A-side is matched with the least preferred partner among all stable matchings

$$m_2 = \{(A_1, B_2), (A_2, B_1), (A_3, B_3)\}$$

How to find a stable matching?

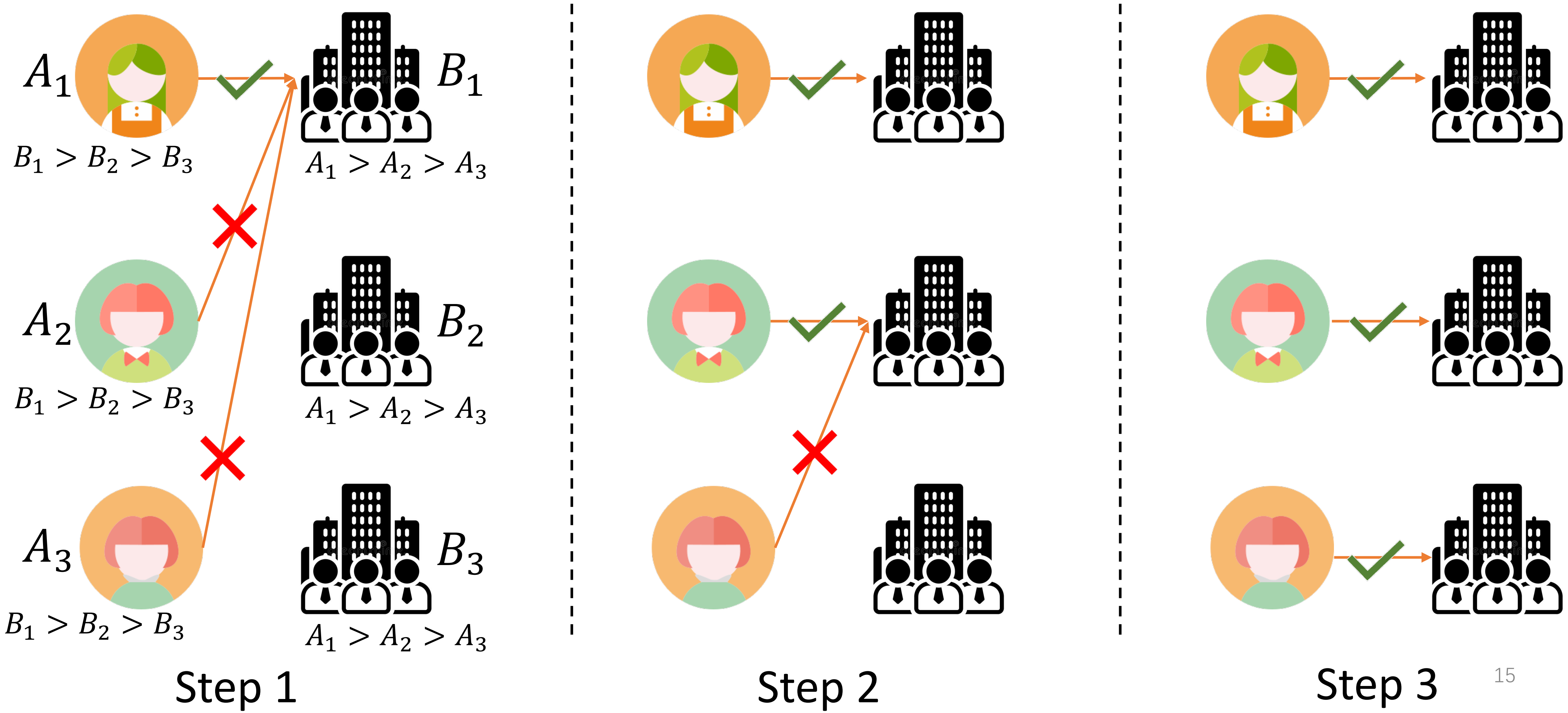


No rejection happens!

Gale-Shapley (GS) algorithm
[Gale and Shapley (1962)]

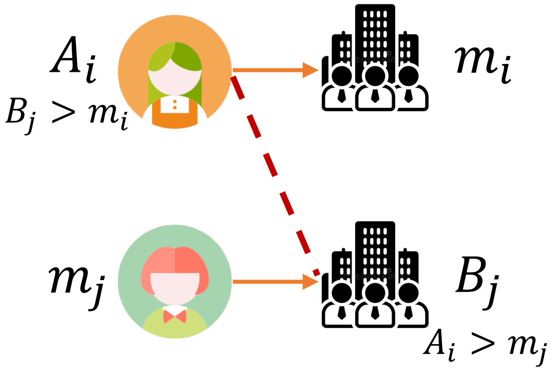
Agents on one side independently propose to agents on the other side according to their preference ranking until no rejection happens

Gale-Shapley (GS) algorithm: Case 2



GS properties: Stability

- The GS algorithm returns the stable matching



- Proof sketch
- Suppose there exists blocking pair (A_i, B_j) such that
  - A_i prefers B_j than its current partner m_i
  - B_j prefers A_i than its current partner m_j
- For A_i, it first proposes to B_j, but is rejected, then proposes to m_i
- This means that B_j must prefers m_j than A_i
- Contradiction!

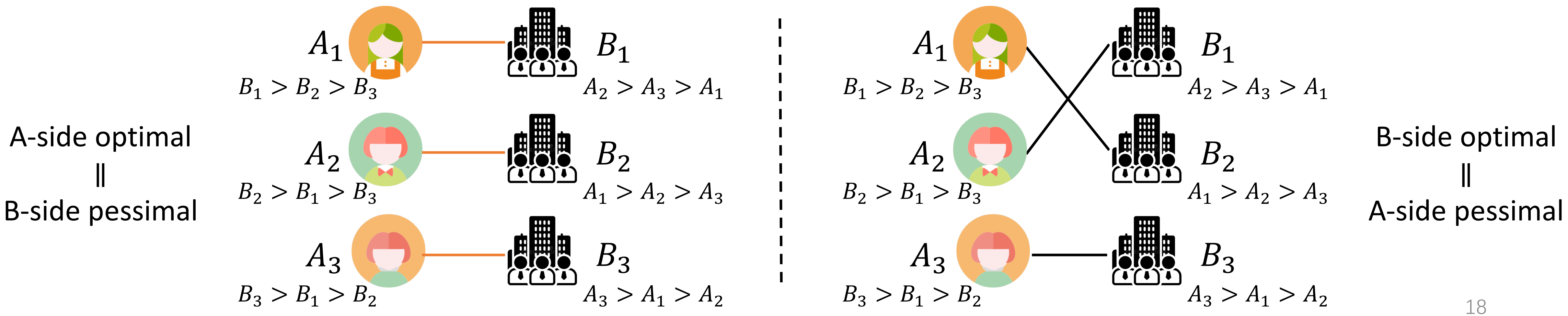
GS properties: Time complexity

- Each B-side agent can reject each A-side agent at most once
- At least one rejection happens at each step before stop
- $N =$ # {proposing-side agents}, $K =$ # {acceptance-side agents}
- $\Longrightarrow$ GS will stop in at most NK steps

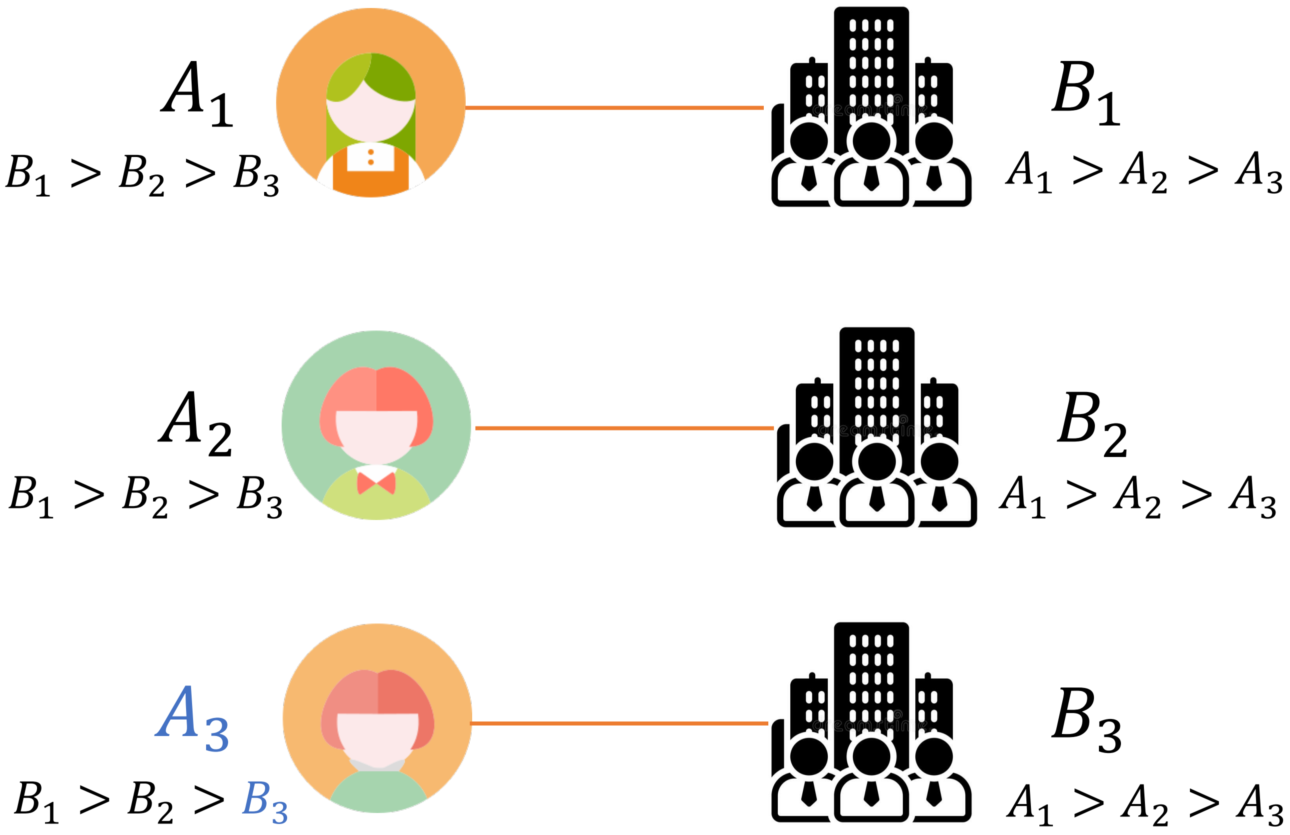
The time complexity can be improved as N^2 if $N \leq K$ [Kong and Li, 2023, arXiv version]

GS properties: Optimality

- Who proposes matters
  - Each proposing-side agent is happiest, matched with the most preferred partner among all stable matchings
  - Each acceptance-side agent is only matched with the least preferred partner among all stable matchings
  - A-side optimal stable matching = B-side pessimal stable matching



GS properties: Strategic behavior



Strategy-proof: Each participant is optimal to be truthful

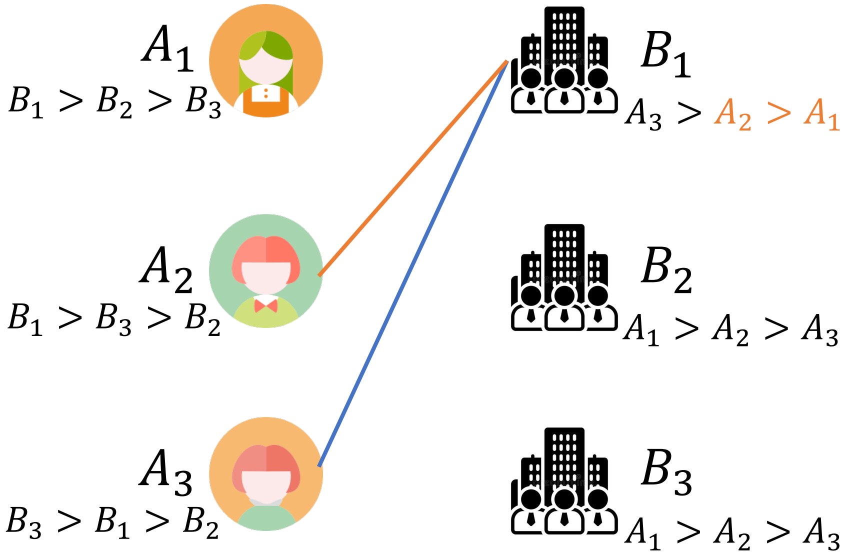
Deviating results in sub-optimal assignments

A_3 is matched with the least preferred partner B_3

Whether it is possible to match a better partner by misreporting?

GS properties: Strategic behavior (cont.)

- GS is strategy-proof for the proposing side [DF (1981); Roth (1982)]
  - Best for the proposing-side agents to report truthfully
- GS is not strategy-proof for the acceptance side



If B_1 reports truthfully:

Matching: $\{(A_1, B_2), (A_2, B_1), (A_3, B_3)\}$

If B_1 misreports preference $A_3 > A_1 > A_2$

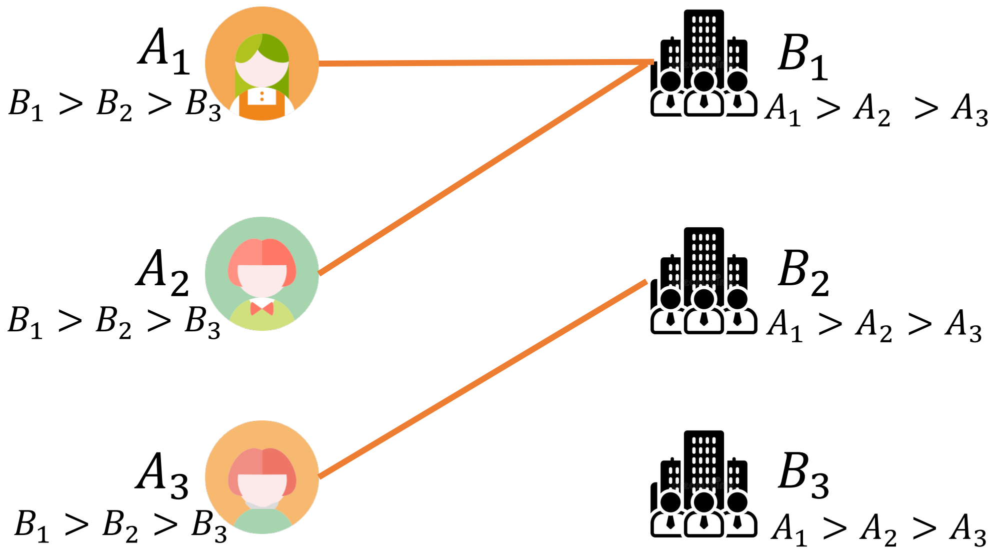
Matching[1]: $\{(A_1, B_1), (A_2, B_3), (A_3, B_1)\}$

B_1: $A_3 > A_2$, better partner!

