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 Matching Markets

Shuai Li

2024.12.3 at TBSI

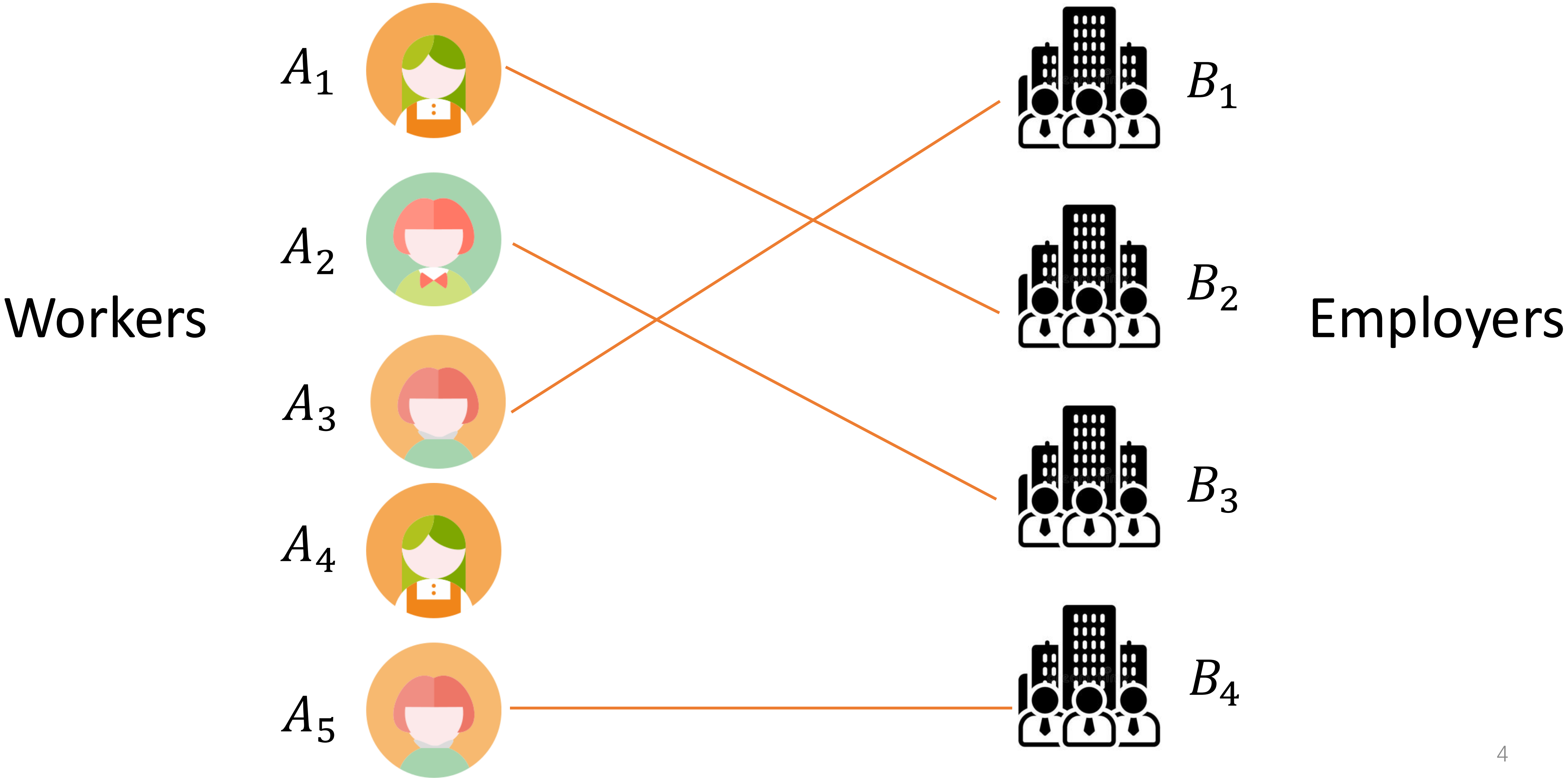
Part 1: Two-sided Matching Markets

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

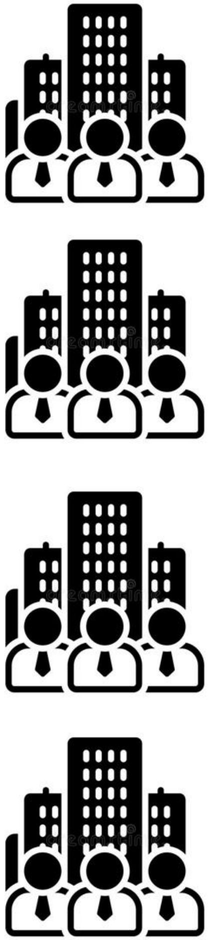
A_1 : $B_2 > B_3 > B_1 > B_4$

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

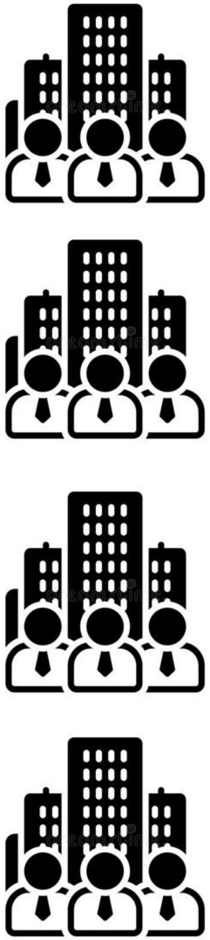
A_1

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)]

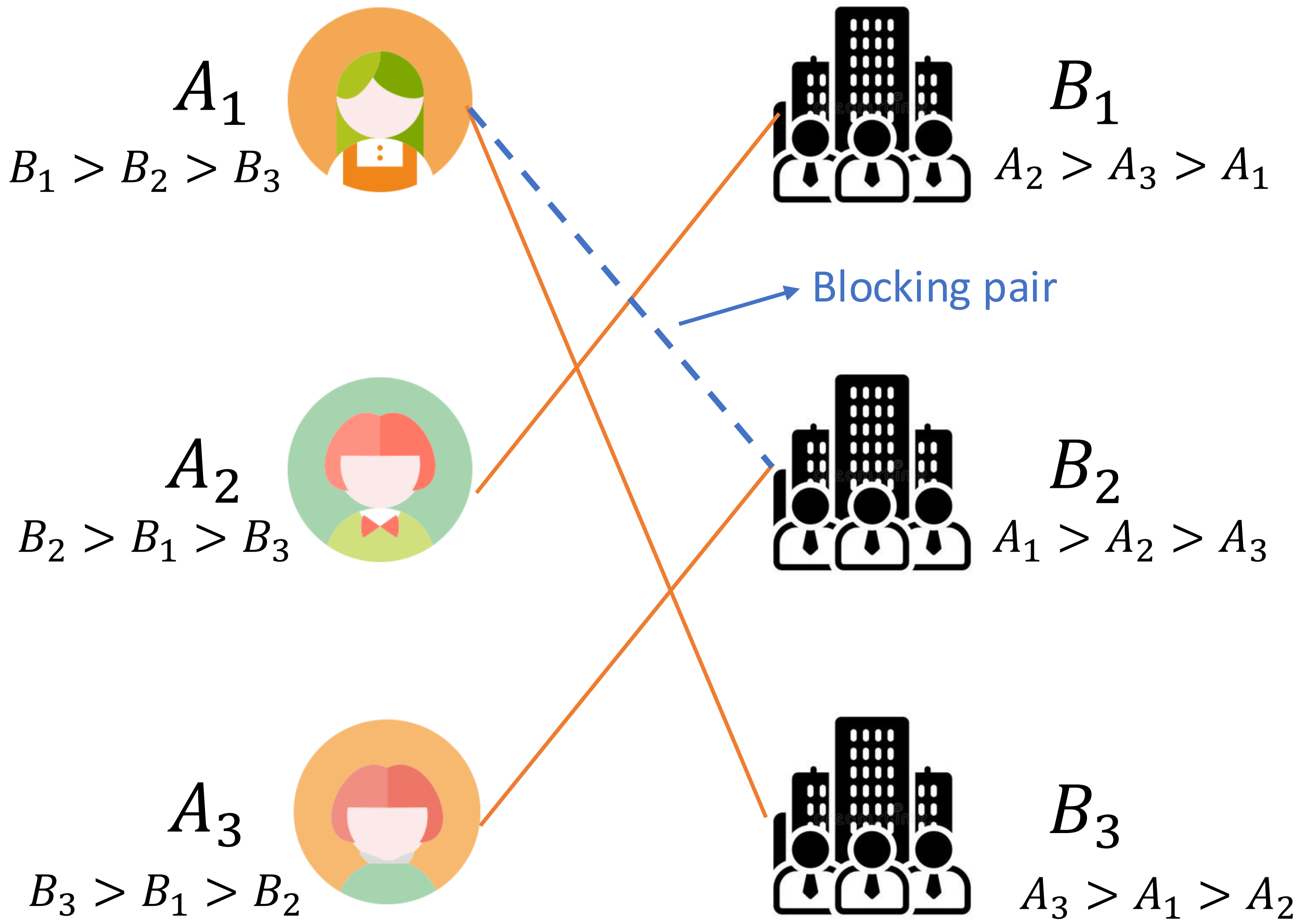
- Hospital side
  - Internship has relatively low cost
- Student side
  - closely engage with clinical medicine through internships

