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

上海创智学院
SHANGHAI INNOVATION INSTITUTE
2024

Bandit Learning in Matching Markets

Shuai Li

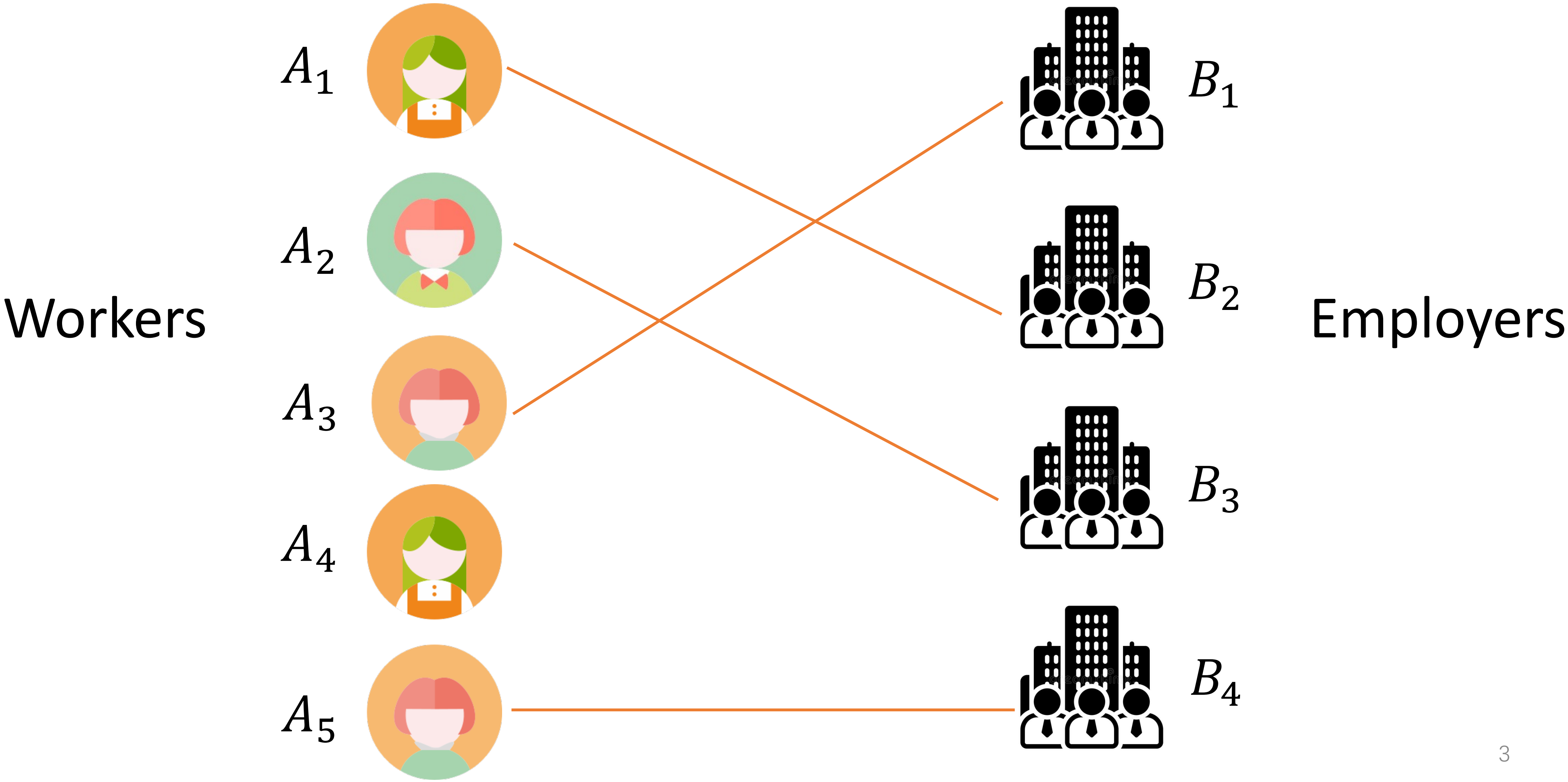
Shanghai Jiao Tong University

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

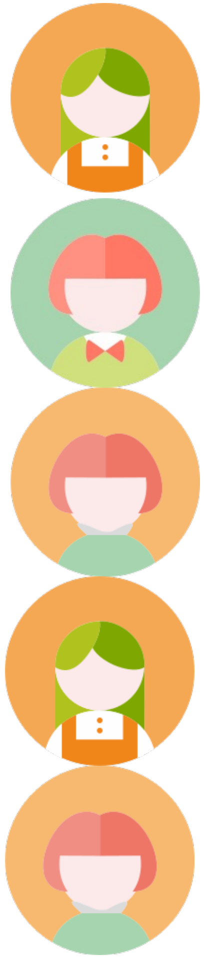