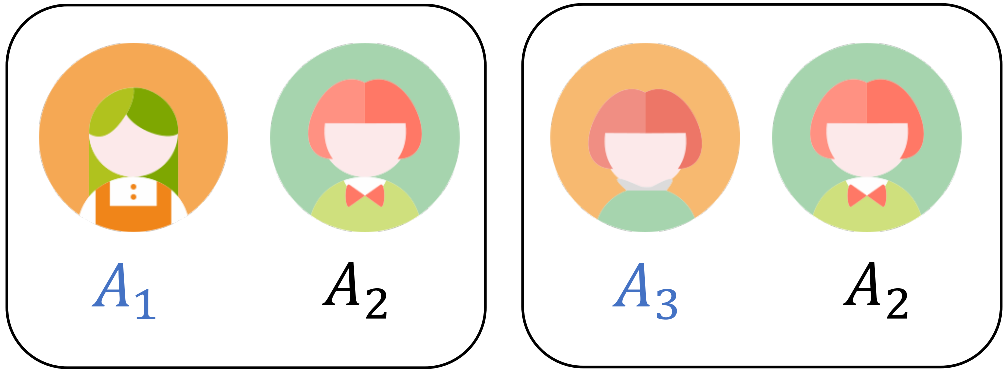
[1]Assume all of other agents report truthfully

Extension with sets: Many-to-one markets

- An agent may match more than one partner
  - Applications
    - An employer can hire a group of workers
    - A school can admit multiple students



Preferences over sets: Responsiveness



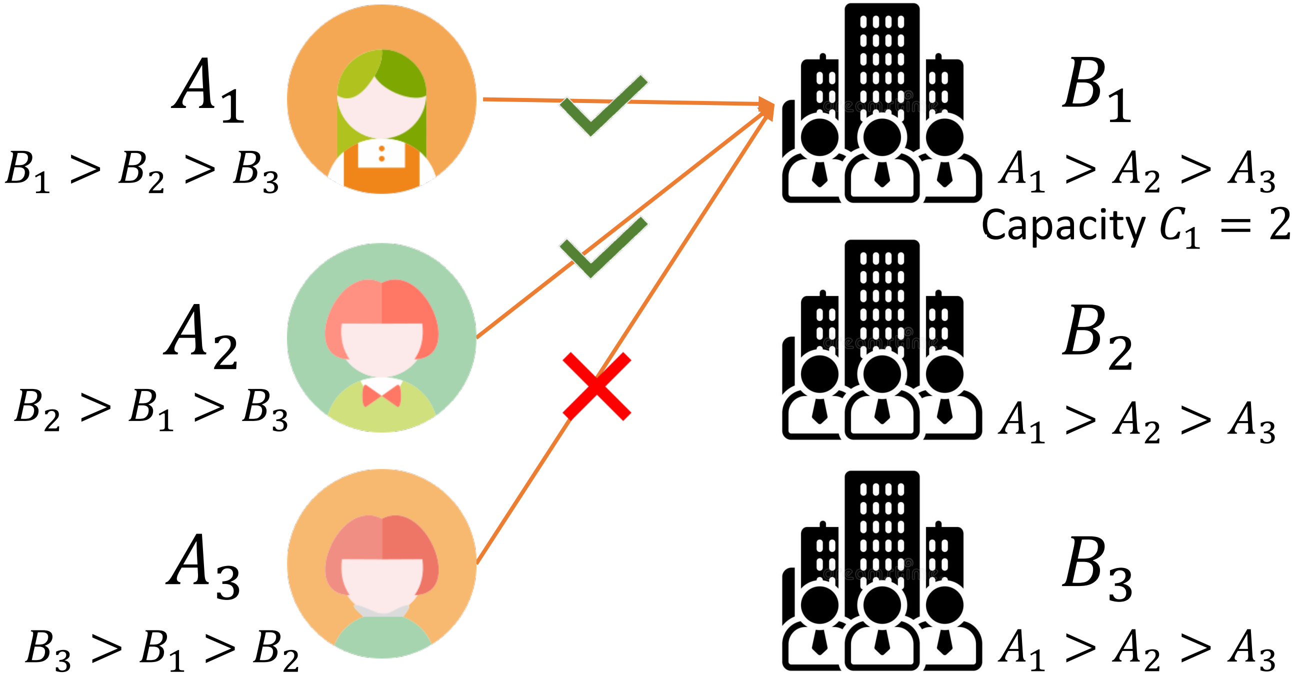
Set 1 Set 2

Group preferences are responsive to individual preferences:

Set 1 > Set 2 $\Leftrightarrow A_1 > A_3$

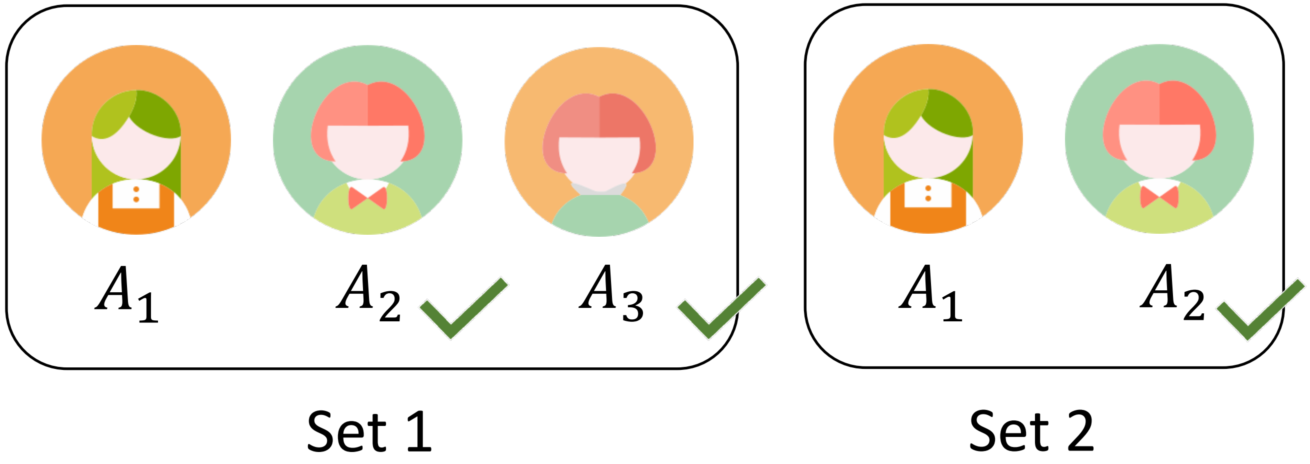
Common realization:

- Each agent B_j has a capacity C_j and preferences over individual partners
- Accept top C_j of them



Preferences over sets: Substitutability

- Agents have preferences over groups (instead of simply individuals)



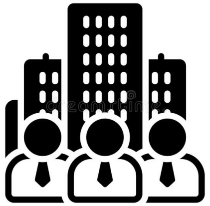
- Naturally holds under responsiveness
- One of the most generally known conditions to ensure the existence of a stable matching

- Regarding participants as substitutes over complementary:
  - Keeps accepting A_2 even if its colleague A_3 becomes unavailable

Substitutable preferences: An example

A_1

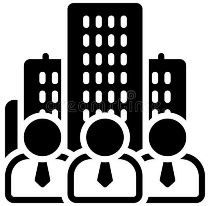
$B_2 > B_1 > B_3$



B_1

A_2

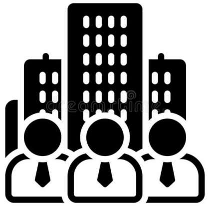
$B_2 > B_1 > B_3$



B_2

A_3

$B_2 > B_1 > B_3$



B_3

Agents' preference rankings:

B_1: $\{A_1, A_2\} > \{A_1, A_3\} > \{A_2, A_3\} > \{A_3\} > \{A_2\} > \{A_1\}$

B_2: $\{A_3\} > \emptyset$

B_3: $\{A_3\} > \emptyset$

When B_j is selected, it accepts the most preferred subset of agents proposing to B_j

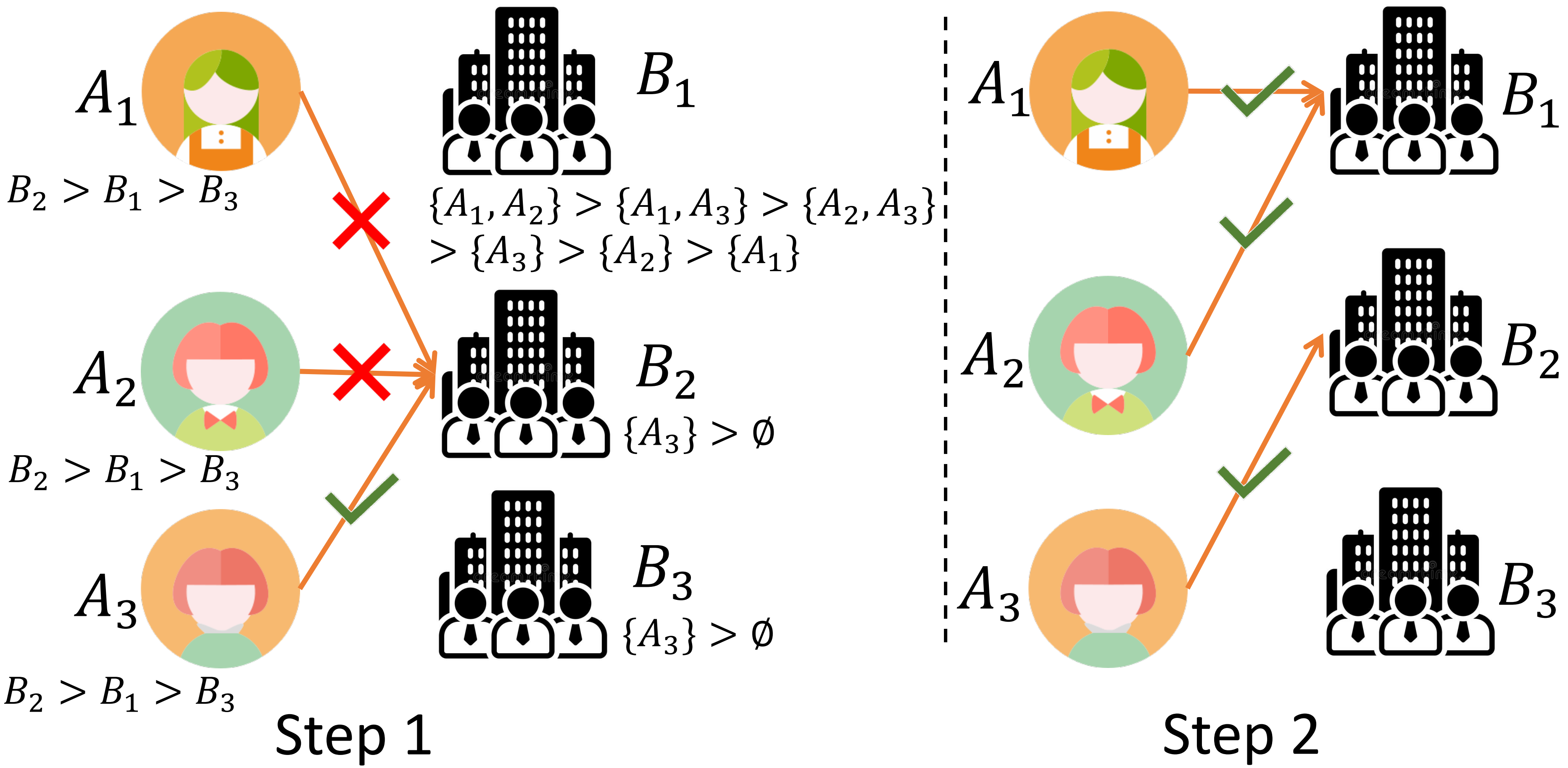
For example, for agent B_2:

If A_3 is in the proposing set, then B_2 accepts A_3;

Otherwise, B_2 accepts none of them

Deferred acceptance (DA) for substitutability

- The extension of GS under substitutability



The same properties as GS:

- Stability
- Time complexity
- Optimality
- Strategic behavior (When A-side propose)

[KC (1982); Roth (1984b); RS (1992)]

Summary of Part 1: Two-sided matching markets

- Introduction to matching markets
- Stable matching
- Gale-Shapley algorithm: Procedure and properties
  - Stability
  - Time complexity
  - Optimality
  - Strategic behavior
- Extension to many-to-one markets
  - Responsiveness
  - Substitutability
  - Deferred-acceptance algorithm

But agents usually have unknown preferences in practice

Can learn them from iterative interactions !

Part 2: Multi-armed Bandits

Shuai Li, Fang Kong

Shanghai Jiao Tong University

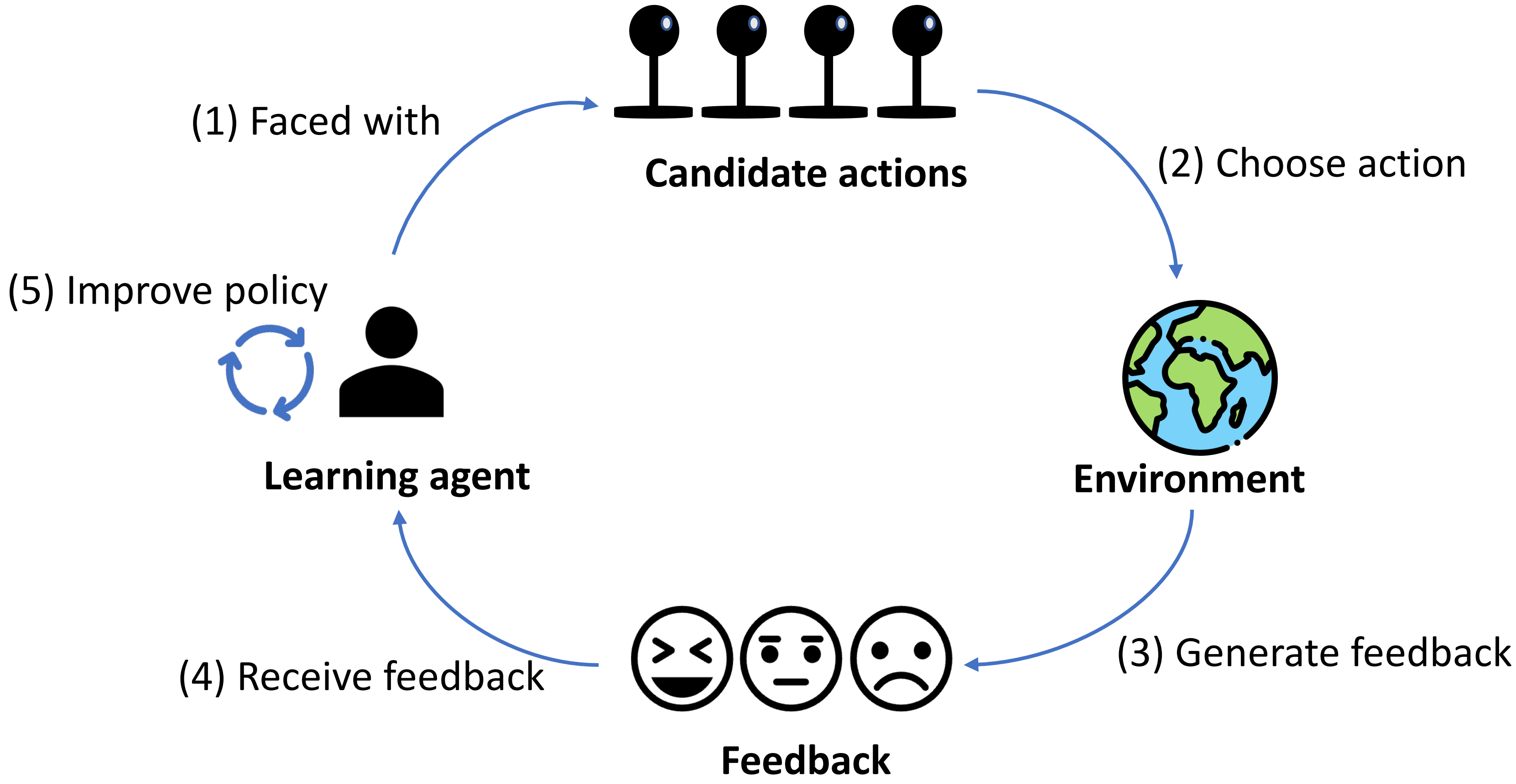
What are bandits? [Lattimore and Szepesvári, 2020]

Time	1	2	3	4	5	6	7	8	9	10
Arm 1	\$1	\$0			\$1	\$1	\$0			
Arm 2			\$1	\$0						

To accumulate as many rewards, which arm would you choose next?