- 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



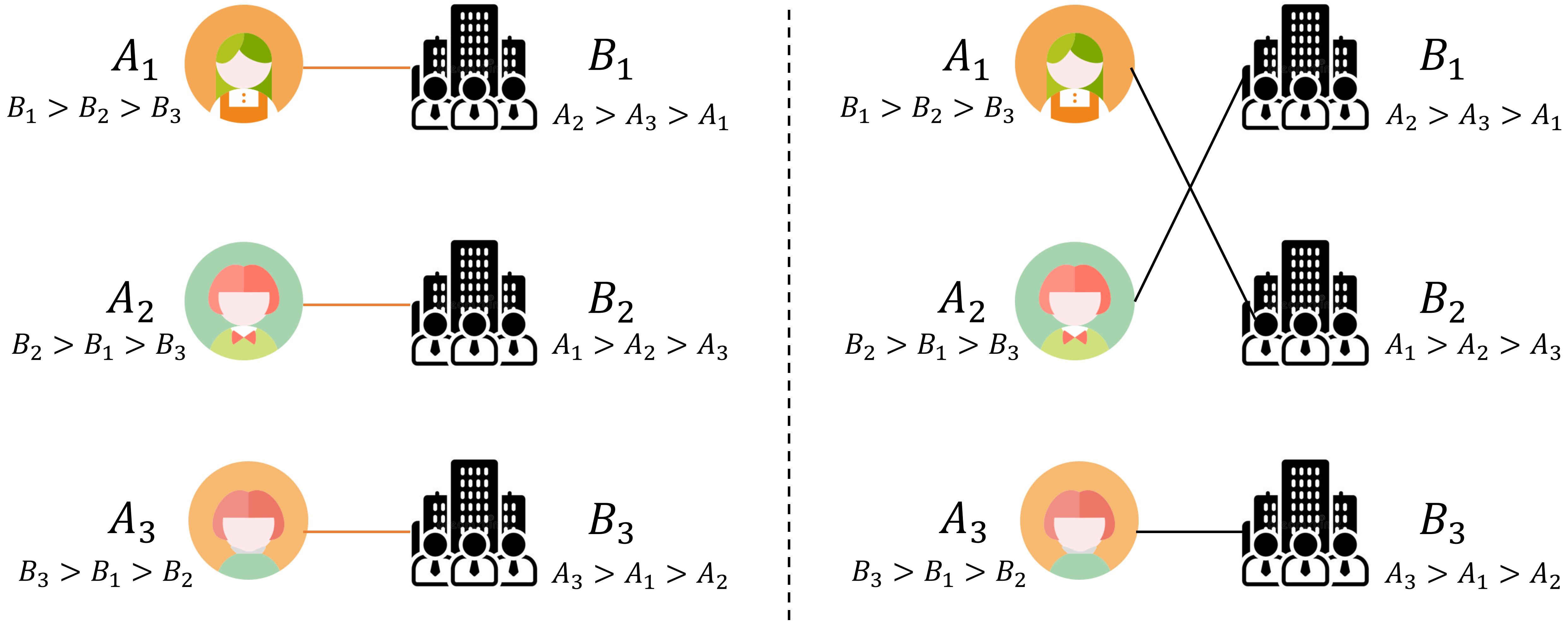
Participants have no incentive to abandon their current partner,

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.

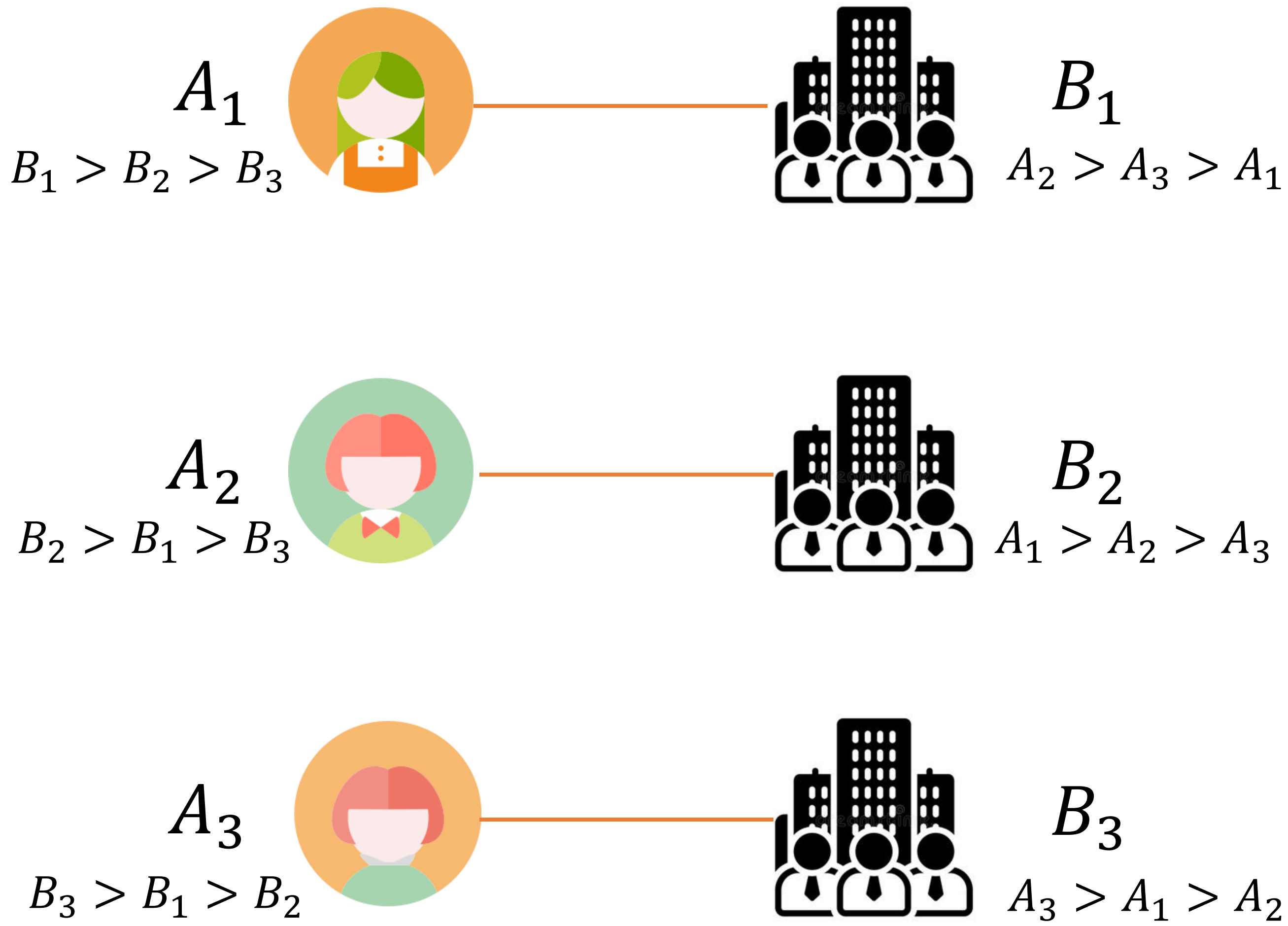
May be more than one stable matchings



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

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

A-side optimal stable matching[1]