Matching market has two sides



Both sides have preferences over the other side

Worker side

Based on payment or prior familiarity of the task



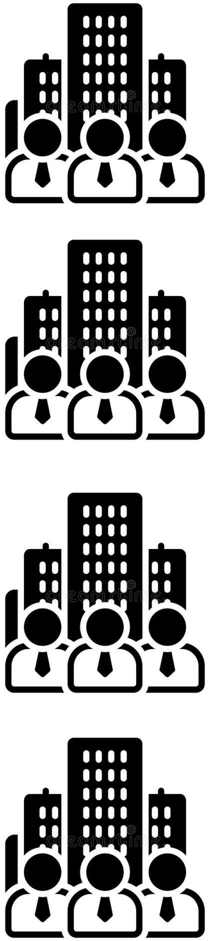
A_1 : $B_2 > B_3 > B_1 > B_4$

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

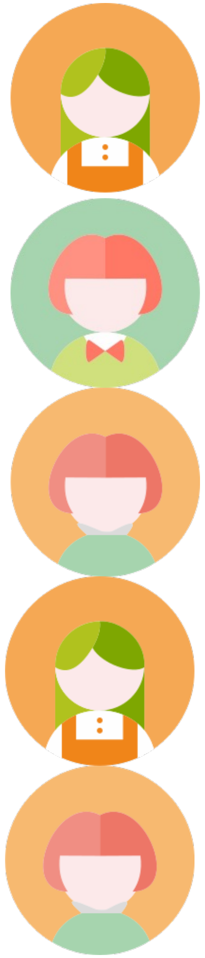