Exploitation V.S. Exploration

Interactive machine learning

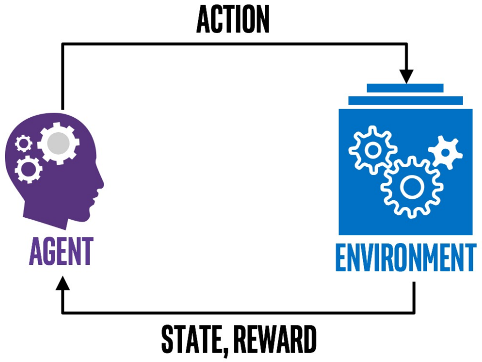


Provide insights for agents in matching markets to learn their unknown preferences through iterative interactions

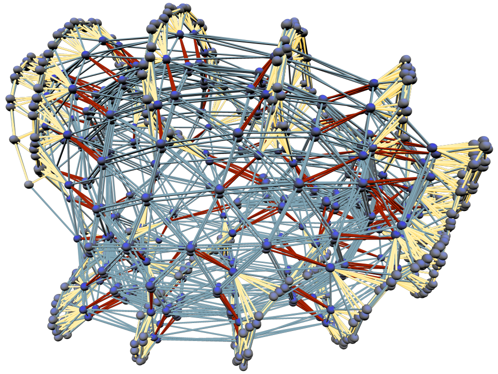
Applications

Recommendation systems
[Li et al., 2010]

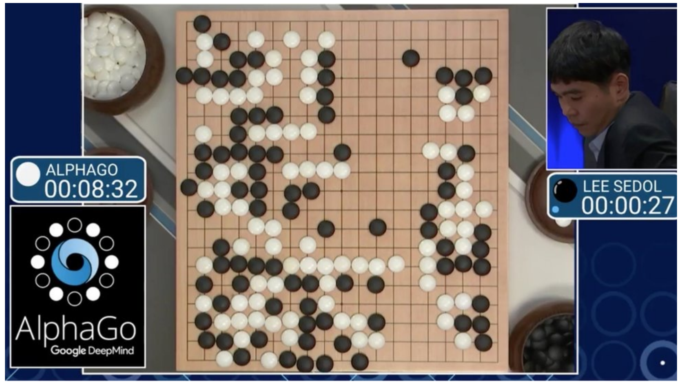
Advertisement placement
[Yu et al., 2016]



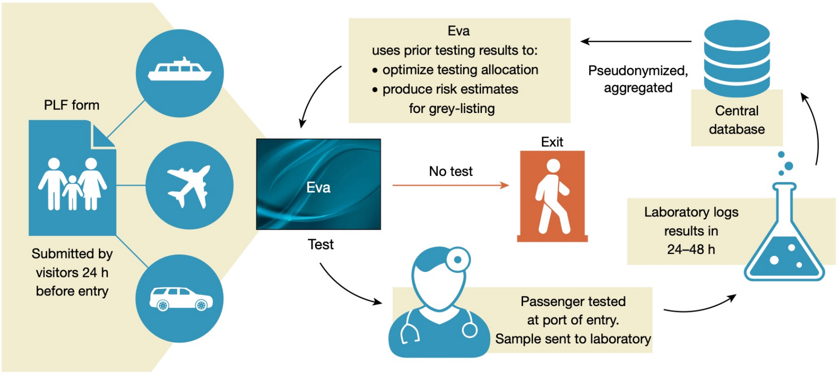
Key part of reinforcement learning
[Hu et al., 2018]



SAT solvers
[Liang et al., 2016]

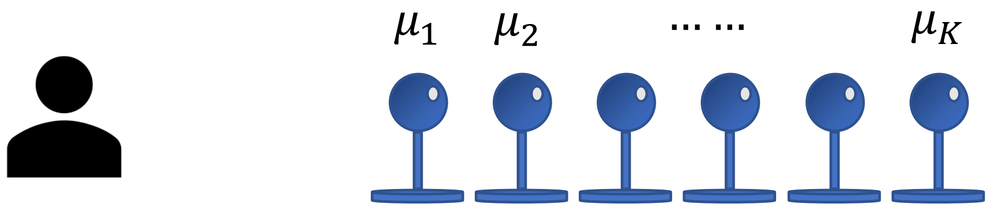


Monte-carlo Tree Search (MCTS) in AlphaGo
[Kocsis and Szepesvári, 2006; Silver et al., 2016]



Public health: COVID-19 border testing in Greece
[Bastani et al., 2021]

Multi-armed bandits (MAB)



- A player and K arms

Items, products, movies, companies, …

- Each arm a_j has an unknown reward distribution P_j with unknown mean μ_j

CTR, preference value, …

- In each round $t = 1,2,\ldots$:
  - The agent selects an arm $A_t \in \{1,2,\ldots,K\}$
  - Observes reward $X_t \sim P_{A_t}$

Click information, satisfaction, …

Assume P_j is supported on [0,1]

Objective

- Maximize the expected cumulative reward in T rounds

$$\mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

- Minimize the regret in T rounds
  - Denote $j^* \in \mathrm{argmax}_j\ \mu_j$ as the best arm

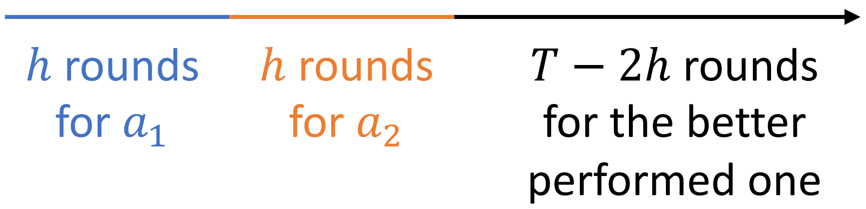
$$Reg(T) = T \cdot \mu_{j^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

Explore-then-commit (ETC) [Garivier et al., 2016]

- There are $K = 2$ arms (choices/plans/...)
- Suppose
  - $\mu_1 > \mu_2$
  - $\Delta = \mu_1 - \mu_2$

A/B testing

- Explore-then-commit (ETC) algorithm
  - Select each arm h times
  - Find the empirically best arm A
  - Choose $A_t = A$ for all remaining rounds



Explore-then-commit (cont.)

h rounds for a_1 | h rounds for a_2 | $T-2h$ rounds for the better performed one

- Regret analysis:

$$Reg(T) = T\cdot\mu_1 - \mathbb{E}\left[\sum_{t=1}^{T}\mu_{A_t}\right]$$

Sample mean

$$= h\Delta + (T-2h)\cdot\Delta\cdot\mathbb{P}(\hat{\mu}_1 < \hat{\mu}_2)$$

$$= h\Delta + (T-2h)\cdot\Delta\cdot\mathbb{P}\big((\hat{\mu}_2-\mu_2)-(\hat{\mu}_1-\mu_1) > \Delta\big)$$

$$\le \underbrace{h\Delta}_{\text{Exploration}} + \underbrace{T\cdot\Delta\cdot\exp\left(-\frac{h\Delta^2}{4}\right)}_{\text{Exploitation}}$$

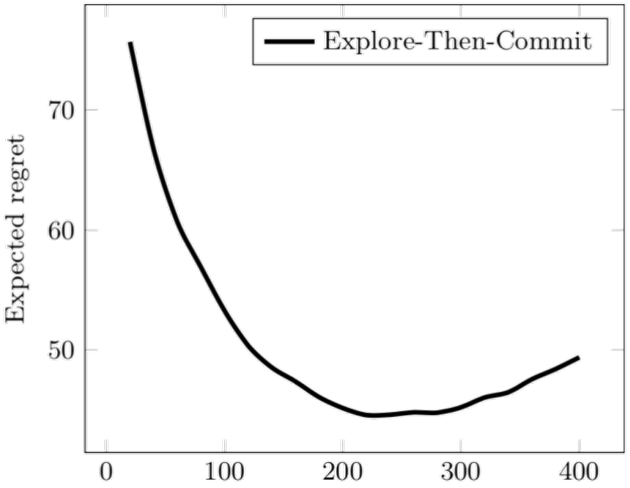
Hoeffding's inequality

$$\le O\left(\frac{\log T}{\Delta}\right)$$

Choose $h = \left\lceil \frac{4}{\Delta^2}\log\left(\frac{T\Delta^2}{4}\right)\right\rceil$

require the knowledge of Δ

- $Reg(T) = \Omega(T\Delta)$ if $h = 100$
- $Reg(T) = \Omega(T\Delta)$ if $h = T/10$



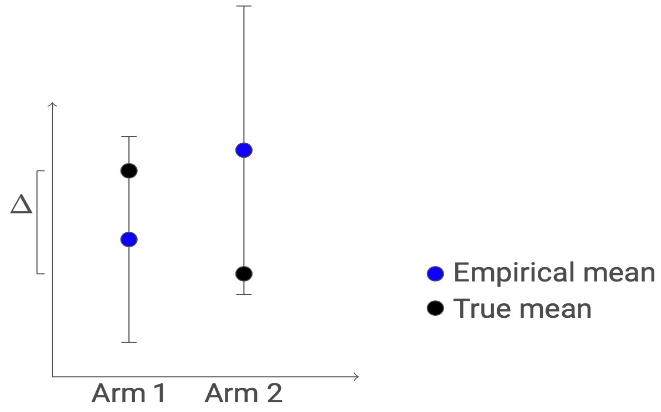
Only with the best choice of h the regret would be smallest

Upper confidence bound (UCB) [Auer et al., 2002]

- With high probability $\geq 1-\delta$ By Hoeffding's inequality

$$\mu_j \in \left[\hat{\mu}_j - \sqrt{\frac{\log 1/\delta}{T_j}}, \hat{\mu}_j + \sqrt{\frac{\log 1/\delta}{T_j}}\right]$$

Sample mean Number of selections of a_j



- Optimism: Believe arms have higher rewards, encourage exploration
  - The UCB value represents the reward estimates
- For each round t, select the arm

Upper confidence bound (UCB)

$$A(t) \in \mathrm{argmax}_{j\in[K]} \left\{ \hat{\mu}_j + \sqrt{\frac{\log 1/\delta}{T_j(t)}} \right\}$$

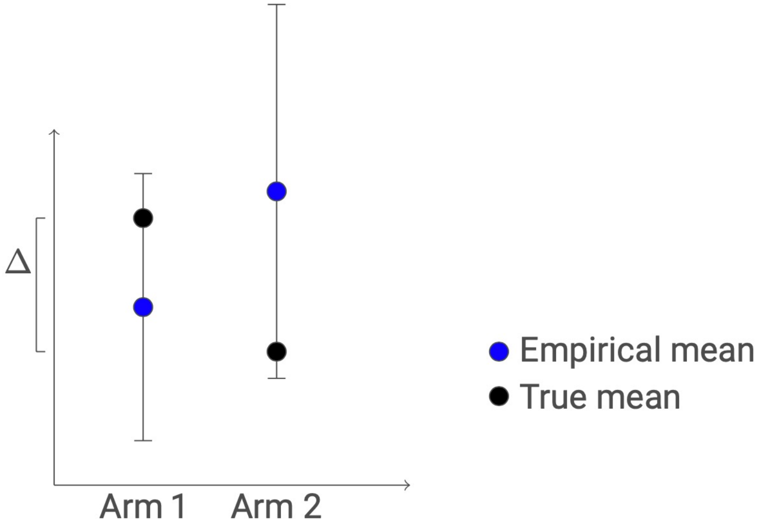
Exploitation Exploration

Upper confidence bound (UCB) (cont.)

- Assume arm a_1 is the best arm
- If sub-optimal arm a_j is selected
  - w/ high probability

    $$\mu_1 \le \text{UCB}_1 \le \text{UCB}_j \le \mu_j + 2\sqrt{\frac{\log 1/\delta}{T_j(t)}}$$

  - $\Longrightarrow 2\sqrt{\frac{\log 1/\delta}{T_j(t)}} \ge \Delta_j := \mu_1 - \mu_j$
  - $\Longrightarrow T_j(t) \le O\left(\frac{\log 1/\delta}{\Delta_j^2}\right)$



Can choose δ adaptive to time t

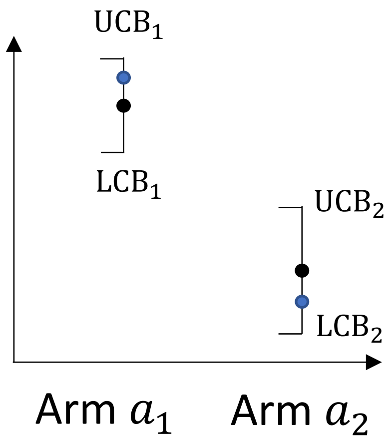
- By choosing $\delta = 1/T$, cumulative regret:

$$O\left(\sum_{j\ne 1} \frac{\log T}{\Delta_j^2} \cdot \Delta_j\right) = O(K \log T / \Delta)$$

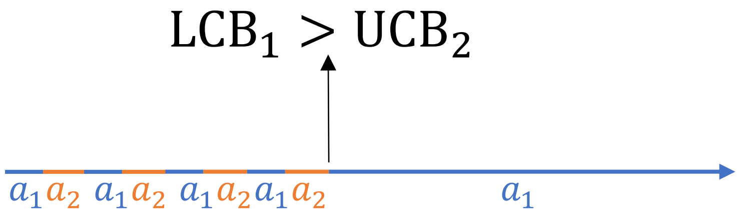
$\Delta := \min_{j\ne 1} \Delta_j$
Without knowing Δ

Improve ETC: Elimination [Audibert and Bubeck, 2010]

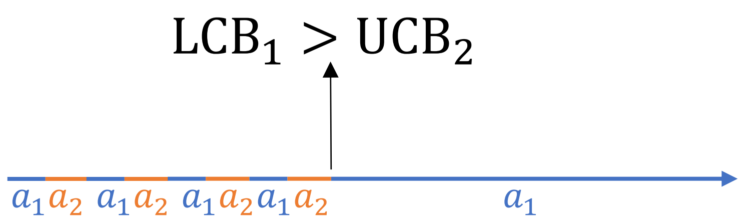
- Use confidence bound idea to remove requirement of Δ in ETC
- Recall that with high probability $\geq 1-\delta$
  - $\mu_j \in \left[\hat{\mu}_j - \sqrt{\frac{\log 1/\delta}{T_j}}, \hat{\mu}_j + \sqrt{\frac{\log 1/\delta}{T_j}}\right]$
  - Once $\mathrm{LCB}_1 > \mathrm{UCB}_2$ (disjoint confidence intervals)
    - Believes arm a_1 has higher rewards



- Uniformly select all active arms
- Once an arm is determined to be sub-optimal (its UCB is smaller than someone’ LCB values)
  - Delete it from the active set



Improve ETC: Elimination (cont.)



- Assume arm a_1 is the best arm
- If sub-optimal arm a_j is selected
  - w/ high probability

$$\mu_1 - 2\sqrt{\frac{\log 1/\delta}{T_1(t)}} \le \mathrm{LCB}_1 \le \mathrm{UCB}_j \le \mu_j + 2\sqrt{\frac{\log 1/\delta}{T_j(t)}}$$

  - $\Rightarrow \Delta \le 4\sqrt{\frac{\log 1/\delta}{\min\{T_1(t), T_j(t)\}}}$

Uniform exploration

  - $\Rightarrow T_j(t) \le O\left(\frac{\log 1/\delta}{\Delta^2}\right)$
- By choosing $\delta = 1/T$, cumulative regret:

$$O\left(\sum_{j \ne 1} \frac{\log T}{\Delta_j^2} \cdot \Delta_j\right) = O(K \log T/\Delta)$$

Without knowing Δ

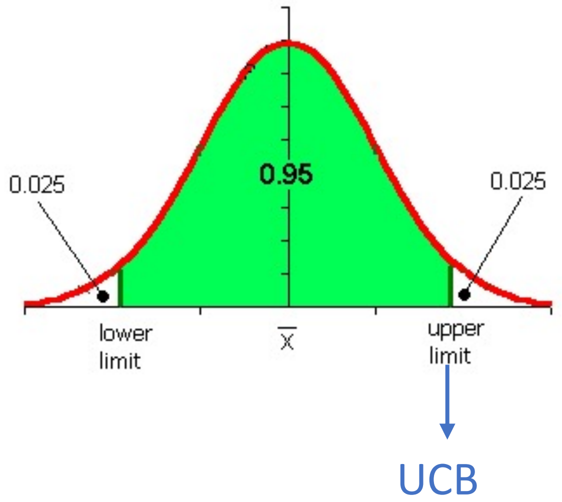
Thompson sampling (TS) [Agrawal and Goyal, 2013]

- Assume each arm has prior $\mathrm{Gaussian}(0,1)$
- Sample an estimate $\tilde{\mu}_j$ from the posterior distribution

$$\tilde{\mu}_j \sim \mathrm{Gaussian}\left(\hat{\mu}_j, \frac{1}{1+T_j(t)}\right)$$

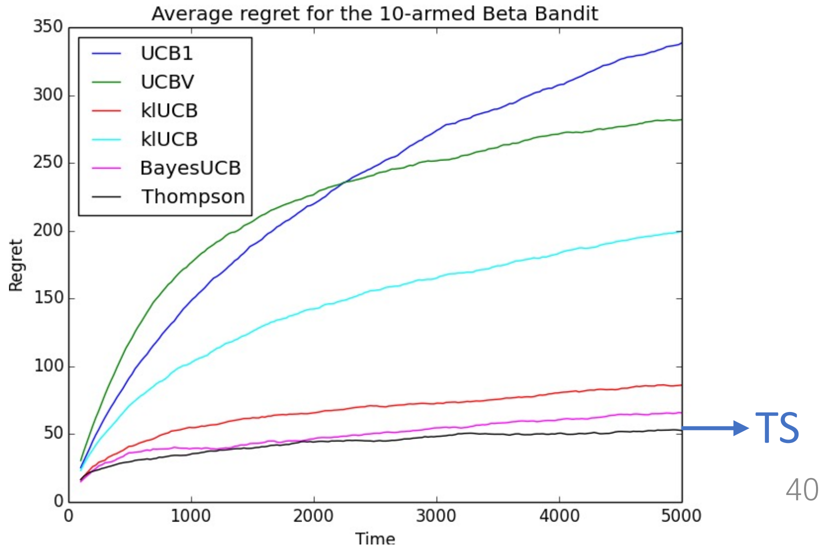
Exploitation

Exploration



- Select the arm $A(t) \in \mathrm{argmax}_{j\in[K]} \tilde{\mu}_j$

- Also have $O(K \log T/\Delta)$ regret
- Usually outperforms UCB



Lower bound [Lai and Robbins, 1985]

- An algorithm is consistent on class of bandits $\mathcal{E}$ if $Reg(T) = o(T)$ for all bandits in $\mathcal{E}$
- If the algorithm is consistent, then

$$\liminf_{T\to\infty} \frac{Reg(T)}{\log T} \geq \Omega\left(\sum_{j\neq 1} \frac{1}{\Delta_j^2} \cdot \Delta_j\right) = \Omega\left(\sum_j \frac{1}{\Delta_j}\right)$$

- Intuition
  - To distinguish sub-optimal arm a_j from the optimal one, it needs to be observed $\Omega\left(\log T/\Delta_j^2\right)$ times

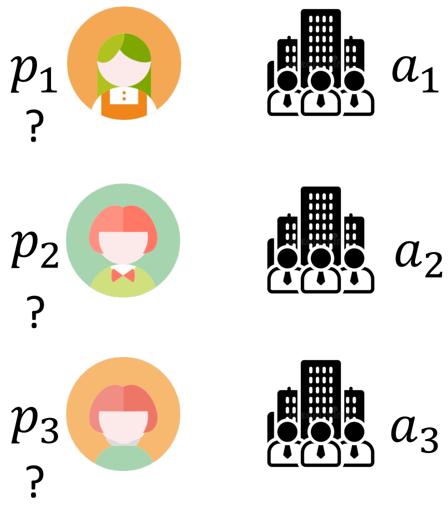
Bandit learning in matching markets [Liu et al., 2020]

- N players: $\mathcal{N} = \{p_1, p_2, \ldots, p_N\}$
- K arms: $\mathcal{K} = \{a_1, a_2, \ldots, a_K\}$
- $N \leq K$ to ensure players can be matched
- $\mu_{i,j} > 0$: (unknown) preference of player p_i towards arm a_j
- For each player p_i
  - $\{\mu_{i,j}\}_{j \in [K]}$ forms its preference ranking
  - For simplicity, the preference values of any player are distinct
- For each round t:
  - Player p_i selects arm $A_i(t)$
  - If p_i is accepted by $A_i(t)$: receive $X_{i,A_i(t)}(t)$ with

    Satisfaction over this matching experience

    $$\mathbb{E}\left[X_{i,A_i(t)}(t)\right] = \mu_{i,A_i(t)}$$
  - If p_i is rejected: receive $X_{i,A_i(t)}(t) = 0$

When would p_i be rejected?