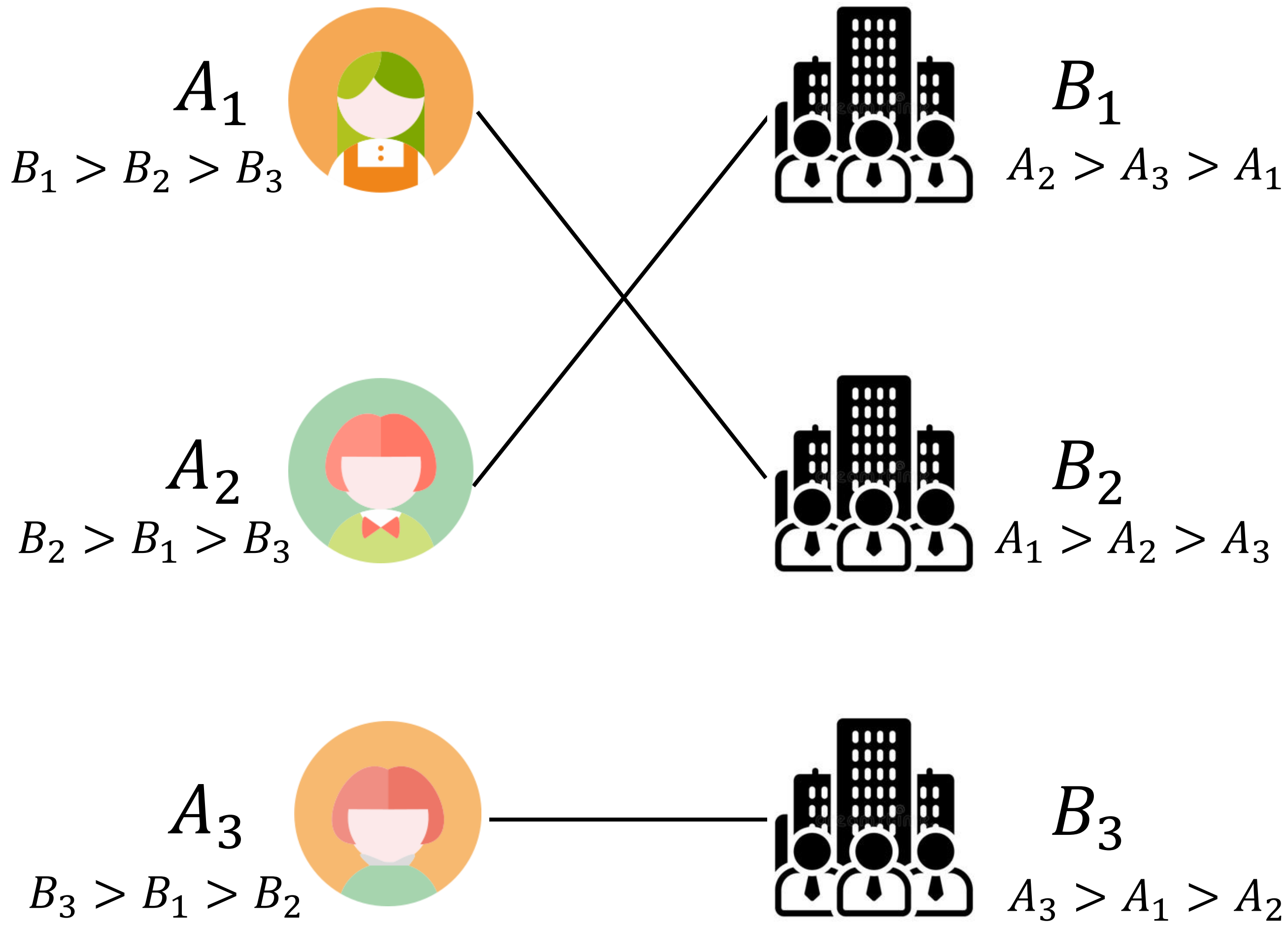


Each agent on A-side is matched with the most preferred partner among all stable matchings

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

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

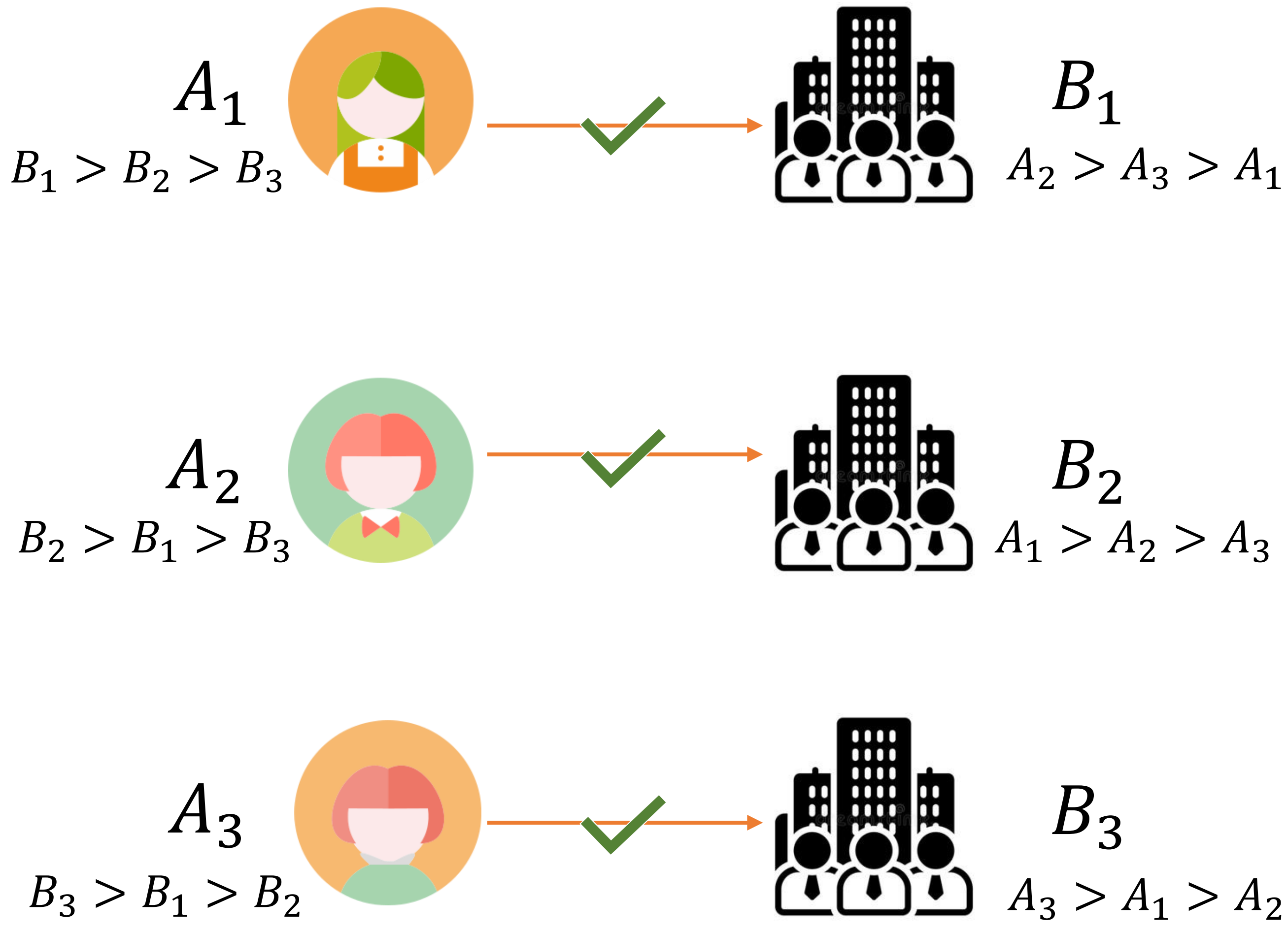
A-side pessimal stable matching



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?

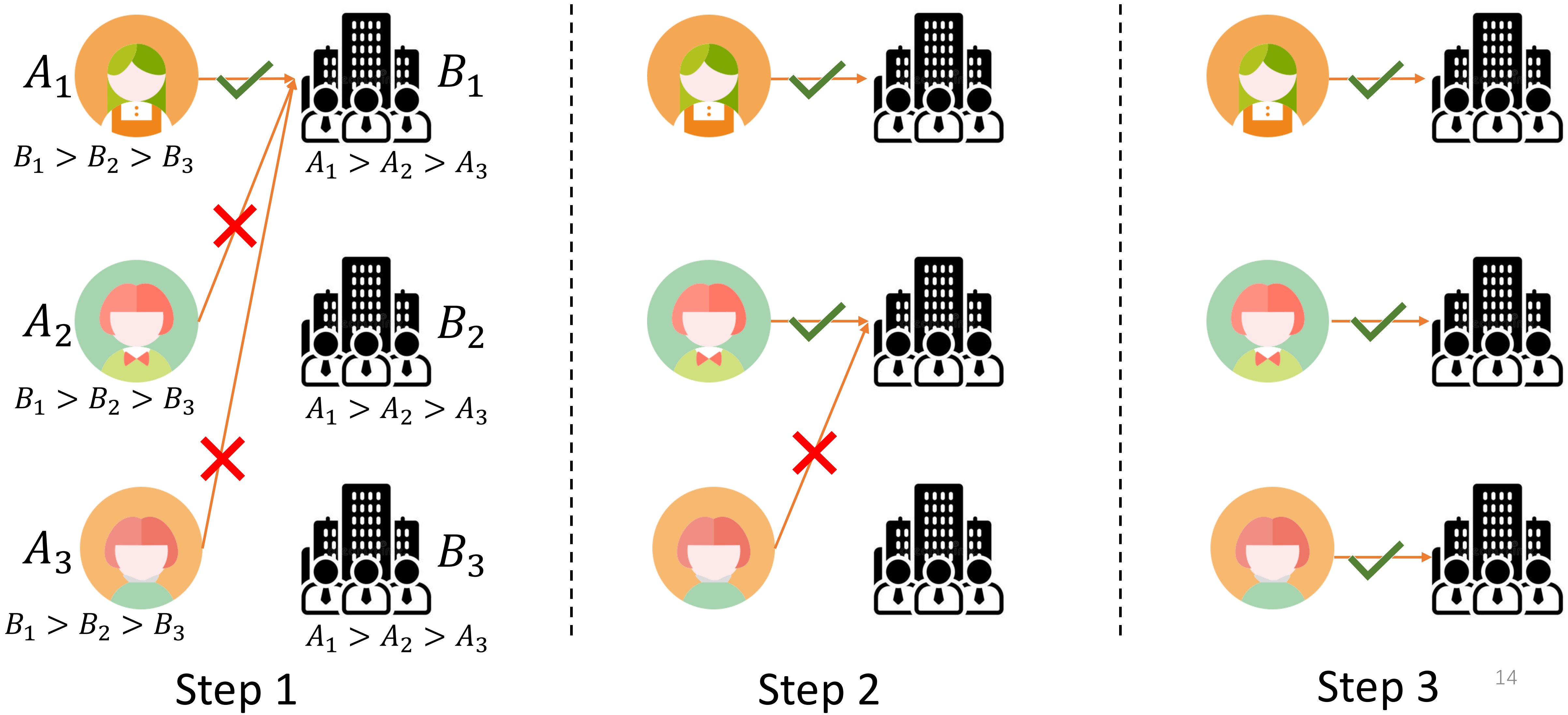


No rejection happens!