A_1

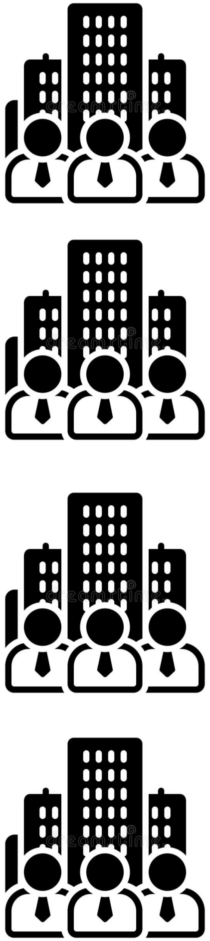
A_2

A_3

A_4

A_5





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

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

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

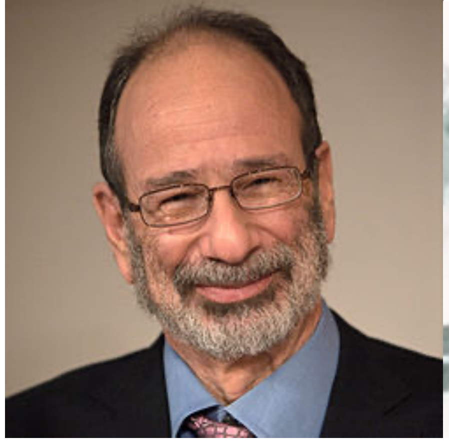
$B_4 : A_4 > A_5 > A_1 > A_2 > A_3$