For simplicity, assume arms know their preferences

Conflict resolution: One-to-one setting

- Each arm a_j has a preference ranking π_j
- $\pi_j(p_i)$: the position of p_i in the preference ranking of a_j
- $\pi_j(p_i) < \pi_j(p_{i'})$: a_j prefers p_i than $p_{i'}$

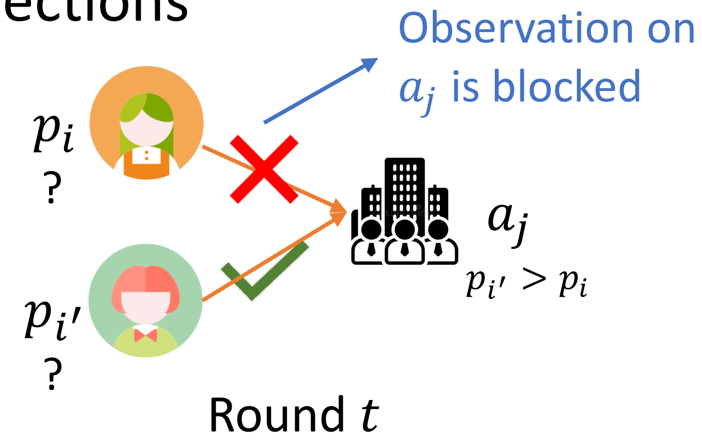
- At each round t, when multiple players select arm a_j
- a_j only accepts the most preferred one $p_i \in \mathrm{argmin}_{p_{i'}:A_{i'}(t)=a_j} \pi_j(p_{i'})$ and rejects others

Objective

- Minimize the stable regret
  - The player-optimal stable matching
$$\overline{m} = \{(i, \overline{m}_i) : i \in [N]\}$$
  - The player-optimal stable regret of player p_i is
$$\overline{Reg}_i(T) = T\mu_{i,\overline{m}_i} - \mathbb{E}\left[\sum_{t=1}^{T} X_{i,A_i(t)}(t)\right]$$
  - The player-pessimal stable regret $\underline{Reg}_i(T)$
    - Use the objective of the player-pessimal stable matching $\underline{m}$
- Guarantee strategy-proofness
  - Single player can not achieve $O(T)$ reward increase by deviating when others follow the algorithm

Challenge in matching markets

- Learning process: Other players will block observations
  - Once the player selects an arm based on its exploration-exploitation (EE) strategy, this arm may reject the player due to others' selections
  - The individual player's EE trade-off is interrupted



- Objective: Cannot maximize a single player's utility
  - Aim to find the optimal equilibrium of the market

How to control agents' blockings?

- Centralized
  - All participants submit their estimations to the platform
  - The platform computes an assignment
  - All players follow this assignment

- Decentralized
  - Each player independently computes the target arm
  - Necessary information to communicate:
    - common index of arms, matching outcomes in each round, etc.

Summary of Part 2: Multi-armed bandits

- Multi-armed bandits (MAB)
  - Applications
  - Explore-then-commit (ETC)
  - Upper confidence bound (UCB)
  - Successive elimination
  - Thompson sampling (TS)
  - Lower bound
- Bandit learning in matching markets
  - Setting
  - Challenge

上海交通大学
约翰·霍普克罗夫特计算机科学中心
John Hopcroft Center for Computer Science

Part 3: Bandit Algorithms in Matching Markets

Shuai Li, Fang Kong

Shanghai Jiao Tong University

Outline

- Centralized algorithms
  - ETC, UCB
  - The failure of UCB
- Decentralized algorithms
  - General markets
  - Markets with unique stable matching
  - Explore-then-GS (ETGS) strategies
- Lower bound
- Many-to-one markets
- Strategic behavior
  - Adaptive ETGS
- Other variants

Warm up: Centralized ETC [Liu et al., 2020]

- Input: An exploration budget h

Exploration | $t = hK$ | Exploitation

hK rounds: explore all arms in a round-robin manner

GS with estimated ranking

Remaining rounds: Follow GS's choice

- For round $t = 1,2, \dots,$
  - $t < hK$:
    - $A_i(t) = a_{(t+i) \bmod K}$ //No conflict
    - Update the corresponding rewards
  - $t = hK$:
    - Receive the estimated rankings $\hat{\rho}_i$
    - Using GS to compute the matching $m \coloneqq (m_i)_{i\in[N]}$ based on $(\hat{\rho}_i)_{i\in[N]}$
    - $A_i(t) = m_i$
  - $t > hK$
    - $A_i(t) = m_i$

Centralized ETC: Analysis

- If any player can estimate their preference ranking accurately
- Then the GS algorithm can output the player-optimal stable matching

- Define $\Delta_{i,j,j'} = \left|\mu_{i,j} - \mu_{i,j'}\right|$
- Further define $\Delta = \min_{i,j\neq j'} \Delta_{i,j,j'}$

Larger than 0 due to distinct preferences

- By choosing $h = \left\lceil \frac{4}{\Delta^2} \log\left(1 + \frac{TN\Delta^2}{4}\right) \right\rceil$, all players can estimate their ranking well w.h.p.
- The player-optimal stable regret satisfies

$$\overline{Reg}_i(T) = O(hK) = O\left(\frac{K \log T}{\Delta^2}\right)$$

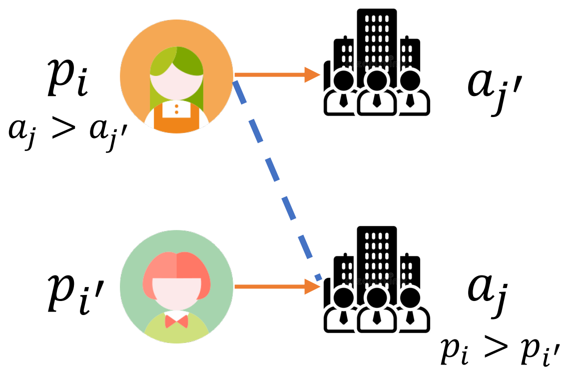
Needs to know Δ

Remark: Δ can be improved as the minimum gap between the player-optimal stable arm and the next preferred one among all players.

Centralized UCB [Liu et al., 2020]

- For round $t = 1,2, \dots,$
  - Each player estimates a UCB ranking towards all arms
  - The GS platform returns an assignment m_t under these UCB rankings
  - Each player selects the assigned arm

Centralized UCB: Analysis



- When is m_t unstable?
  - Exists blocking pair (p_i, a_j), p_i is actually matched with $a_{j'}$
  - What causes this blocking pair to appear?
    - p_i wrongly estimate UCB rankings: $\mathrm{UCB}_{i,j} < \mathrm{UCB}_{i,j'}$
- This scenario happens at most $O(\log T/\Delta^2)$ times

- Converge to the player-pessimal stable matching

$$Reg_i(T) = O\left(\frac{NK\log T}{\Delta^2}\right)$$

Do not require Δ, but can only achieve pessimal stable matching

Decentralized algorithms: UCB and TS

- Players select the arm based on the UCB ranking and TS estimates
- Coordinate players' selections to control conflicts

p_i

$A_i(t)$

w.p. $1-\lambda$ Available arm set: $S_i(t) =$ {arms that would accept p_i at $t-1$ given others selections}; the arm with the largest UCB/TS estimate in $S_i(t)$ →Exploration

w.p. λ Last-round choice $A_i(t-1)$ →Exploitation

Can successfully match the target arm w.p.

$$\kappa = (1-\lambda)\,\lambda^{N-1}$$

Regret type	Regret bound	Algorithm type	References
Player-pessimal stable matching	$O\left(\frac{N^5K^2\log^2 T}{\kappa^{N^4}\Delta^2}\right)$	UCB	[Liu et al., 2021]
		TS	[Kong et al., 2022]

Pessimal stable matching
Exponentially large term

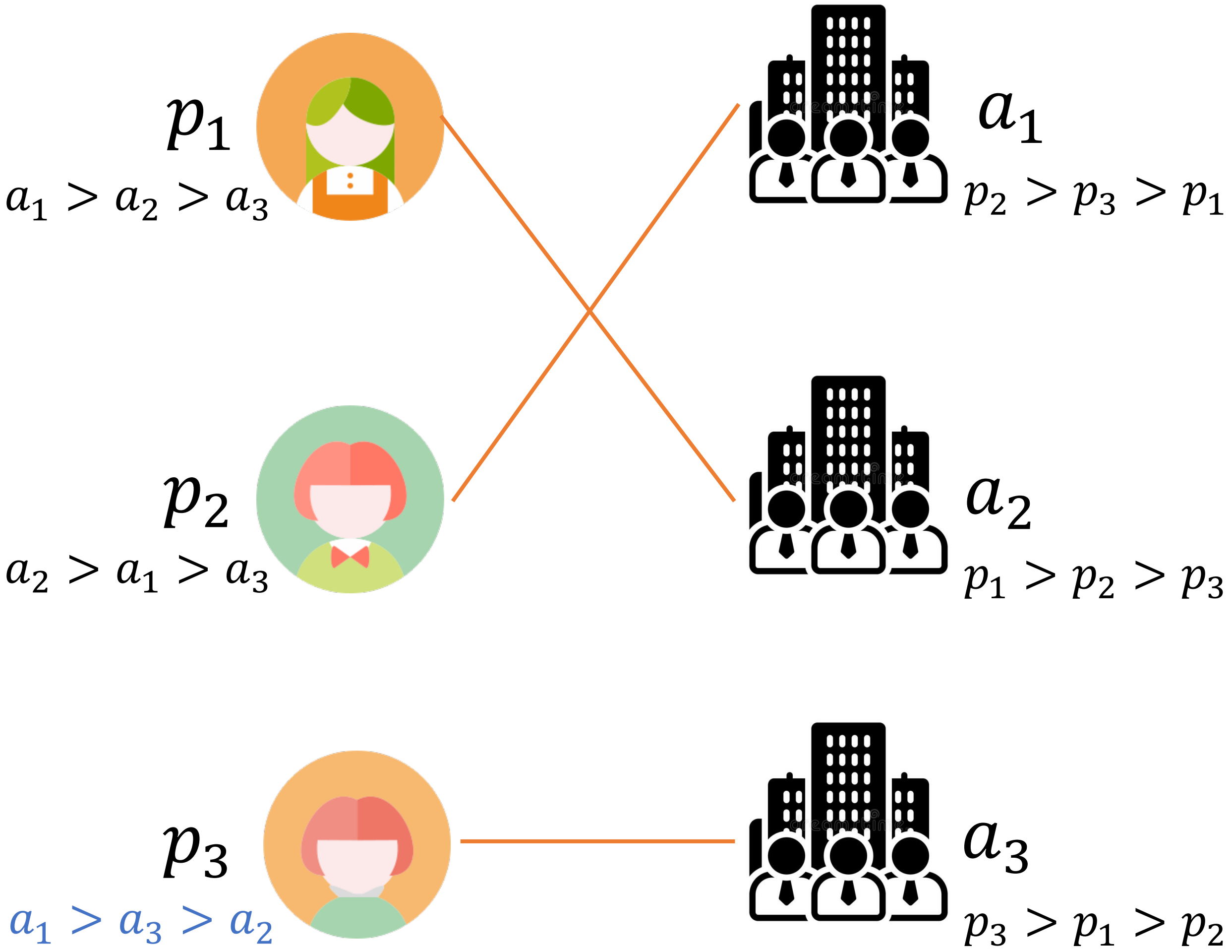
Unique stable matching

- When there is only one stable matching
  - Player-optimal stable matching = Player-pessimal stable matching
  - The blocking relationship becomes simpler

Regret type	Regret bound	Uniqueness condition	References
Unique stable matching	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Serial dictatorship	[Sankararaman et al., 2021]
		α-reducible condition	[Maheshwari et al., 2022]
		Uniqueness consistency (The most general)	[Basu et al., 2021]

Remark: Δ can be improved as the minimum gap between the player-optimal stable arm and the next preferred one among all players.

Why UCB fails to achieve player-optimality?



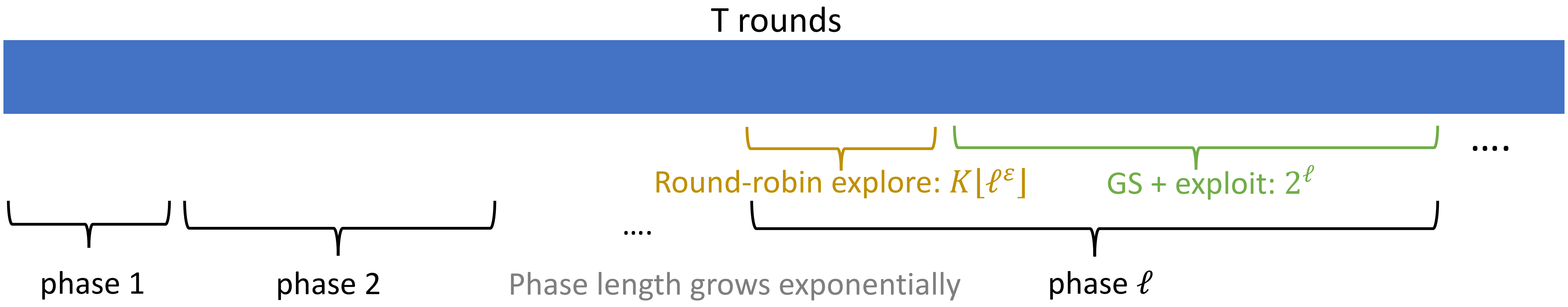
- When p_3 lacks exploration on a_1 with $a_1 > a_3 > a_2$ on UCB, GS outputs the matching[1] $(p_1, a_2), (p_2, a_1), (p_3, a_3)$
- p_3 fails to observe a_1
- UCB vectors do not help on exploration here
- Not consistent with the principle of *optimism in face of uncertainty*

1. When p_1 and p_2 submit the correct rankings

How to balance EE in a more appropriate way?