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

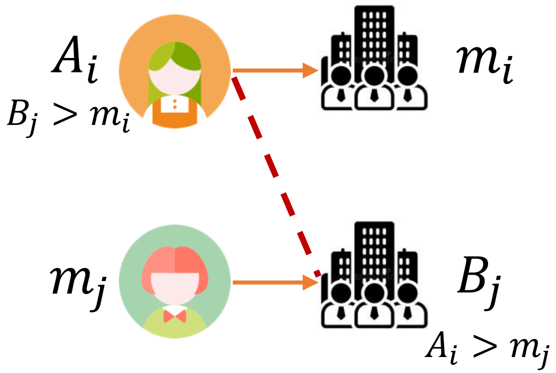
Agents on one side independently propose to agents on the other side according to their preference ranking until no rejection happens

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!

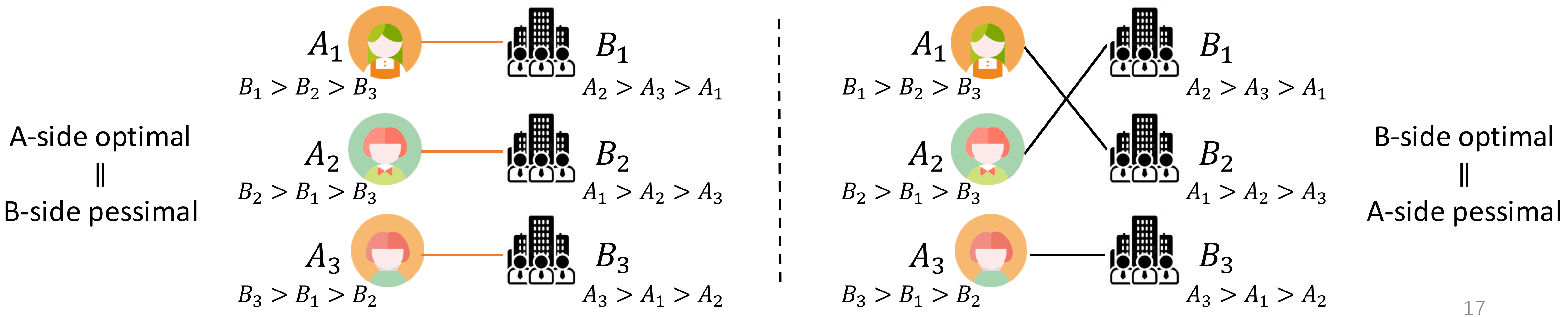
GS properties: Time complexity

- 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, Wang and Li, NeurIPS 2024]

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



Summary of Part 1: Two-sided matching markets

- Introduction to matching markets
- Stable matching
- Gale-Shapley algorithm: Procedure and properties
    - Stability
    - Time complexity
    - Optimality

But agents usually have unknown preferences in practice

Can learn them from iterative interactions !

Part 2: Multi-armed Bandits

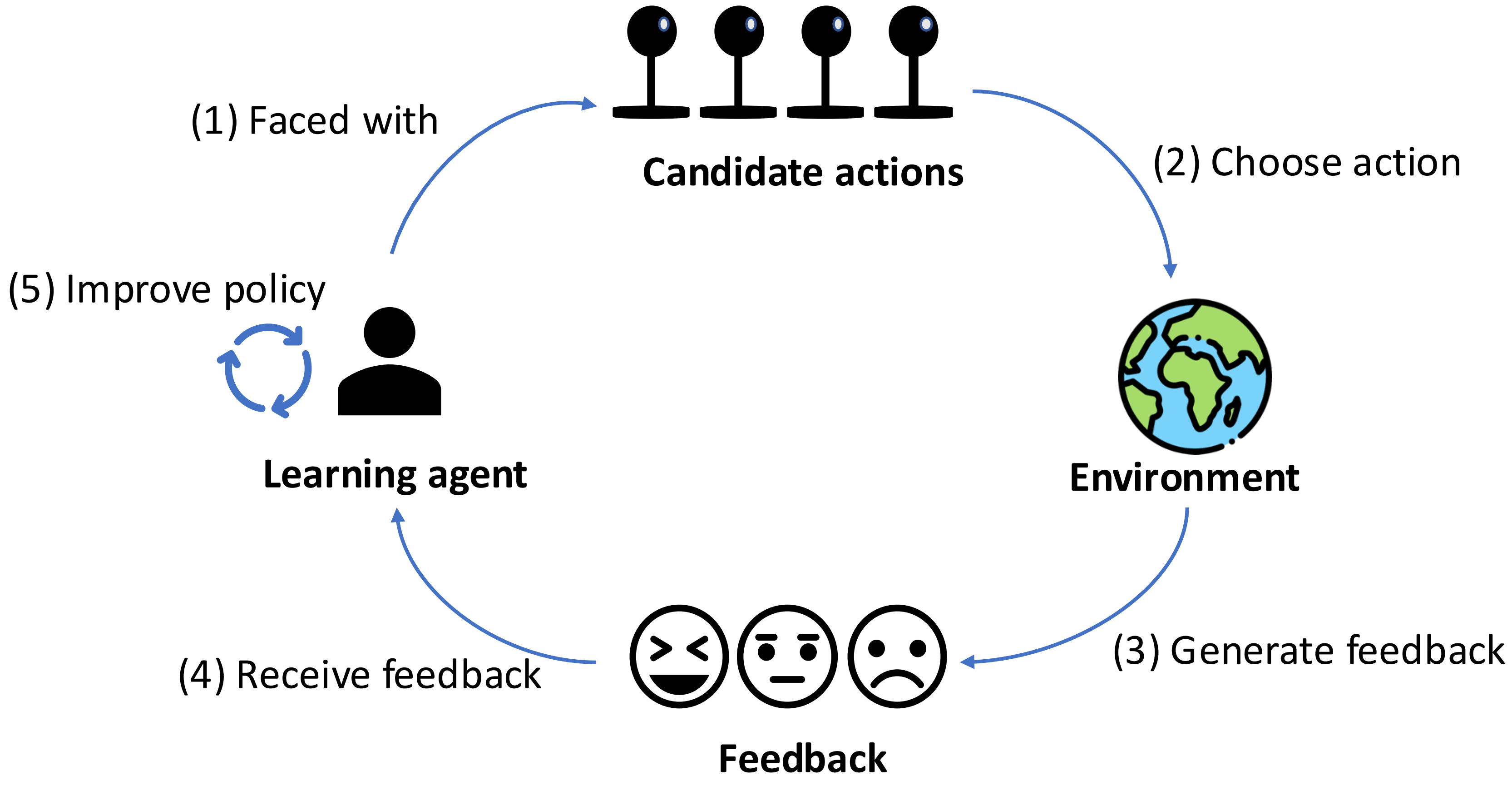
What are bandits? [Lattimore and Szepesvári, 2020]

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

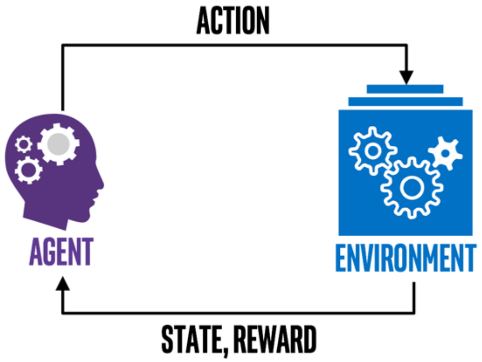


Provide insights for agents in matching markets to learn their unknown preferences through iterative interactions

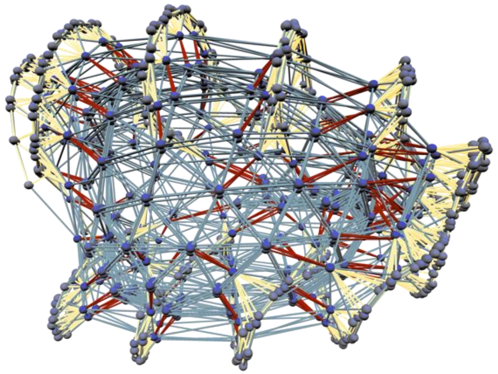
Applications

Recommendation systems
[Li et al., 2010]