Employer side

Based on the skill levels of workers

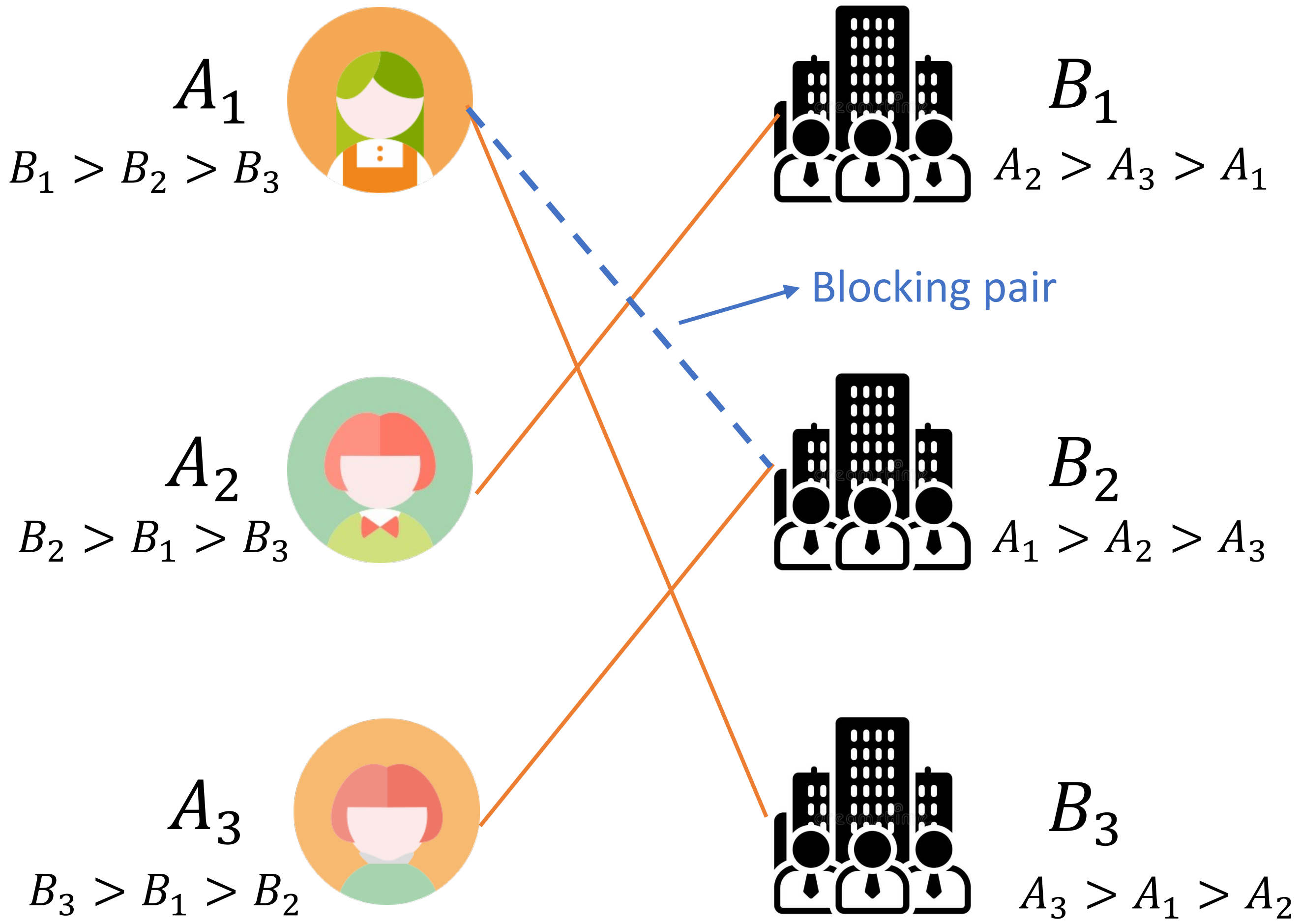
Stable matching

Nobel Memorial Prize
in Economic Sciences
2012



Alvin E. Roth

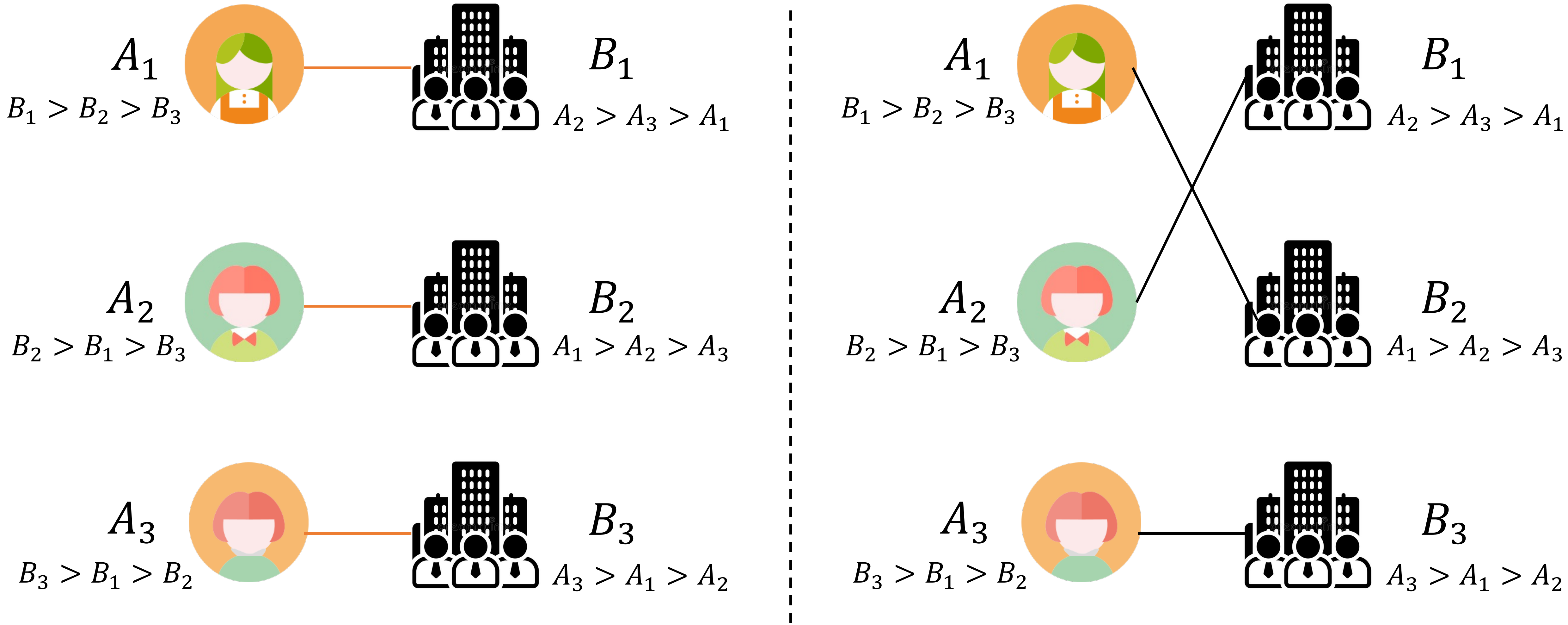
Lloyd Shapley



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.