- Exploration-Exploitation trade-off
  - Exploitation goes though with correct rankings by following GS
  - Require enough exploration to estimate the correct rankings
- The UCB ranking does not guarantee enough exploration
- Perhaps design manually?
- To avoid other players’ block: Coordinate selections in a round-robin way

PhasedETC [Basu et al., 2021]



Implementation: Exploration

For round t:

$A_i(t) = a_{(t+i) \bmod K}$

$//p_1$: 1 2 3 1 2 3 1 2 3

$//p_2$: 2 3 1 2 3 1 2 3 1

Update the estimated ranking based on the received rewards

Implementation: GS + exploitation

//Follow GS to find the matching with the estimated ranking ρ based on the empirical mean

Initialize $s_i = 1$ for each player p_i

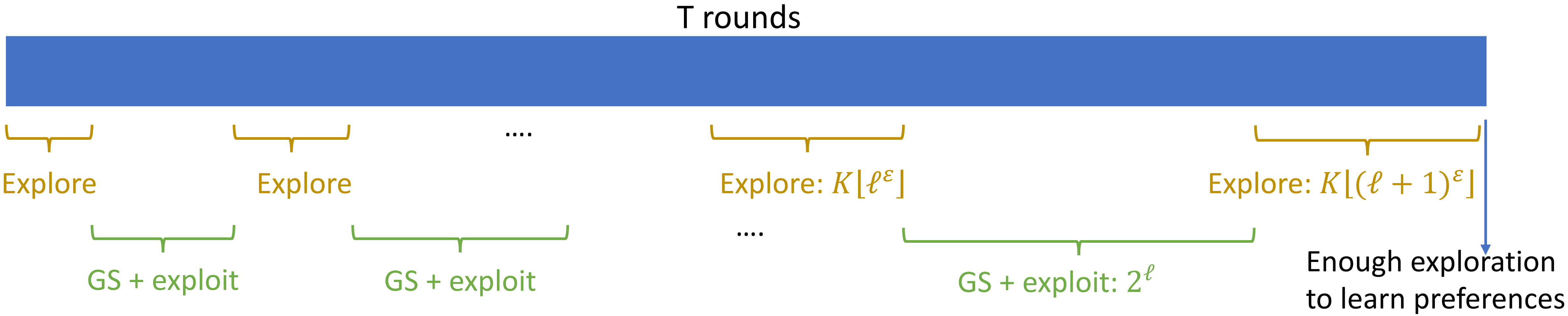
For round t:

$A_i(t) = a_{\rho_{s_i}}$

If p_i is not matched, $s_i = s_i + 1$

PhasedETC: Regret analysis

- Exploration is enough $\Longrightarrow$ Estimated ranking is correct $\Longrightarrow$ In the corresponding phase: GS returns the player-optimal stable matching



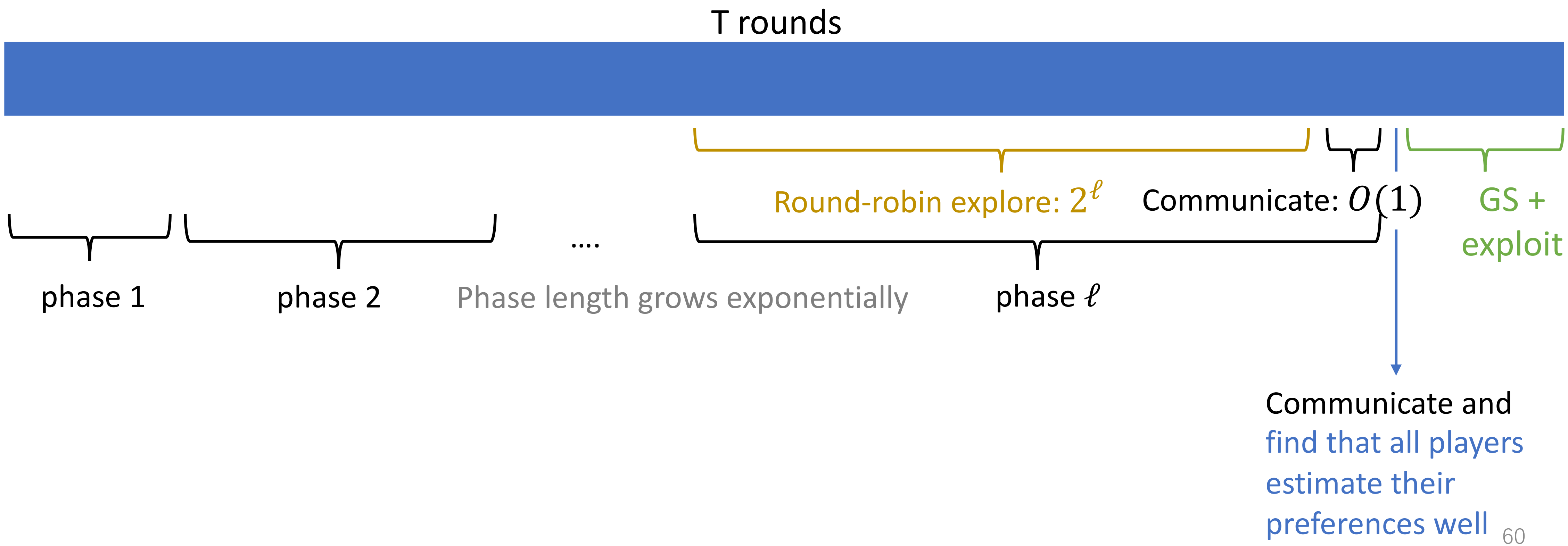
- The player-optimal regret comes from exploration and exploitation before estimating well

$$\overline{Reg}_i(T) = O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{1/\varepsilon}}\right)$$

Exponentially large term

Explore-then-GS (ETGS) [Kong and Li, 2023]

- Avoid unnecessary exploitation before estimating preferences well
  - Only when all players estimate well, enter GS + exploit

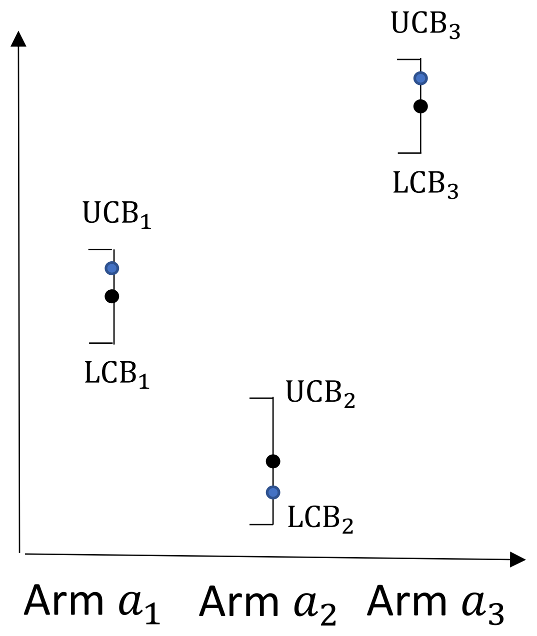


ETGS implementation: Communication

- At communication block: players determine whether all players estimate their preference rankings well

- For p_i
  - If there exists a ranking ρ_i over arms such that
    - The confidence intervals of all arms are disjoint
  - Note: this estimated ranking is accurate w.h.p.

- How to communicate with others?

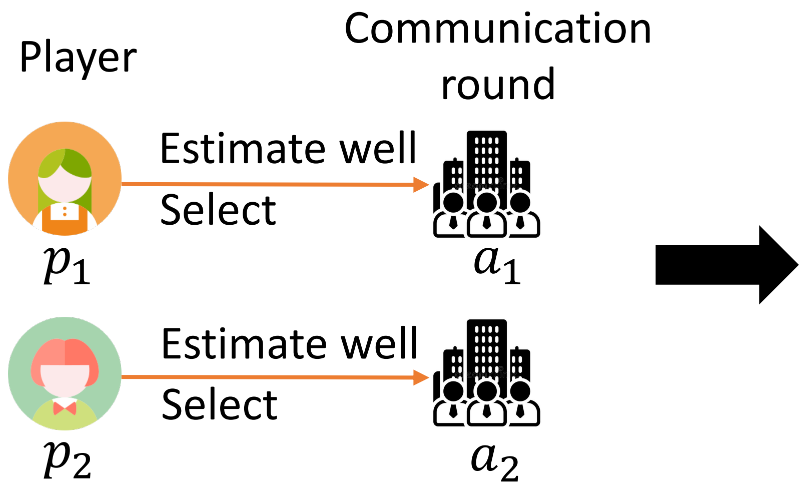


player $p_i's$ preference values

Remark: each player identifying the arms ranked in the first N+1 is enough to find the player-optimal stable matching.

ETGS implementation: Communication (cont.)

- Based on observed all players' matching outcomes [KL, 2023]
  - If p_i has estimated well with ranking ρ_i: select arm a_i
  - Else: Select nothing

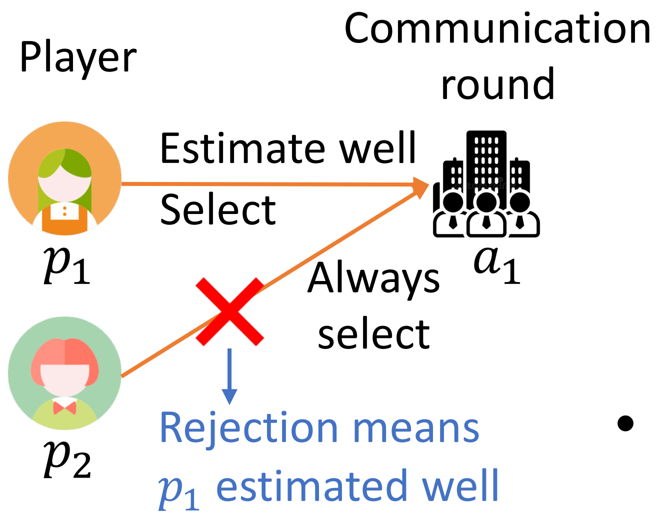


At the communication round, if p_i observes that all players have been matched:

Then all players estimate their preference well

ETGS implementation: Communication (cont.)

- Based on players' own matching outcomes [Zhang et al., 2022]



- Communicate based on every pair of players
  - p_i can transmit information $\{0,1\}$ to $p_{i'}$ based on a_j $(p_i > p_{i'})$
  - In the corresponding round, $p_{i'}$ always selects a_j
  - If p_i finished exploration, selects a_j
    - $p_{i'}$ is rejected, receives information 1
  - Otherwise, p_i do not select a_j
    - $p_{i'}$ is accepted, receive information 0
- If a player cannot receive others' information (all arms prefer this player than others)
  - The player can directly exploit the stable arm
  - Others cannot block it

ETGS: Regret analysis [Kong and Li, 2023]

- Exploration is enough $\Longrightarrow$ Estimated ranking is correct $\Longrightarrow$ All players enter the GS + exploit phase and find the player-optimal stable matching
- The player-optimal regret comes from exploration and communication

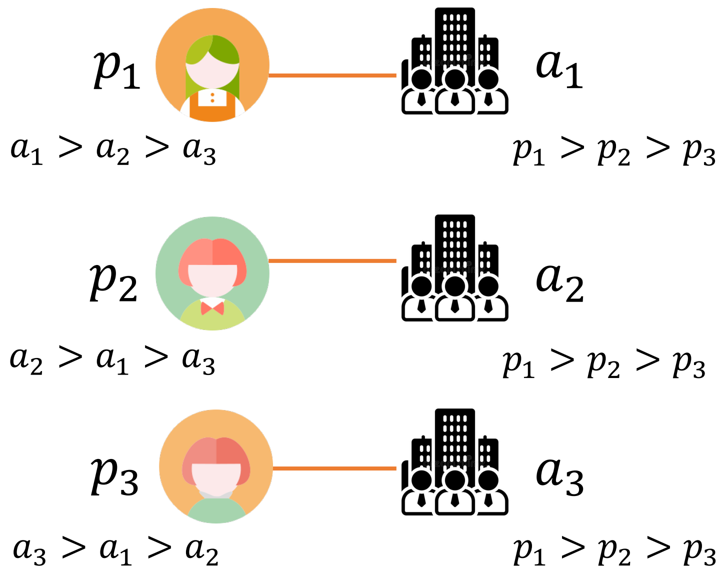
$$\overline{Reg}_i(T) = O\left(\frac{K\log T}{\Delta^2} + \log\left(\frac{K\log T}{\Delta^2}\right)\right)$$

- What is the optimal regret that an algorithm can achieve?

Remark: Δ can be improved as the minimum gap between the first N+1 ranked arms among all players.

Lower bound [Sankararaman et al., 2021]

- Optimally stable bandits
  - All arms have the same preferences
  - $\Longrightarrow$ Unique stable matching exists
  - The stable arm of each player is its optimal arm
- For any player p_i
  - Its stable arm is a_i
  - a_i prefers $p_1, p_2 \dots\dots p_{i-1}$ than p_i
  - $T_{i,j}$: the number of times that p_i selects a_j



$$\overline{Reg}_i(T) \geq \max\left\{\Delta_{i,i,j}\underbrace{\sum_{j\neq i} T_{i,j}}_{p_i \text{ selects sub-optimal arm } a_j},\quad \overbrace{\Delta_{i,\min}}^{\text{The minimum regret that } p_i \text{ may suffer at any round}}\underbrace{\sum_{i'<i} T_{i',i}}_{\text{The optimal arm } a_i \text{ is occupied by a higher-priority player}}\right\}$$

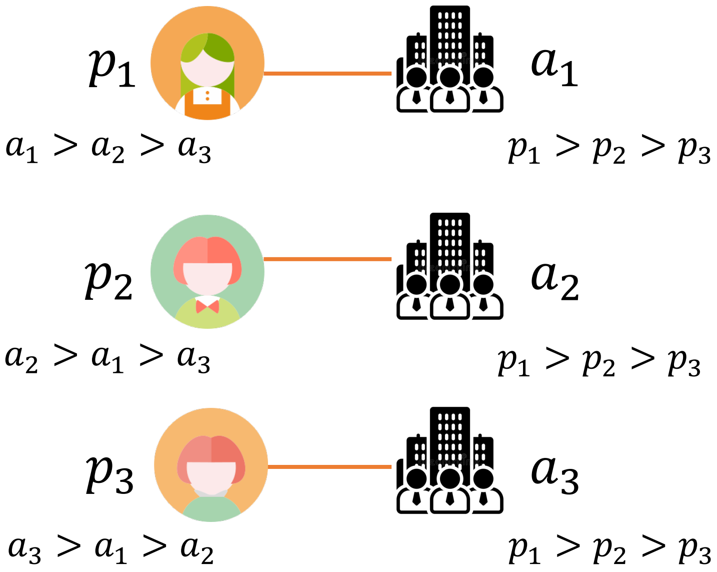
Lower bound (cont.)

- How many times does p_i select a sub-optimal arm a_j ?
  - To distinguish the sub-optimal arm a_j from the optimal arm a_i
  - p_i needs to observe this arm

$$\Omega\left(\frac{\log T}{\Delta_{i,i,j}^2}\right) \text{ times}$$

- K sub-optimal arms cause regret

$$\Omega\left(\sum_{j\neq i}\frac{\log T}{\Delta_{i,i,j}^2}\cdot\Delta_{i,i,j}\right)=\Omega\left(\frac{K\log T}{\Delta}\right)$$



Lower bound (cont.)