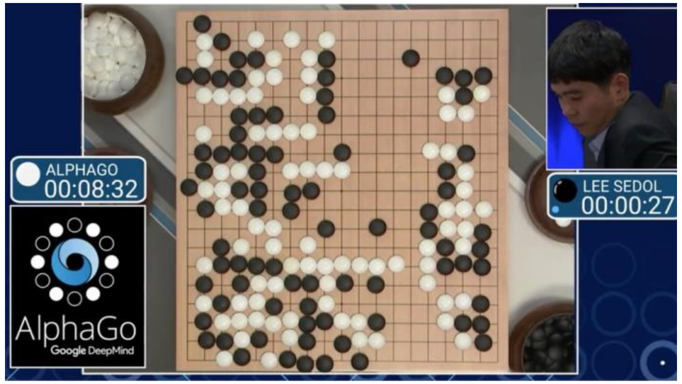
Advertisement placement
[Yu et al., 2016]



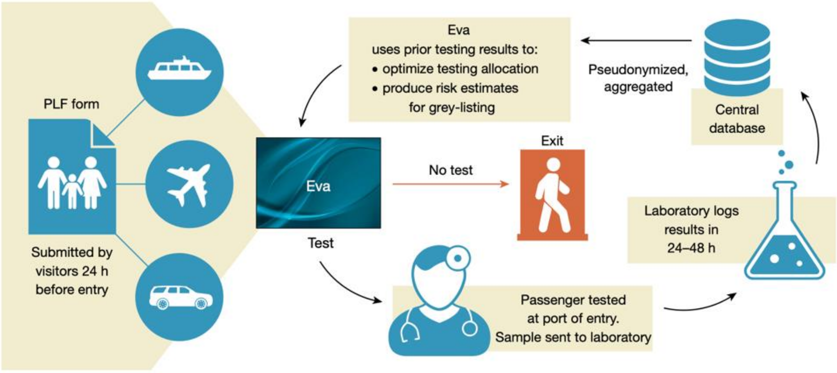
Key part of reinforcement learning
[Hu et al., 2018]



SAT solvers
[Liang et al., 2016]

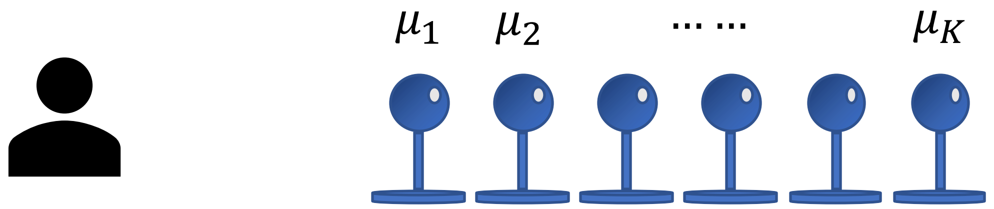


Monte-carlo Tree Search (MCTS) in AlphaGo
[Kocsis and Szepesvári, 2006; Silver et al., 2016]



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 $j^* \in \operatorname{argmax}_j \mu_j$ as the best arm

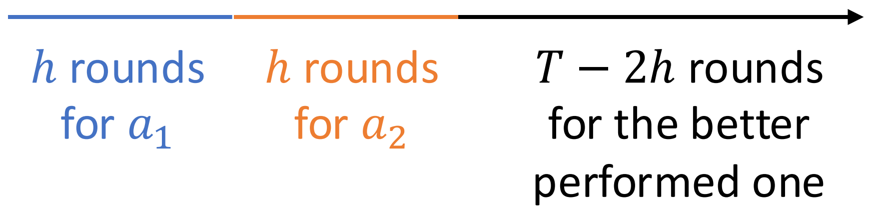
$$Reg(T) = T \cdot \mu_{j^*} - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{A_t}\right]$$

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.)

- Regret analysis:

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

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

Sample mean

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

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

Hoeffding's inequality

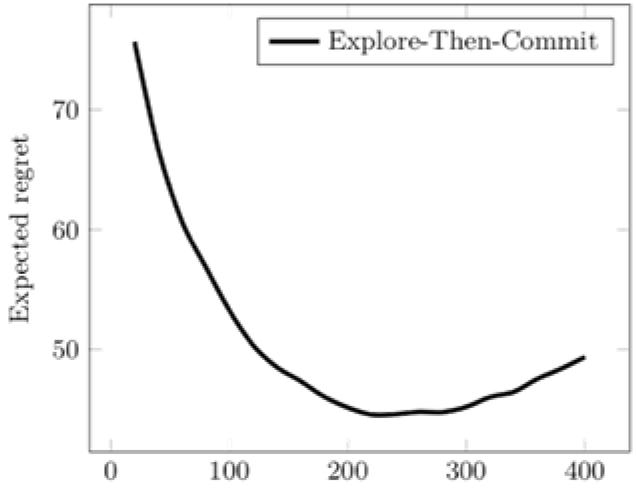
$$\le O\left(\frac{\log T}{\Delta}\right)$$

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 $h = T/10$

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



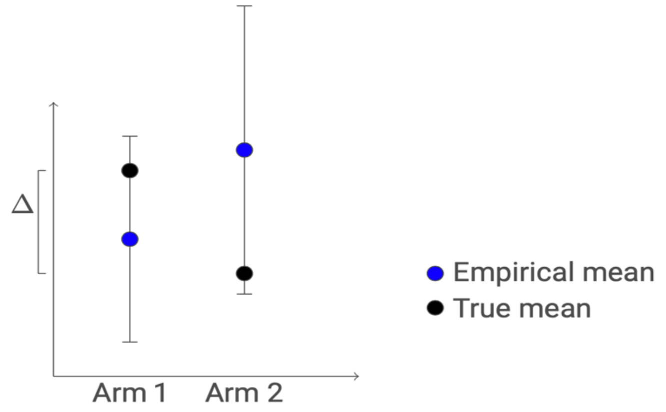
Only with the best choice of h the regret would be smallest

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)

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

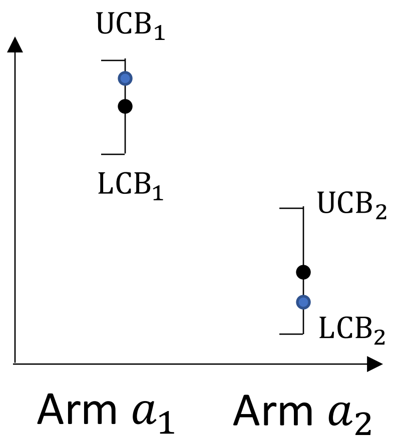
Without knowing Δ

- Regret $O(K\log T/\Delta)$

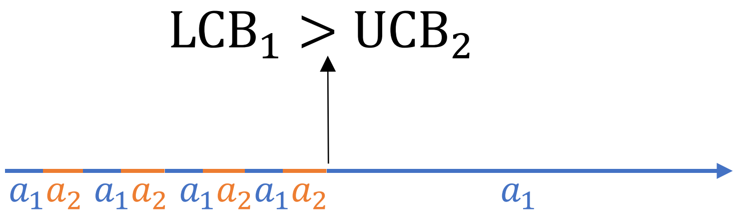
Exploitation Exploration

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
- Regret $O(K \log T/\Delta)$



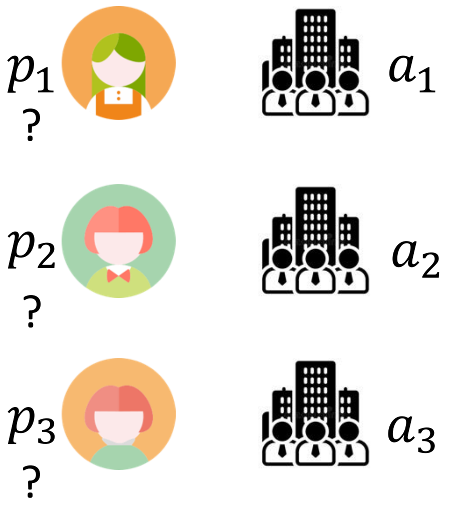
Bandit learning in matching markets [Liu et al., 2020]

- N players: $\mathcal{N} = \{p_1, p_2, \dots, p_N\}$
- K arms: $\mathcal{K} = \{a_1, a_2, \dots, a_K\}$
- $N \leq K$ 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

- 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 $p_{i'}$

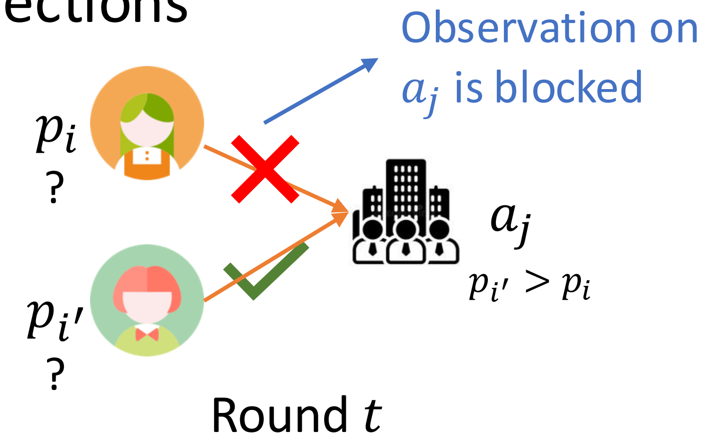
- At each round t, when multiple players select arm a_j
- a_j only accepts the most preferred one $p_i \in \mathrm{argmin}_{p_{i'}: A_{i'}(t)=a_j} \pi_j(p_{i'})$ and rejects others