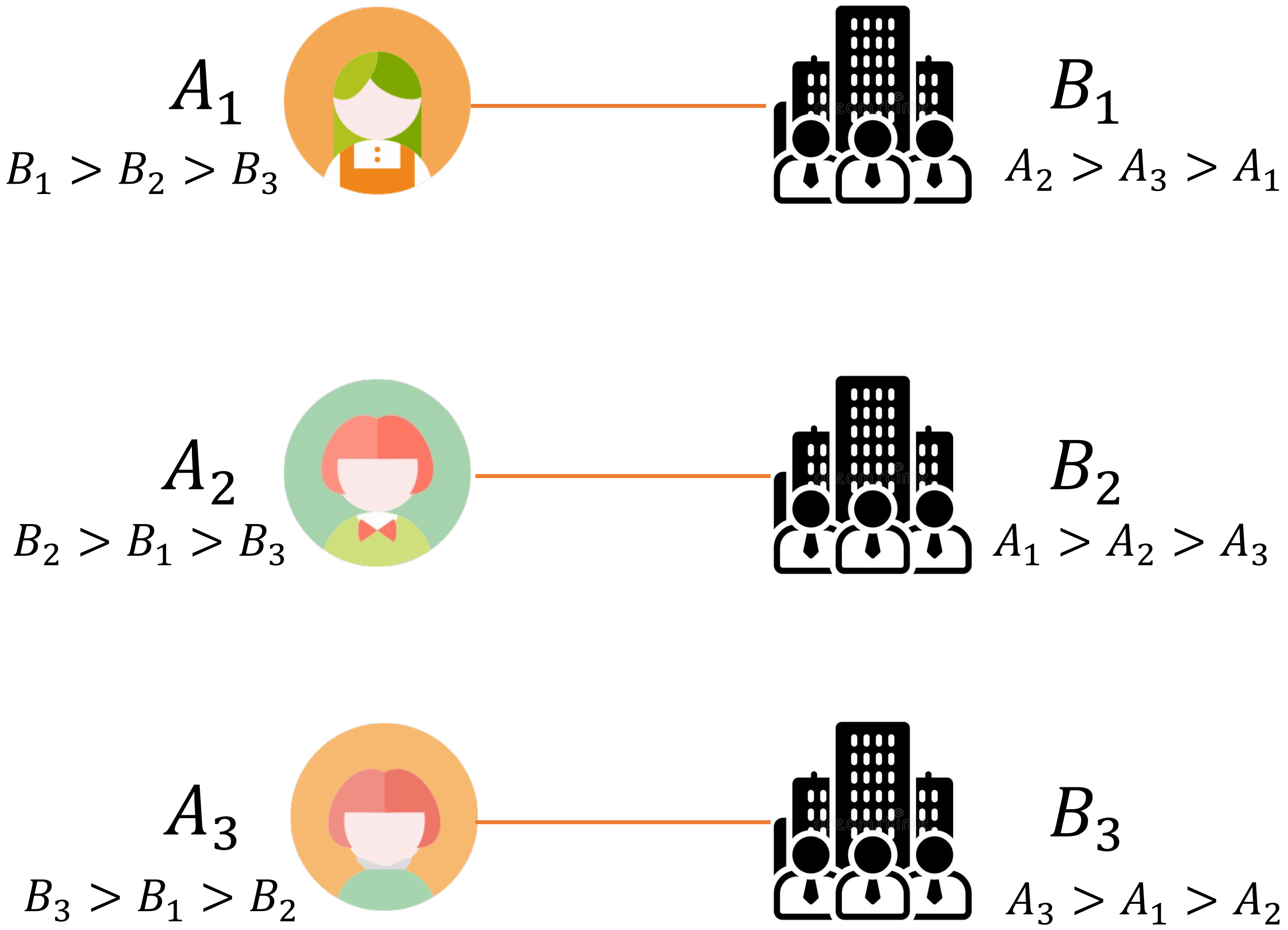
May be more than one stable matchings



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

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

A-side optimal stable matching[1]