- How many times does a_i is occupied by a higher-priority player $p_{i'}$?
  - To distinguish the sub-optimal arm a_i from the optimal arm $a_{i'}$
  - $p_{i'}$ needs to observe this arm

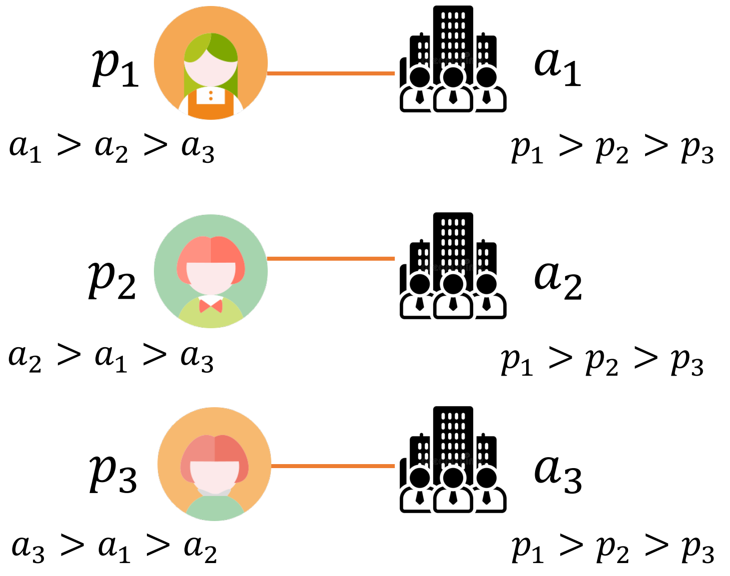
$$\Omega\left(\frac{\log T}{\Delta^2_{i',i',i}}\right) \text{ times}$$

- N higher-priority players cause regret

$$\Omega\left(\sum\nolimits_{i'<i} \frac{\log T}{\Delta^2_{i',i',i}} \cdot \Delta_{i,\min}\right) = \Omega\left(\frac{N\log T}{\Delta^2}\right)$$

- The stable regret satisfies

$$\overline{Reg}_i(T) \geq \Omega\left(\max\left\{\frac{N\log T}{\Delta^2}, \frac{K\log T}{\Delta}\right\}\right)$$



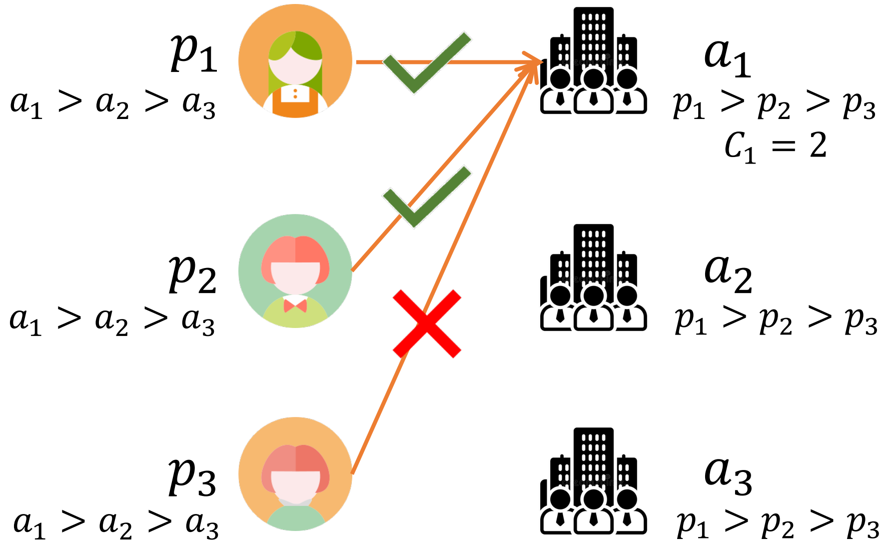
Remark: Δ can be improved as the minimum gap between the player-optimal stable arm and the next preferred one among all players.

One-to-one markets: Results overview

<table>
<tr><th>Regret type</th><th>Regret Bound</th><th>Communication type</th><th>References</th></tr>
<tr><td>Player-optimal</td><td>$O\left(\frac{K\log T}{\Delta^2}\right)$</td><td>Centralized, known Δ</td><td rowspan="2">[Liu et al., 2020]</td></tr>
<tr><td rowspan="3">Player-pessimal</td><td>$O\left(\frac{NK\log T}{\Delta^2}\right)$</td><td>Centralized</td></tr>
<tr><td rowspan="2">$O\left(\frac{N^5K^2\log^2 T}{\rho^{N^4}\Delta^2}\right)$</td><td rowspan="2">Decentralized, observed matching outcomes</td><td>[Liu et al., 2021]</td></tr>
<tr><td>[Kong et al., 2022]</td></tr>
<tr><td>Unique</td><td>$O\left(\frac{NK\log T}{\Delta^2}\right)$</td><td>Decentralized</td><td>[Sankararaman et al., 2021; Basu et al., 2021; Maheshwari et al., 2022]</td></tr>
<tr><td>Optimal stable bandits (Unique)</td><td>$\Omega\left(\frac{N\log T}{\Delta^2}\right)$</td><td>Decentralized</td><td>[Sankararaman et al., 2021]</td></tr>
<tr><td rowspan="3">Player-optimal</td><td>$O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{1/\varepsilon}}\right)$</td><td>Decentralized</td><td>[Basu et al., 2021]</td></tr>
<tr><td rowspan="2">$O\left(\frac{K\log T}{\Delta^2}\right)$</td><td>Decentralized, observed matching outcomes</td><td>[Kong and Li, 2023]</td></tr>
<tr><td>Decentralized</td><td>[Zhang et al., 2022]</td></tr>
</table>

How about many-to-one markets?

- Responsiveness:
  - Each arm a_j has preferences over individual players and a capacity C_j
  - Accept the most preferred C_j players among those who propose to it



Extension of one-to-one algorithms

Centralized ETC/UCB [Wang et al., 2022]

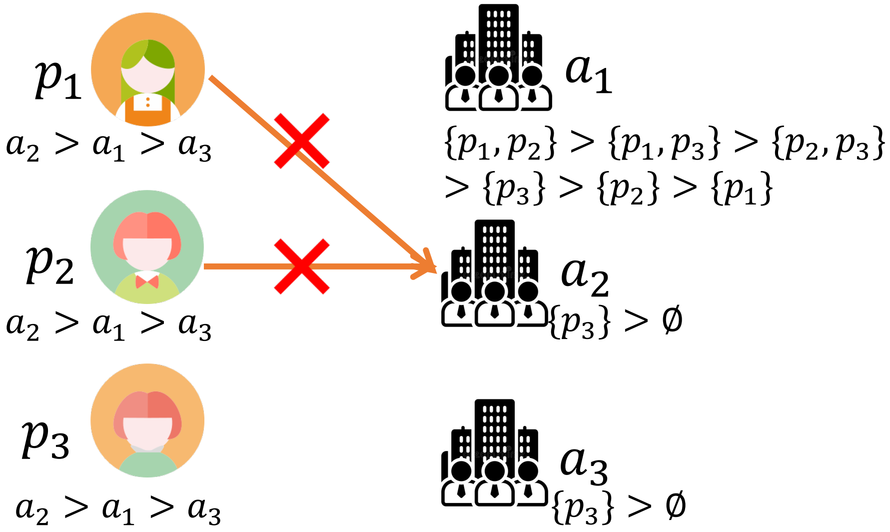
Decentralized UCB [Wang et al., 2022]

ETGS [Kong and Li, 2024]

Results in the same regret upper bounds

Many-to-one markets: Substitutability

- Challenge: Arms may reject all applications, players fail to explore in a round-robin manner



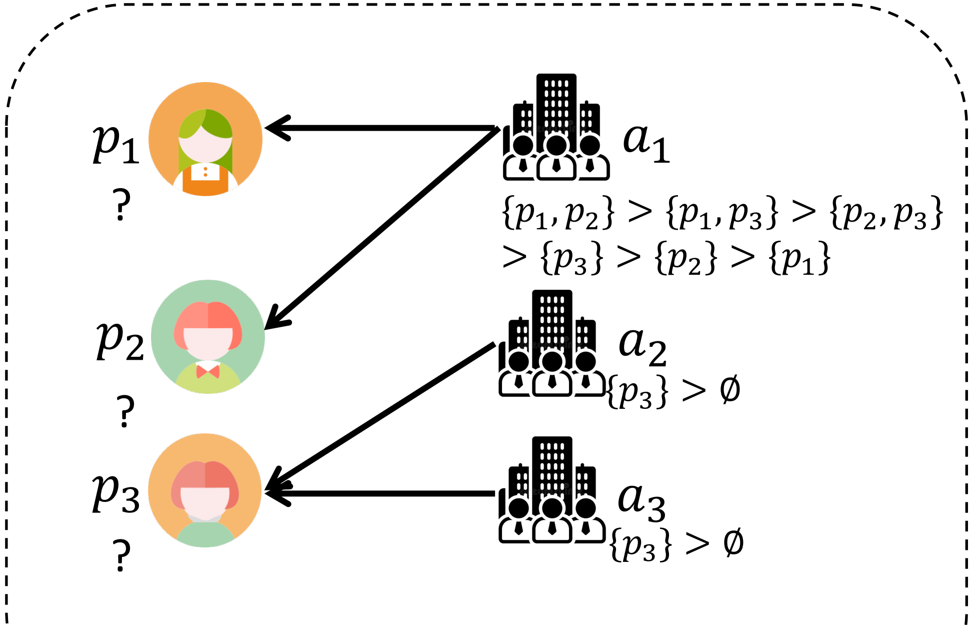
When p_1 or p_2 selects a_2, a_2 reject them

Neither p_1 nor p_2 can receive rewards and learn their unknown preferences over a_2

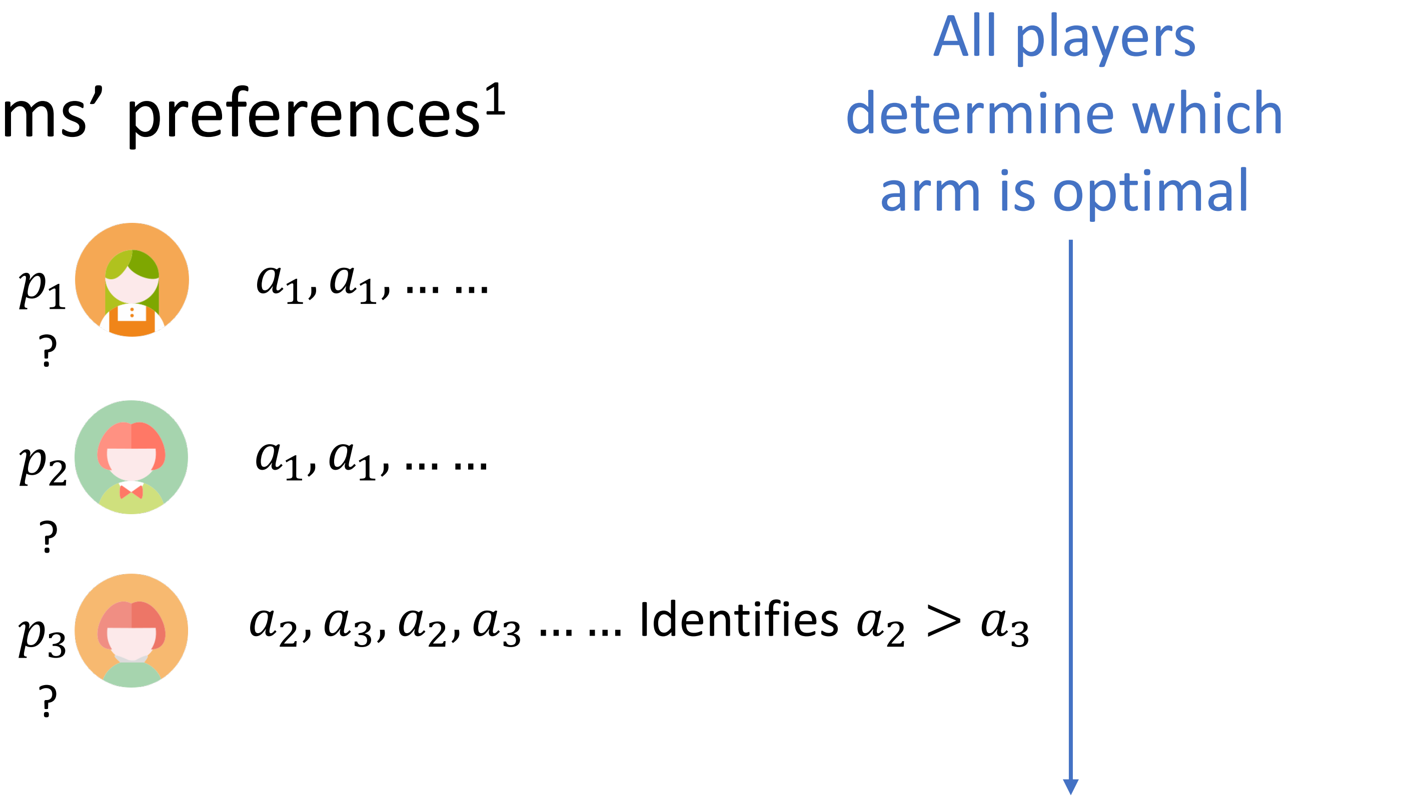
- Idea: Determine which match to explore from the arm side
- From arm-proposal DA to design learning process

Substitutability: Algorithm [KL, 2024]

- First assume players know arms' preferences[1]



Step 1 of arm-proposal DA



Step 1 of the online algorithm

[1]Could use $O(NK^2)$ rounds to learn each arm's most preferred player set at the start of each step of arm-proposal DA.

Substitutability: Theoretical analysis

- Arm-proposal DA produces the player-pessimal stable matching
- Each rejection requires $O(\log T/\Delta^2)$ rounds
  - At most NK rejections happen
- The player-pessimal stable regret of each player p_i satisfies

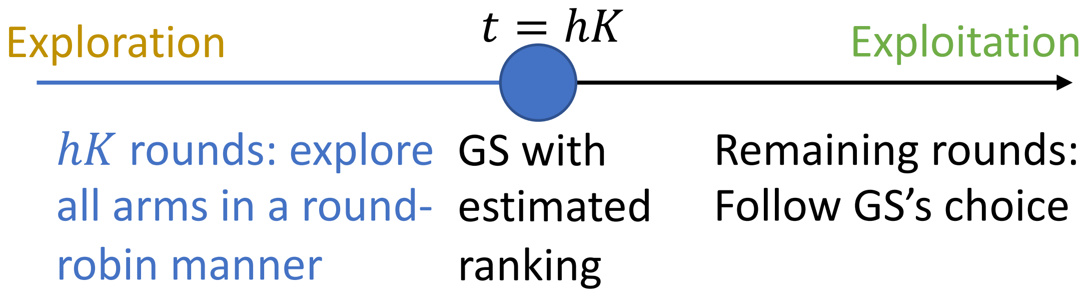
$$\underline{Reg}_i(T) \leq O\left(\frac{NK\log T}{\Delta^2}\right)$$

The first result for combinatorial preferences

Remark: Δ can be improved as the minimum gap between the player-pessimal stable arm and other less-preferred arms among all players.

Strategic behavior: One-to-one setting

- Can players improve their rewards by deviating from the algorithm?



- Warm up: Centralized ETC [Liu et al., 2020]
  - At time $t = hK$: players report the estimated preference ranking
  - In other rounds: players have no freedom of choice
  - Based on the property of GS
    - Single player's deviation cannot improve the matching results (obtain linear reward increase)
  - Is strategy-proof
  - Also holds for the many-to-one setting with responsiveness [Wang et al., 2022]

Strategic behavior: Centralized UCB [Liu et al., 2020]

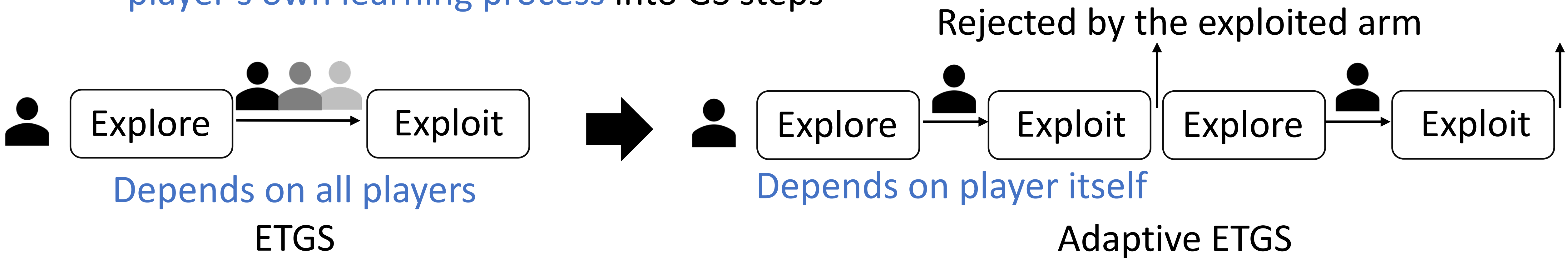
- At each round: players report their UCB rankings
- Open: Not sure whether a single player' deviation can obtain $O(T)$ reward increase
- A weaker result
  - A single player can not match a better arm than the optimal stable matching in $O(T)$ times (Note the regret is only guaranteed for the pessimal stable matching)

Strategic behavior: ETGS [KL, 2023; Zhang et al., 2022]

- If their exists a player whose stable arm is the least preferred one
- He can always report that he has not finished exploration
- All players fail to enter the exploitation phase
- This player: Always match better arms during exploration, $O(T)$ reward increase
- Other players: $O(T/K)$ times match worse arms, $O(T)$ reward decrease
- Not strategy-proof!