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

Summary of Part 2: Multi-armed bandits

- Multi-armed bandits (MAB)
  - Applications
  - Explore-then-commit (ETC)
  - Upper confidence bound (UCB)
  - Successive elimination
  - 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

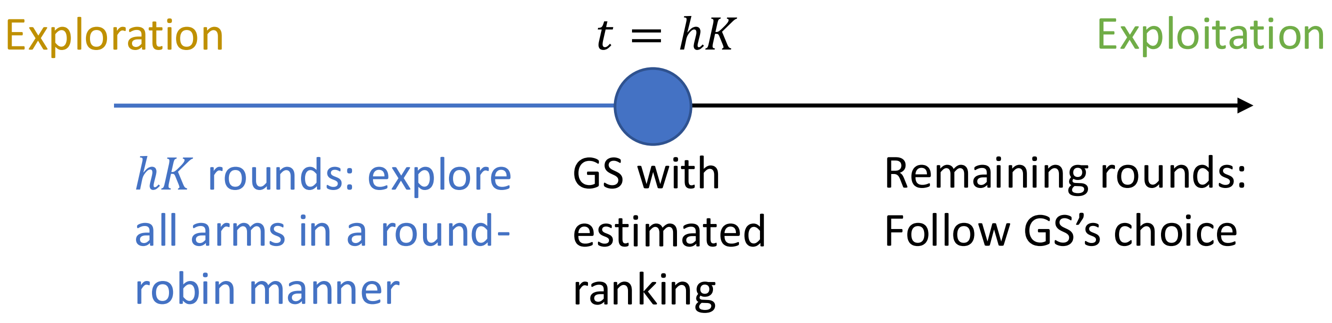
Part 3: Bandit Algorithms in Matching Markets

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
- Other variants

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

- Input: An exploration budget h



- 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 $\Delta_{i,j,j'} = \left|\mu_{i,j} - \mu_{i,j'}\right|$

Larger than 0 due to distinct preferences

- Further define $\Delta = \min_{i,j\neq j'} \Delta_{i,j,j'}$
- 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

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

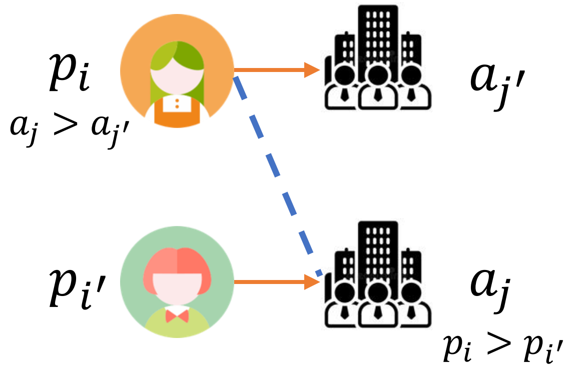
Needs to know Δ

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 $t = 1,2, \ldots,$
  - 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

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

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

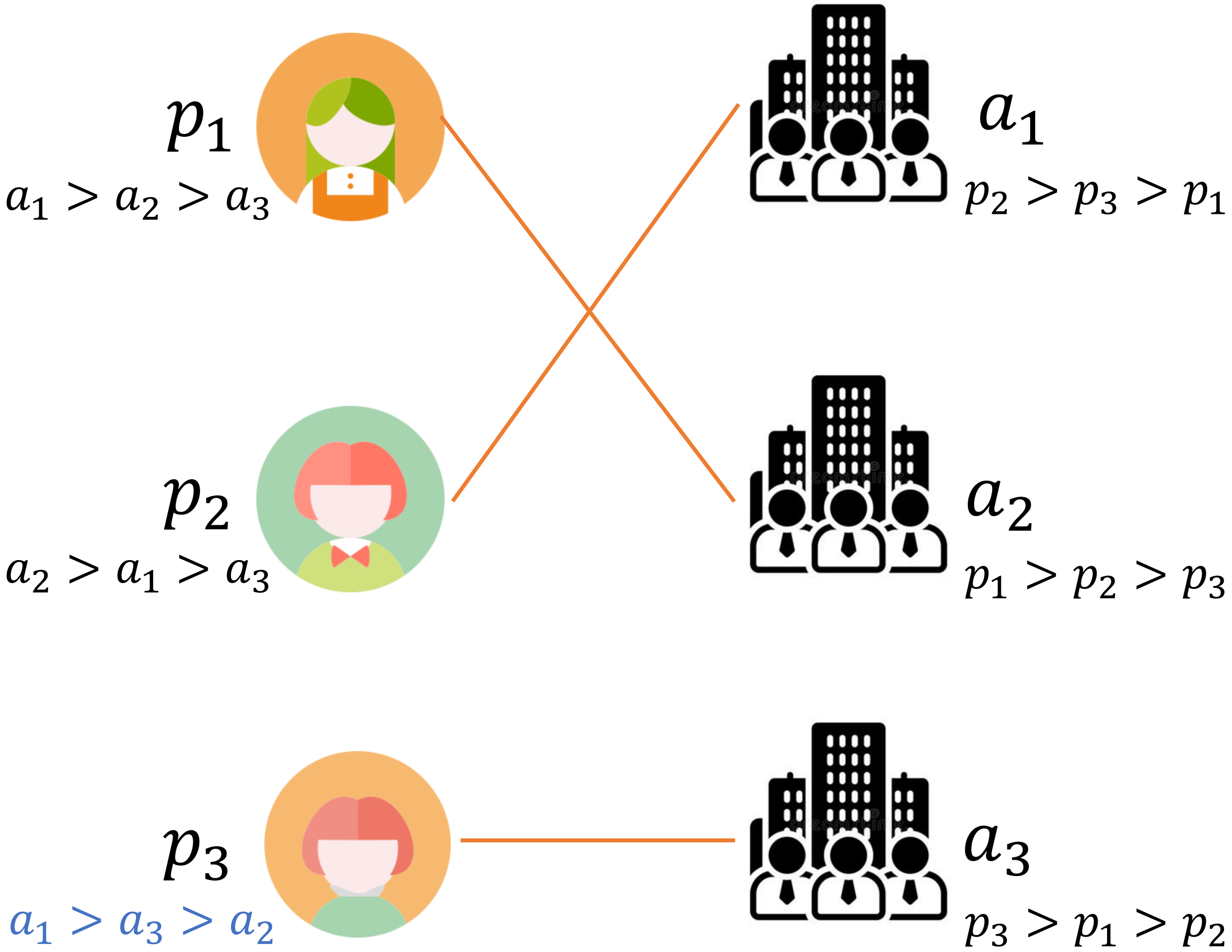
Unique stable matching

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

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*

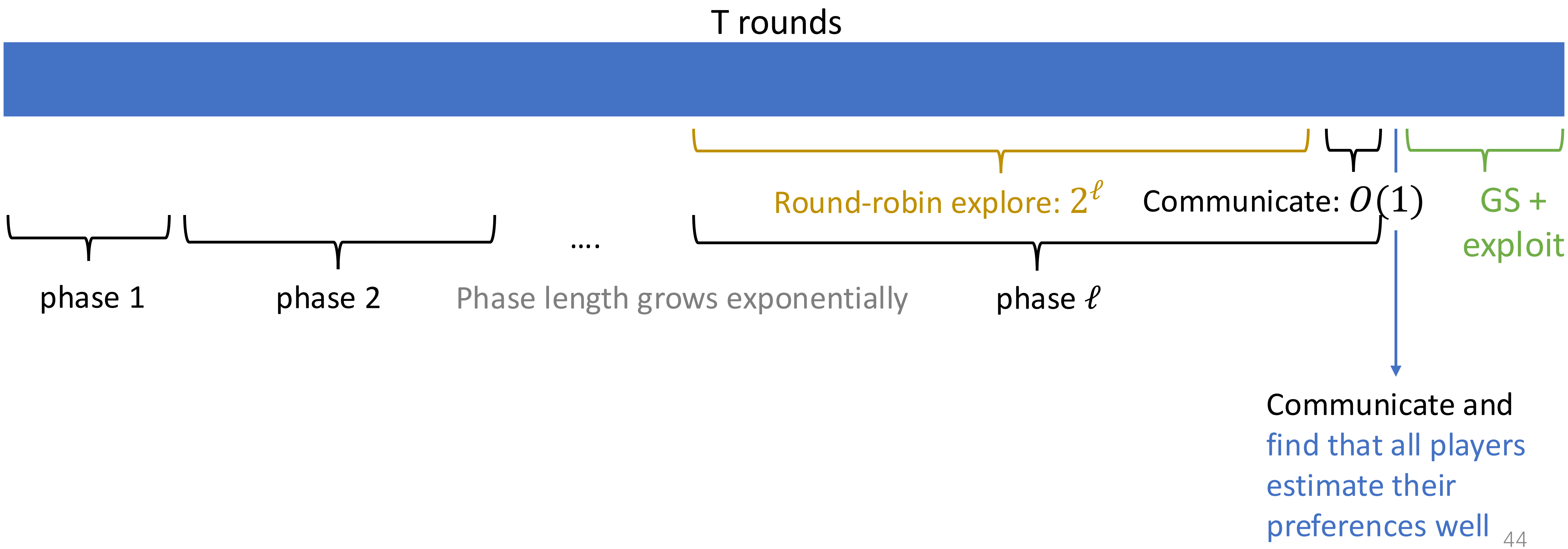
1. When p_1 and p_2 submit the correct rankings