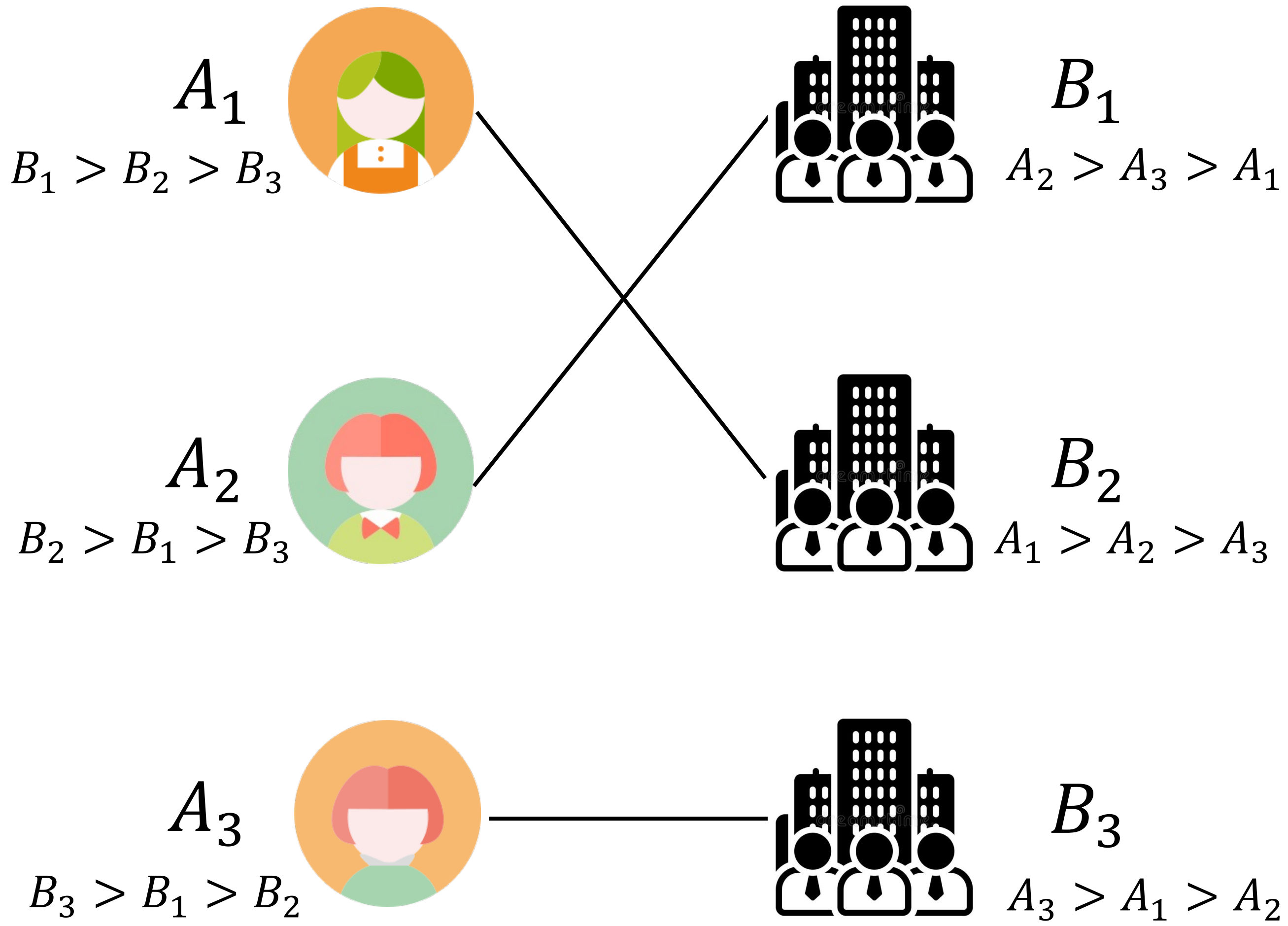


Each agent on A-side is matched with the most preferred partner among all stable matchings