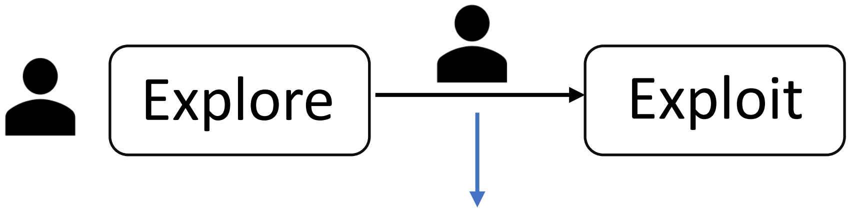
Adaptive ETGS [Kong and Li, 2024]

- Idea: Instead of starting GS + exploitation with all players' agreement, integrating each player's own learning process into GS steps



- Each player explores arms in a round-robin manner
- Once the player identifies the most preferred one, always exploits this arm
- If the exploited arm is occupied by a higher-priority player (the arm "rejects" the player)
    - Enter the next step of GS (explore the next most preferred arm)

Adaptive ETGS: Strategic behavior



Have identified the optimal arm. What to report?

How about reporting NOT?

- Equivalent to delayed entering GS in the offline setting
- Cannot change the final matching results

How about reporting a non-optimal arm?

- Equivalent to misreporting rankings in the offline GS
- Cannot improve the final matched partner

- Is strategy-proof: Single player can not obtain $O(T)$ reward increase (improve the final matched arm) by misreporting the exploration status
- Also can extend to many-to-one markets with responsiveness

Adaptive ETGS: Regret

- Arrangement of round-robin exploration under responsiveness
  - $C := \sum_j C_j$
  - In every C rounds, each player can match each available arm once
- Each step of GS executes $O(C\log T/\Delta^2)$ times
- At most NK steps

- The player-optimal stable regret of each player p_i satisfies

$$\overline{Reg}_i(T) \le O\left(\frac{NKC\log T}{\Delta^2}\right)$$

When reduced to one-to-one setting, the result is $O(N^2K\log T/\Delta^2)$

The coefficient NKC can be improved as $N\min\{N,K\}C$ by using a tight time complexity of offline GS under responsiveness [Kong and Li, 2024]; Δ can be improved as the minimum preference gap between any arms that have higher ranking than the arm after the player-optimal stable one.

Many-to-one markets: Results overview

Setting	Regret type	Regret Bound	Communication type	References
Responsiveness	Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Centralized, known Δ	[Wang et al., 2022]
	Player-pessimal	$O\left(\frac{NK^3\log T}{\Delta^2}\right)$	Centralized	
		$O\left(\frac{N^5K^2\log^2 T}{\kappa^{N^4}\Delta^2}\right)$	Decentralized, observed matching outcomes	
	Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Decentralized, observed matching outcomes, $N \leq K \cdot \min_j C_j$	[Kong and Li, 2024]
		$O\left(\frac{N\min\{N, K\}C\log T}{\Delta^2}\right)$	Decentralized, observed matching outcomes	
Substitutability	Player-pessimal	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Decentralized, observed matching outcomes, known arms' preferences	

Other setting variants

- Contextual information [Li et al., 2022]
- Non-stationary preferences [Ghosh et al., 2022; Muthirayan et al., 2023]
- Two-sided unknown preferences [PD, 2023; PG, 2023]
- Markov matching markets [Min et al., 2022]
- Multi-sided matching markets [Mordig et al., 2021]
- Money transfer [Jagadeesan et al., 2021]
- P2P: matching with budget [Sarkar, 2021]

Summary of Part 3: Bandit algorithms in matching markets

- Centralized algorithms
  - ETC, UCB
  - The failure of UCB
- Decentralized algorithms
  - General markets
  - Markets with unique stable matching
  - Explore-then-GS (ETGS) strategies
- Lower bound
- Many-to-one markets
- Strategic behavior
  - Adaptive ETGS
- Other variants

Part 4: Beyond Matching Markets

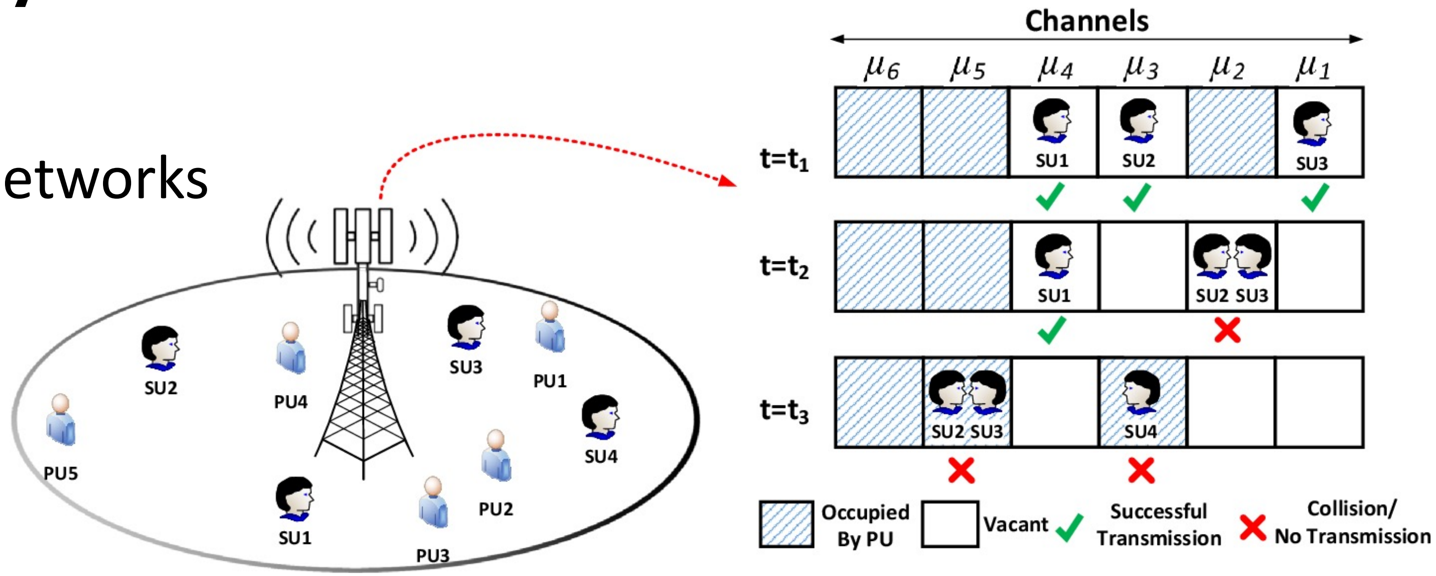
Shuai Li, Fang Kong
Shanghai Jiao Tong University

Outline

- Multi-player bandits
  - Example: Cognitive radio networks
  - Centralized settings
  - Decentralized settings
- Learning in auctions
  - One seller and multiple buyers
  - Multiple sellers and buyers
  - Dynamic sellers and buyers
  - ……

Multi-player bandits

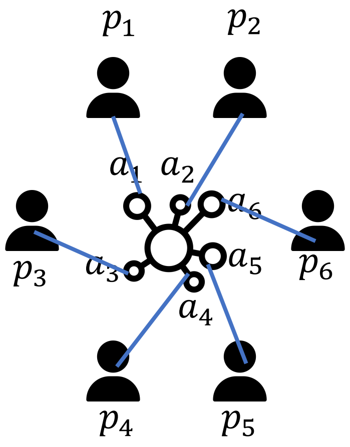
Cognitive radio networks



- N users (players) hope to use K channels (arms) for transmission
- A single user repeatedly chooses among a choice of K channels
- At each round $t = 1,2, \dots T$
  - Each player p_i selects an arm $A_i(t)$
  - $X_{i,j}(t)$: Information transmission quality, with unknown expectation $\mu_{i,j}$
  - If collied with other players, only receive reward 0

Darak, Sumit J., and Manjesh K. Hanawal. "Distributed Learning in Ad-Hoc Networks: A Multi-player Multi-armed Bandit Framework." arXiv preprint arXiv:2004.00367 (2020).

Multi-player bandits: Objective



- A matching m is a one-to-one function: $[N] \to [K]$
- The expected utility of m:

$$U(m) \coloneqq \sum_i \mu_{i,m_i}$$

- Objective: Minimize the collective regret

Collision indicator:
1 if collide; 0 otherwise

$$Reg(T) = T \cdot \max_m U(m) - \mathbb{E}\left[\sum_{t=1}^{T} \sum_i \underbrace{\mu_{i,A_i(t)}\left(1 - \eta_{A_i(t)}(t)\right)}_{\text{Final reward of player } p_i \text{ at time } t}\right]$$

Comparison: Multi-player V.S. Matching markets

- Collision
  - Multi-player bandits: Players receive no reward
  - Matching markets: Accepted player(s) receive the reward (based on arms' preferences)

- Objective
  - Multi-player bandits: Collective utilities
  - Matching markets: Equilibrium state of the market

Multi-player bandits: Settings

- Centralized setting:
  - All players follow a central platform to avoid conflicts
- Decentralized setting:
  - Different levels of observed information
    - Pre-agreement
    - Collision information
    - Only observe the final reward
    - ……

Multi-player bandits: Centralized setting

- Homogeneous setting [Anantharam et al., 1987]:
  - All players have the same preferences over arms
  - The problem reduces to bandits with multiple plays [Komiyama et al., 2015]
    - A single player selects N over K arms in each round
- Heterogeneous setting:
  - Players have different preferences over arms
  - The problem reduces to combinatorial bandits problem [Chen et al., 2013]:
    - A single player and NK arms (original player-arm pairs)
    - At each round: The player selects an action (a matching), and receives the corresponding reward

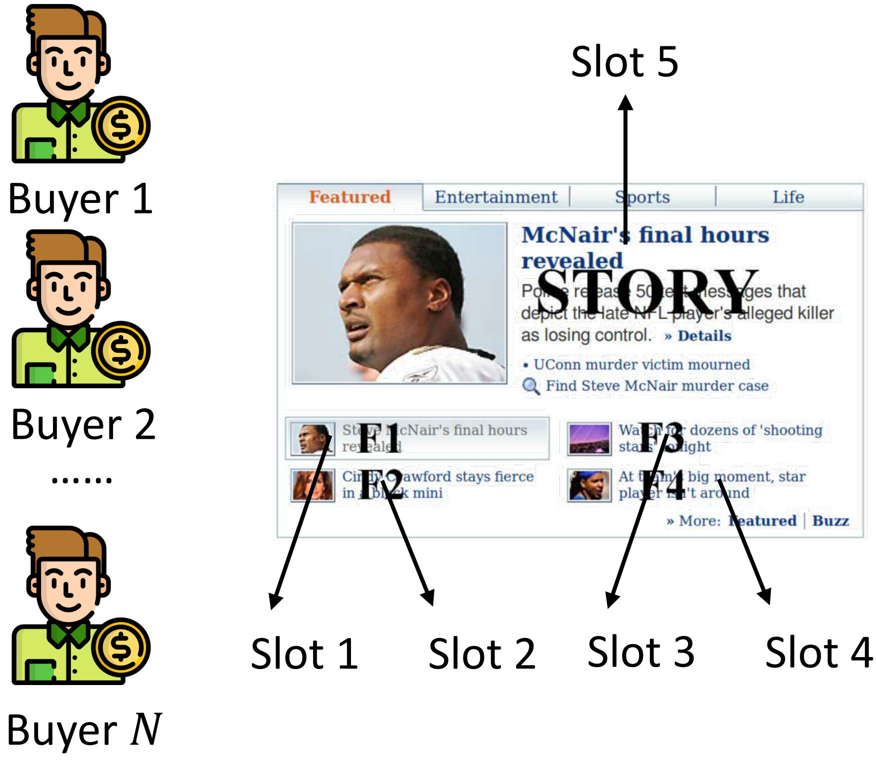
Multi-player bandits: Decentralized setting

- Key point: avoid conflicts among players
- Based on pre-agreement:
  - Each player has a rank i and aims to focus on the i-th best arm [Anandkumar et al., 2010]
- Based on the collision information:
  - Musical chair [Rosenski et al., 2016]:
    - A player uniformly sample arms and focus on this arm until no collision
    - After some time, with high probability, players focus on different arms
  - Communication [Boursier et al., 2019]:
    - Collision: receive 1; no collision: receive 0
  - ……
- Without collision information [Bubeck et al., 2020; 2021]

- Other multi-agent interaction rules?

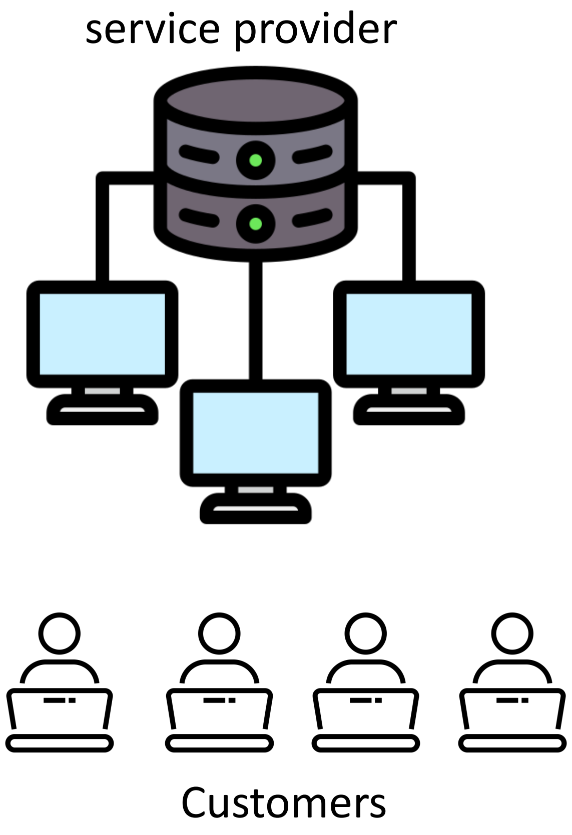
Example of auction: Online advertising

- A publisher (mechanism) has a set of advertising slots
- Assigns them to N buyers
- When a slot is assigned to a buyer, its reward corresponds to the click-through-rate (CTR) ...
- Buyers do not know their exact values towards an assignment



Example of auction: platform-as-a-service

- The service provider (seller) serves multiple customers (buyer) using the same compute cluster
- The seller chooses a service level for each buyer, and charge them accordingly
- The buyer's experience is affected by exogenous stochastic factors such as traffic, machine failures
- Buyers do not know their values towards an assignment



Formulation: Repeated auction [Kandasamy et al., 2023]

- 1 seller and N buyers (players)
- The seller chooses an assignment ω, charge a price P_i to player $i \in [N]$
- For each assignment ω
  - profits, satisfaction, …
  - Player i's value is $v_i(\omega)$ (unknown)
  - Seller's value is $v_0(\omega)$
  - cost
- In each round t:
  - The seller chooses an assignment $\omega(t)$ and charge price $P_i(t)$
  - Player i receives a reward $X_i(t)$ with expectation $v_i(\omega_t)$

Repeated auction: Objective

- Social welfare

$$V(\omega_t) = v_0(\omega_t) + \sum_i v_i(\omega_t)$$

- Optimal assignment $\omega_* \in \arg\max_\omega V(\omega)$

- Minimize the social welfare regret

$$Reg(T) = T \cdot V(\omega_*) - \mathbb{E}\left[\sum_{t=1}^{T} V(\omega_t)\right]$$

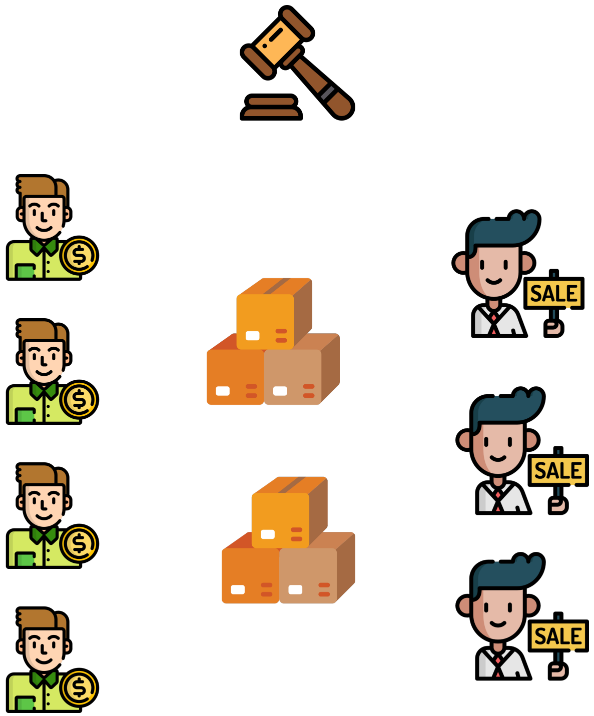
Repeated auction: Objective (cont.)