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

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

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

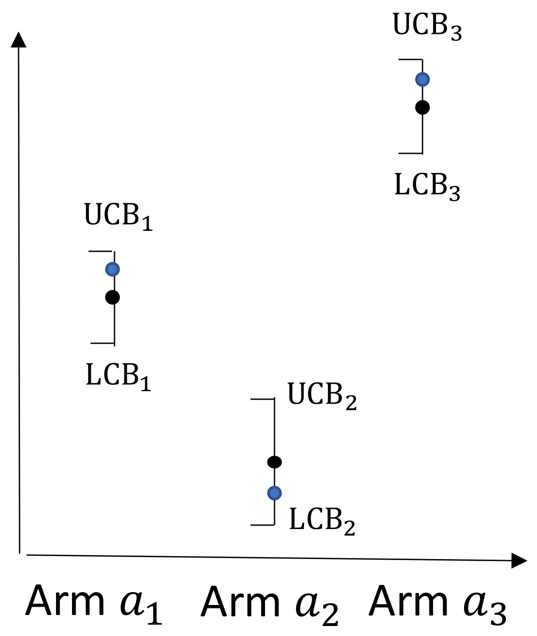


ETGS implementation: Communication

- 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?

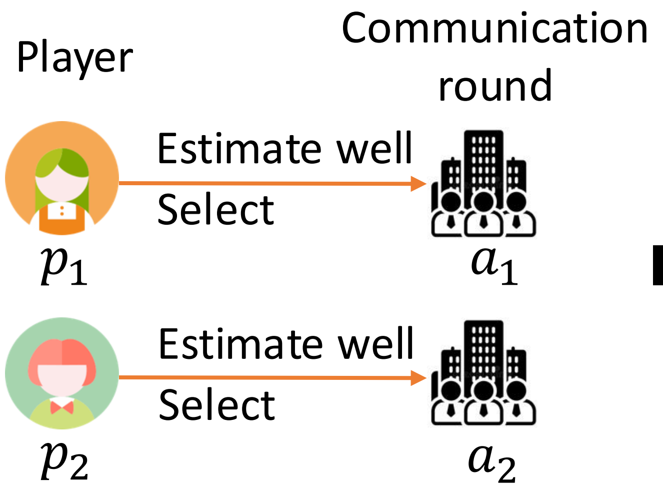


player $p_i's$ preference values

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

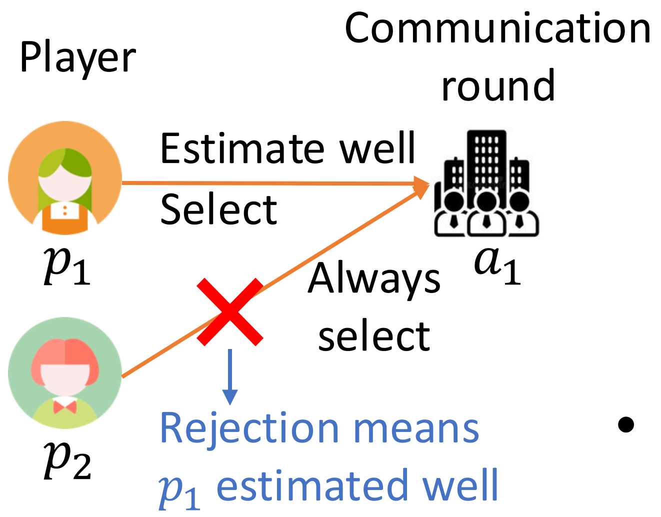


At the communication round, if p_i observes that all players have been matched:

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

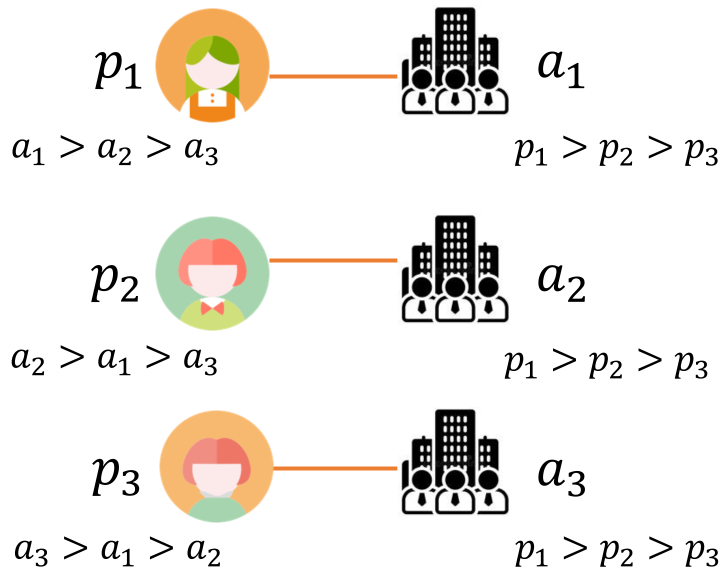
$$\overline{Reg}_i(T) = O\left(\frac{K\log T}{\Delta^2} + \log\left(\frac{K\log T}{\Delta^2}\right)\right)$$

- 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



The minimum regret that p_i may suffer at any round

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

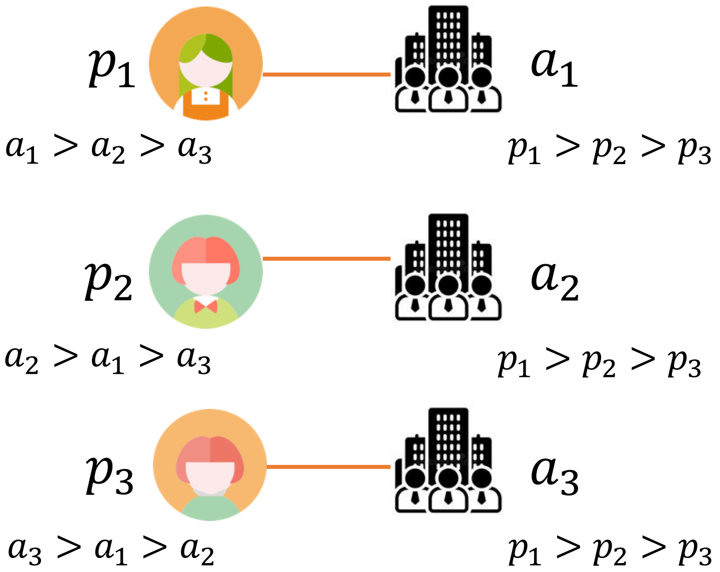
Lower bound (cont.)

- 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

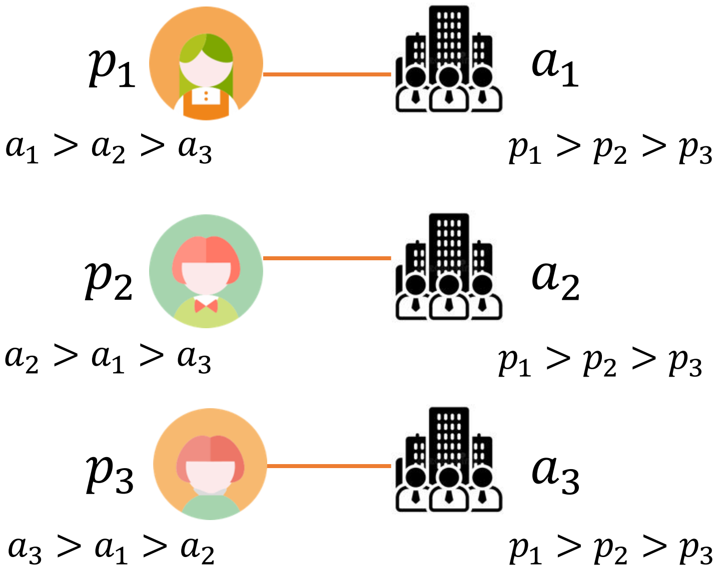
$$\Omega\left(\frac{\log T}{\Delta^2_{i',i',i}}\right) \text{ times}$$