- Players' own utilities
  - Given assignment ω_t and price $P_i(t)$
  - The player i's expected utility is $u_i(t) = v_i(\omega_t) - p_i(t)$
  - Cumulative utilities: $\sum_t u_i(t)$
- Truthfulness
  - A single player cannot improve its cumulative utilities by deviating from the algorithm
- Individual rationality
  - Do not charge a player more than her bid
  - The cumulative utilities of any player is non-negative

Multiple sellers: Double auctions

- N buyers, K sellers
- Single type of good

- Each buyer $i \in [N]$ has a unknown valuation B_i
- Each seller $j \in [K]$ has a unknown valuation S_j



Multiple sellers: Double auctions' setting [Basu and Sankararaman, 2023]

- In each round t:
  - Each buyer i submits bid $b_i(t)$, each seller j submits bid $s_j(t)$
  - The mechanism outputs:
    - Participants subsets $\mathcal{P}_b(t), \mathcal{P}_s(t)$ with the same size $K(t) \leq \min\{N, K\}$
    - Trading price $P(t)$
  - Participating buyer i receives utility $u_i(t) = X_i(t) - P(t)$
  - Participating seller j receives utility $u_j(t) = P(t) - X_j(t)$
  - Here $X_i(t)$ is with expectation B_i, $X_j(t)$ is with expectation S_j
  - Other buyers and sellers receive utility 0

Multiple sellers: Double auctions' objective

- Social welfare
  - Cumulative values of agents who hold the goods

$$\sum_{i\in\mathcal{P}_b(t)} B_i + \sum_{j\in[K]\backslash\mathcal{P}_s(t)} S_j$$

- Minimize the social welfare regret

$$Reg(T)$$
$$= T\Big(\sum_{i\in\mathcal{P}_b^*} B_i + \sum_{j\in[K]\backslash\mathcal{P}_s^*} S_j\Big) - \mathbb{E}\Big[\sum_{t=1}^{T}\Big(\sum_{i\in\mathcal{P}_b(t)} B_i + \sum_{j\in[M]\backslash\mathcal{P}_s(t)} S_j\Big)\Big]$$

Optimal participating buyers　　Optimal participating sellers

- Minimize the individual regret

$$Reg_{b,i}(T) = T(B_i - p^*)\mathbb{I}(i \in \mathcal{P}_b^*) - \mathbb{E}\Big[\sum_{t:i\in\mathcal{P}_b(t)} (B_i - P(t))\Big]$$

  - Similar for the seller side

Optimal trading price

Dynamic sellers and buyers [Cesa-Bianchi et al., 2020]

- At each time t, a seller and a buyer arrive and wish to trade some good
- The seller's and buyer's valuation S_t, B_t
  - Realizations of underlying values s_t,b_t
- The mechanism selects a price P_t
- The trade occurs if and only if $S_t \leq P_t \leq B_t$
- The learner gains a reward $(B_t - S_t)\mathbb{I}\{S_t \leq P_t \leq B_t\}$
- Aim to selecting prices to minimize the regret

$$Reg(T) = \max_p \mathbb{E}\left[\sum_{t=1}^{T}(B_t - S_t)\mathbb{I}\{S_t \leq p \leq B_t\} - \sum_{t=1}^{T}(B_t - S_t)\mathbb{I}\{S_t \leq P_t \leq B_t\}\right]$$

Other variants

- Different auction scenarios
- Different trading mechanisms
- Different learning side
    - The agent side
    - The mechanism side
- ……
- [Gatti et al., 2012; Kakade et al., 2013; Babaioff et al., 2014; Babaioff et al., 2015; Nazerzadeh et al., 2016; Weed et al., 2016; Nedelec et al., 2019; ……]

Summary of Part 4: Beyond matching markets

- Multi-player bandits
  - Example: Cognitive radio networks
  - Centralized settings
  - Decentralized settings
- Learning in auctions
  - One seller and multiple buyers
  - Multiple sellers and buyers
  - Dynamic sellers and buyers
  - Other variants

Open problems: Matching markets

- Optimality
  - Regret
  - Strategic behavior

Open problems: Regret

- What is the optimal regret in the one-to-one setting?

Regret type	Regret Bound	Communication type	References
Optimal stable bandits (Unique stable matching)	$\Omega(N \log T/\Delta^2)$	Decentralized	[Sankararaman et al., 2021]
Player-optimal stable matching	$O(K \log T/\Delta^2)$	Decentralized	[Kong and Li, 2023; Zhang et al., 2022]

Gap

- Recall that to ensure players can be matched, all existing works assume $N \leq K$

Open problems: Regret (cont.)

- What is the optimal regret in the many-to-one setting?

Setting	Regret type	Regret Bound	Communication type	References
Responsiveness	Player-optimal stable matching	$O\left(\frac{K\log T}{\Delta^2}\right)$	Decentralized, known matching outcomes, $N \le K \cdot \min_j C_j$	[Kong and Li, 2024]
	Player-optimal stable matching	$O\left(\frac{N\min\{N,K\}C\log T}{\Delta^2}\right)$	Decentralized, known matching outcomes	
Substitutability	Player-pessimal stable matching	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Decentralized, known matching outcomes	

Is the player-optimal stable matching achievable?

What is the optimal regret under the responsiveness?

Open problems: Regret & Strategic behavior

- What is the optimal regret when guaranteeing strategy-proofness?

Regret type	Regret Bound	Strategy-proof	References
Player-optimal stable matching	$O(K\log T/\Delta^2)$	No	[Kong and Li, 2023; Zhang et al., 2022]
Player-optimal stable matching	$O(N^2K\log T/\Delta^2)$ $O(N^2C\log T/\Delta^2)$ responsiveness	Yes	[Kong and Li, 2024]

Open problems: Matching markets (cont.)

- How to generalize the setting and what is the optimal regret in these settings?
  - How to deal with two-sided unknown preferences?
    - Existing works assume arms have known preferences and use this to conduct coordination/communication. But arms may also have unknown preferences
  - How to deal with players' indifferent preferences?
    - Players may be indifferent over multiple arms
  - How to utilize the contextual information to accelerate the learning efficiency?
    - Agents' features (gender, age, hometown)
  - How to handle asynchronous agents?
    - Agents may enter the system at different times

……

Open problems: Other mechanism design

- Optimality in existing settings
  - What is the optimal social welfare regret, individual regret?
  - How to guarantee strategy-proofness while ensuring efficiency?
- Model generalizations
  - Relax the required assumptions/observation on agents' rewards
  - Consider other trading mechanisms to ensure the desired properties
  - Consider other common auction scenarios

Thanks!
&
Questions?

Shuai Li

- Associate professor at Shanghai Jiao Tong University
- Research interests: Bandit/RL algorithms
- Personal website: https://shuaili8.github.io/

Fang Kong

- Ph.D. candidate student at Shanghai Jiao Tong University
- Research interests: Bandit/RL algorithms
- Personal website: https://fangkongx.github.io

Credit: Some images are from Flaticon.com

References: Part 1

- Gale, David, and Lloyd S. Shapley. "College admissions and the stability of marriage." The American Mathematical Monthly 69.1 (1962): 9-15.
- Roth, Alvin E. "The evolution of the labor market for medical interns and residents: a case study in game theory." Journal of political Economy 92.6 (1984a): 991-1016.
- Dubins, Lester E., and David A. Freedman. "Machiavelli and the Gale-Shapley algorithm." The American Mathematical Monthly 88.7 (1981): 485-494.
- Roth, Alvin E. "The economics of matching: Stability and incentives." Mathematics of operations research 7.4 (1982): 617-628.
- Kelso Jr, Alexander S., and Vincent P. Crawford. "Job matching, coalition formation, and gross substitutes." Econometrica: Journal of the Econometric Society (1982): 1483-1504.
- Roth, Alvin E. "Stability and polarization of interests in job matching." Econometrica: Journal of the Econometric Society (1984b): 47-57.
- Roth, Alvin E., and Marilda Sotomayor. "Two-sided matching." Handbook of game theory with economic applications 1 (1992): 485-541.

References: Part 2

- Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.
- Li, Lihong, et al. "A contextual-bandit approach to personalized news article recommendation." International conference on World wide web. 2010.
- Kocsis, Levente, and Csaba Szepesvári. "Bandit based monte-carlo planning." European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
- Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." nature 529.7587 (2016): 484-489.
- Yu, Baosheng, Meng Fang, and Dacheng Tao. "Linear submodular bandits with a knapsack constraint." Proceedings of the AAAI Conference on Artificial Intelligence. 2016.
- Liang, Jia Hui, et al. "Learning rate based branching heuristic for SAT solvers." Theory and Applications of Satisfiability Testing–SAT 2016: 19th International Conference, Bordeaux, France, July 5-8, 2016, Proceedings 19. Springer International Publishing, 2016.
- Bastani, Hamsa, et al. "Efficient and targeted COVID-19 border testing via reinforcement learning." Nature 599.7883 (2021): 108-113.

References: Part 2

- Hu, Yujing, et al. "Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.
- Garivier, Aurélien, Tor Lattimore, and Emilie Kaufmann. "On explore-then-commit strategies." Advances in Neural Information Processing Systems 29 (2016).
- Audibert, Jean-Yves, and Sébastien Bubeck. "Best arm identification in multi-armed bandits." COLT-23th Conference on learning theory-2010. 2010.
- Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." Machine learning 47 (2002): 235-256.
- Agrawal, Shipra, and Navin Goyal. "Further Optimal Regret Bounds For Thompson Sampling." Sixteenth International Conference on Artificial Intelligence and Statistics. 2013.
- Lai, Tze Leung, and Herbert Robbins. "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics 6.1 (1985): 4-22.

References: Part 3

- Liu, Lydia T., Horia Mania, and Michael Jordan. "Competing bandits in matching markets." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
- Liu, Lydia T., et al. “Bandit learning in decentralized matching markets.” Journal of Machine Learning Research 22.211 (2021): 1-34.
- Maheshwari, Chinmay, Shankar Sastry, and Eric Mazumdar. "Decentralized, communication-and coordination-free learning in structured matching markets." Advances in Neural Information Processing Systems 35 (2022): 15081-15092.
- Kong, Fang, Junming Yin, and Shuai Li. "Thompson Sampling for Bandit Learning in Matching Markets.” International Joint Conference on Artificial Intelligence. 2022.
- Basu, Soumya, Karthik Abinav Sankararaman, and Abishek Sankararaman. "Beyond $ log^ 2 (T) $ regret for decentralized bandits in matching markets." International Conference on Machine Learning. PMLR, 2021.
- Sankararaman, Abishek, Soumya Basu, and Karthik Abinav Sankararaman. "Dominate or delete: Decentralized competing bandits in serial dictatorship." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

References: Part 3

- Kong, Fang, and Shuai Li. "Player-optimal Stable Regret for Bandit Learning in Matching Markets." Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). Society for Industrial and Applied Mathematics, 2023.
- Zhang, Yirui, Siwei Wang, and Zhixuan Fang. "Matching in Multi-arm Bandit with Collision." Advances in Neural Information Processing Systems 35 (2022): 9552-9563.
- Wang, Zilong, et al. "Bandit learning in many-to-one matching markets." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
- Kong, Fang, and Shuai Li. "Improved Bandits in Many-to-one Matching Markets with Incentive Compatibility." AAAI Conference on Artificial Intelligence. 2024.
- Min, Yifei, et al. "Learn to match with no regret: Reinforcement learning in markov matching markets." Advances in Neural Information Processing Systems 35 (2022): 19956-19970.
- Li, Yuantong, et al. "Rate-optimal contextual online matching bandit." arXiv preprint arXiv:2205.03699 (2022).

References: Part 3

- Pagare, Tejas, and Avishek Ghosh. "Two-Sided Bandit Learning in Fully-Decentralized Matching Markets." ICML 2023 Workshop The Many Facets of Preference-Based Learning. 2023.
- Muthirayan, Deepan, et al. "Competing bandits in time varying matching markets." *Learning for Dynamics and Control Conference*. PMLR, 2023.
- Ghosh, Avishek, et al. "Decentralized competing bandits in non-stationary matching markets." arXiv preprint arXiv:2206.00120 (2022).
- Pokharel, Gaurab, and Sanmay Das. "Converging to Stability in Two-Sided Bandits: The Case of Unknown Preferences on Both Sides of a Matching Market." arXiv preprint arXiv:2302.06176 (2023).
- Jagadeesan, Meena, et al. "Learning equilibria in matching markets from bandit feedback." Advances in Neural Information Processing Systems 34 (2021): 3323-3335.
- Mordig, Maximilian, et al. "Finding Stable Matchings in PhD Markets with Consistent Preferences and Cooperative Partners." *arXiv preprint arXiv:2102.11834* (2021).

References: Part 4

- Darak, Sumit J., and Manjesh K. Hanawal. "Distributed Learning in Ad-Hoc Networks: A Multi-player Multi-armed Bandit Framework." arXiv preprint arXiv:2004.00367 (2020).
- Boursier, Etienne, and Vianney Perchet. "A survey on multi-player bandits." arXiv preprint arXiv:2211.16275 (2022).
- Anantharam, Venkatachalam, Pravin Varaiya, and Jean Walrand. "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards." IEEE Transactions on Automatic Control 32.11 (1987): 968-976.
- Komiyama, Junpei, Junya Honda, and Hiroshi Nakagawa. "Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays." International Conference on Machine Learning. PMLR, 2015.
- Chen, Wei, Yajun Wang, and Yang Yuan. "Combinatorial multi-armed bandit: General framework and applications." International conference on machine learning. PMLR, 2013.
- Wang, Siwei, and Wei Chen. "Thompson sampling for combinatorial semi-bandits." International Conference on Machine Learning. PMLR, 2018.

References: Part 4

- Anandkumar, Animashree, Nithin Michael, and Ao Tang. "Opportunistic spectrum access with multiple users: Learning under competition." 2010 Proceedings IEEE INFOCOM. IEEE, 2010.
- Rosenski, Jonathan, Ohad Shamir, and Liran Szlak. "Multi-player bandits–a musical chairs approach." International Conference on Machine Learning. PMLR, 2016.
- Boursier, Etienne, and Vianney Perchet. "SIC-MMAB: Synchronisation involves communication in multiplayer multi-armed bandits." Advances in Neural Information Processing Systems 32 (2019).
- Bubeck, Sébastien, and Thomas Budzinski. "Coordination without communication: optimal regret in two players multi-armed bandits." Conference on Learning Theory. PMLR, 2020.
- Bubeck, Sébastien, Thomas Budzinski, and Mark Sellke. "Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions." Conference on Learning Theory. PMLR, 2021.
- Vickrey, William. "Counterspeculation, auctions, and competitive sealed tenders." The Journal of finance 16.1 (1961): 8-37.

References: Part 4

- Clarke, Edward H. "Multipart pricing of public goods." Public choice (1971): 17-33.
- Groves, Theodore. "Efficient collective choice when compensation is possible." The Review of Economic Studies 46.2 (1979): 227-241.
- Kandasamy, Kirthevasan, et al. "Vcg mechanism design with unknown agent values under stochastic bandit feedback." Journal of Machine Learning Research 24.53 (2023): 1-45.
- Basu, Soumya, and Abishek Sankararaman. "Double Auctions with Two-sided Bandit Feedback." Advances in Neural Information Processing Systems 36 (2023).
- Sarkar, Soumajyoti. "Centralized Borrower and Lender Matching under Uncertainty for P2P Lending." Companion Proceedings of the Web Conference 2021. 2021.
- Cesa-Bianchi, Nicolò, et al. "A regret analysis of bilateral trade." *Proceedings of the 22nd ACM Conference on Economics and Computation*. 2021.
- Weed, Jonathan, Vianney Perchet, and Philippe Rigollet. "Online learning in repeated auctions." Conference on Learning Theory. PMLR, 2016.

References: Part 4

- Nedelec, Thomas, Noureddine El Karoui, and Vianney Perchet. "Learning to bid in revenue-maximizing auctions." International Conference on Machine Learning. PMLR, 2019.
- Babaioff, Moshe, Robert D. Kleinberg, and Aleksandrs Slivkins. "Truthful mechanisms with implicit payment computation." Journal of the ACM (JACM) 62.2 (2015): 1-37.
- Babaioff, Moshe, Yogeshwer Sharma, and Aleksandrs Slivkins. "Characterizing Truthful Multi-armed Bandit Mechanisms." *SIAM Journal on Computing* 43.1 (2014): 194-230.
- Nazerzadeh, Hamid, et al. "Where to sell: Simulating auctions from learning algorithms." Proceedings of the 2016 ACM Conference on Economics and Computation. 2016.
- Gatti, Nicola, Alessandro Lazaric, and Francesco Trovo. "A truthful learning mechanism for contextual multi-slot sponsored search auctions with externalities." Proceedings of the 13th ACM Conference on Electronic Commerce. 2012.
- Kakade, Sham M., Ilan Lobel, and Hamid Nazerzadeh. "Optimal dynamic mechanism design and the virtual-pivot mechanism." Operations Research 61.4 (2013): 837-854.