- N higher-priority players cause regret

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

- The stable regret satisfies

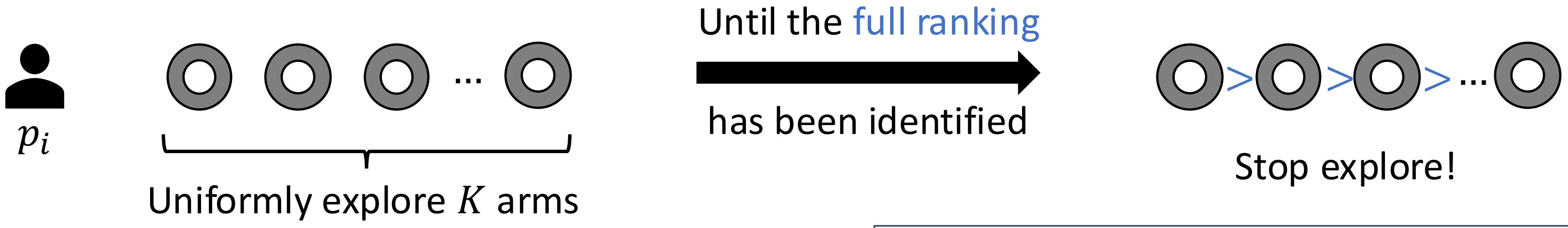
$$\overline{Reg}_i(T) \geq \Omega\left(\max\left\{\frac{N \log T}{\Delta^2}, \frac{K \log T}{\Delta}\right\}\right)$$



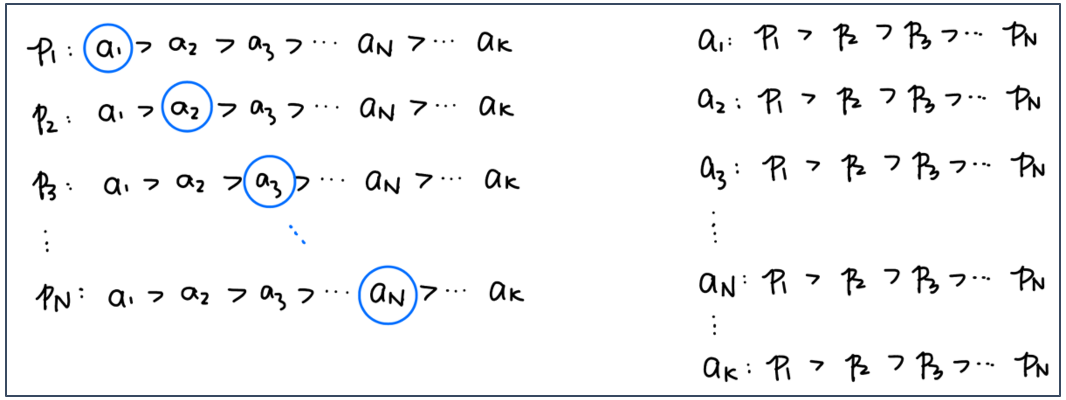
Remark: Δ can be improved as the minimum gap between the player-optimal stable arm and the next preferred one among all players.

Can we close the N and K gap?

- ETGS: $O\left(\frac{K \log T}{\Delta^2}\right)$
- Lower bound: $\Omega\left(\frac{N \log T}{\Delta^2} + \frac{K \log T}{\Delta}\right)$
- Suboptimality: Needs to identify the full ranking among K arms

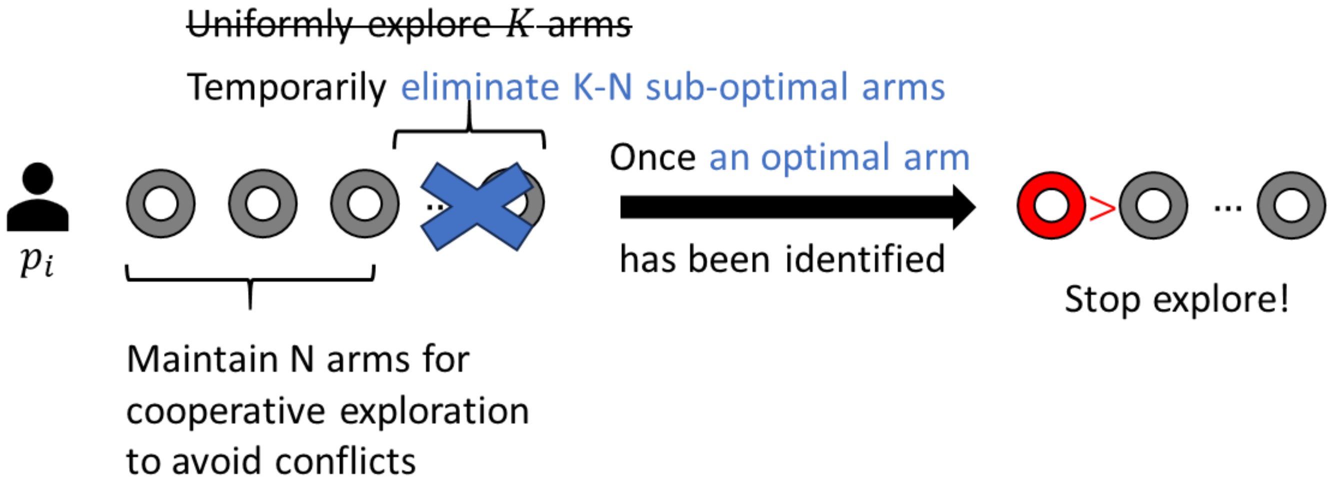


- Key observation:
  N players at most occupy N arms

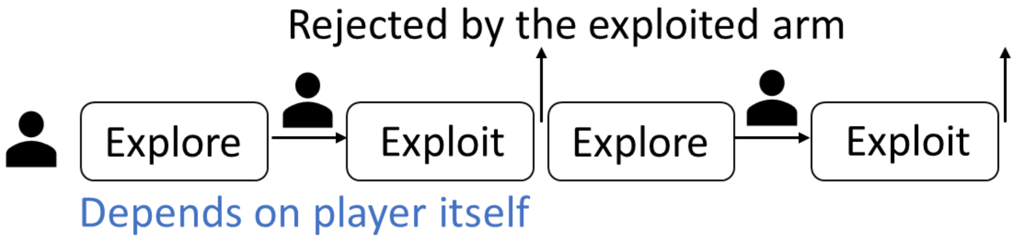


Can we close the N and K gap? (cont.)

- Offline GS + Temporary Elimination



- Independent exploitation



Adaptive ETGS

	Regret bound
Liu *et al.* [19]	$O\left(K\log T/\Delta^2\right) * \#$ $O\left(NK\log T/\Delta^2\right) \#$
Liu *et al.* [20]	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$
Sankararaman *et al.* [27]	$O\left(NK\log T/\Delta^2\right)$ $\Omega\left(\max\left\{N\log T/\Delta^2, K\log T/\Delta\right\}\right)$
Basu *et al.* [4]	$O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{\frac{1}{\varepsilon}}}\right) *$ $O\left(NK\log T/\Delta^2\right)$
Maheshwari *et al.* [21]	$O\left(CNK\log T/\Delta^2\right)$
Kong *et al.* [17]	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$
Zhang *et al.* [30]	$O\left(K\log T/\Delta^2\right) *$
Kong and Li [16]	$O\left(K\log T/\Delta^2\right) *$
Ours NeurIPS 2024	$O\left(N^2\log T/\Delta^2 + K\log T/\Delta\right) *$ $O\left(N\log T/\Delta^2 + K\log T/\Delta\right) \#$

No dependence on K in the main term

Other setting variants

- Many-to-one matching markets
- Strategic behaviors
- Contextual information and indifferent preferences
- Non-stationary preferences
- Two-sided/multi-sided unknown preferences
- Markov matching markets
- Multi-sided matching markets

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
- SOTA result
- Other variants

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

Thanks!
&
Questions?

Shuai Li

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

Credit: Some images are from Flaticon.com

References 1:

- 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.
- Gale, David, and Lloyd S. Shapley. "College admissions and the stability of marriage." The American Mathematical Monthly 69.1 (1962): 9-15.
- Kong, Fang, Zilong Wang and Shuai Li. "Improved Analysis for Bandit Learning in Matching Markets." NeurIPS. 2024.
- 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.
- Yu, Baosheng, Meng Fang, and Dacheng Tao. "Linear submodular bandits with a knapsack constraint." Proceedings of the AAAI Conference on Artificial Intelligence. 2016.

References 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.
- 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.
- 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.
- Bastani, Hamsa, et al. "Efficient and targeted COVID-19 border testing via reinforcement learning." Nature 599.7883 (2021): 108-113.

References 3:

- Garivier, Aurélien, Tor Lattimore, and Emilie Kaufmann. "On explore-then-commit strategies." Advances in Neural Information Processing Systems 29 (2016).
- Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." Machine learning 47 (2002): 235-256.
- Audibert, Jean-Yves, and Sébastien Bubeck. "Best arm identification in multi-armed bandits." COLT-23th Conference on learning theory-2010. 2010.
- Liu, Lydia T., Horia Mania, and Michael Jordan. "Competing bandits in matching markets." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
- 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 4:

- 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.
- 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.
- 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.