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

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

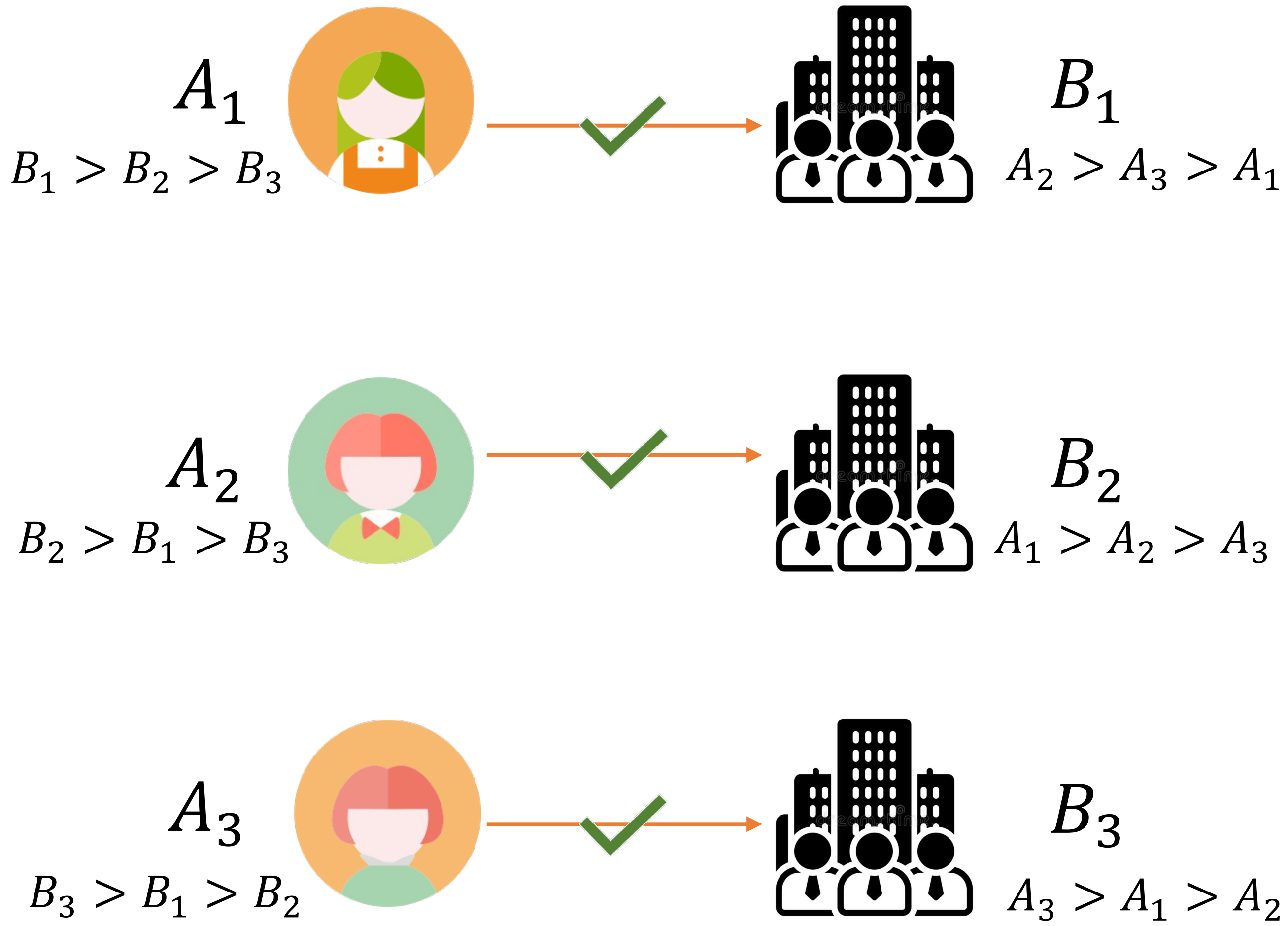
A-side pessimal stable matching



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?



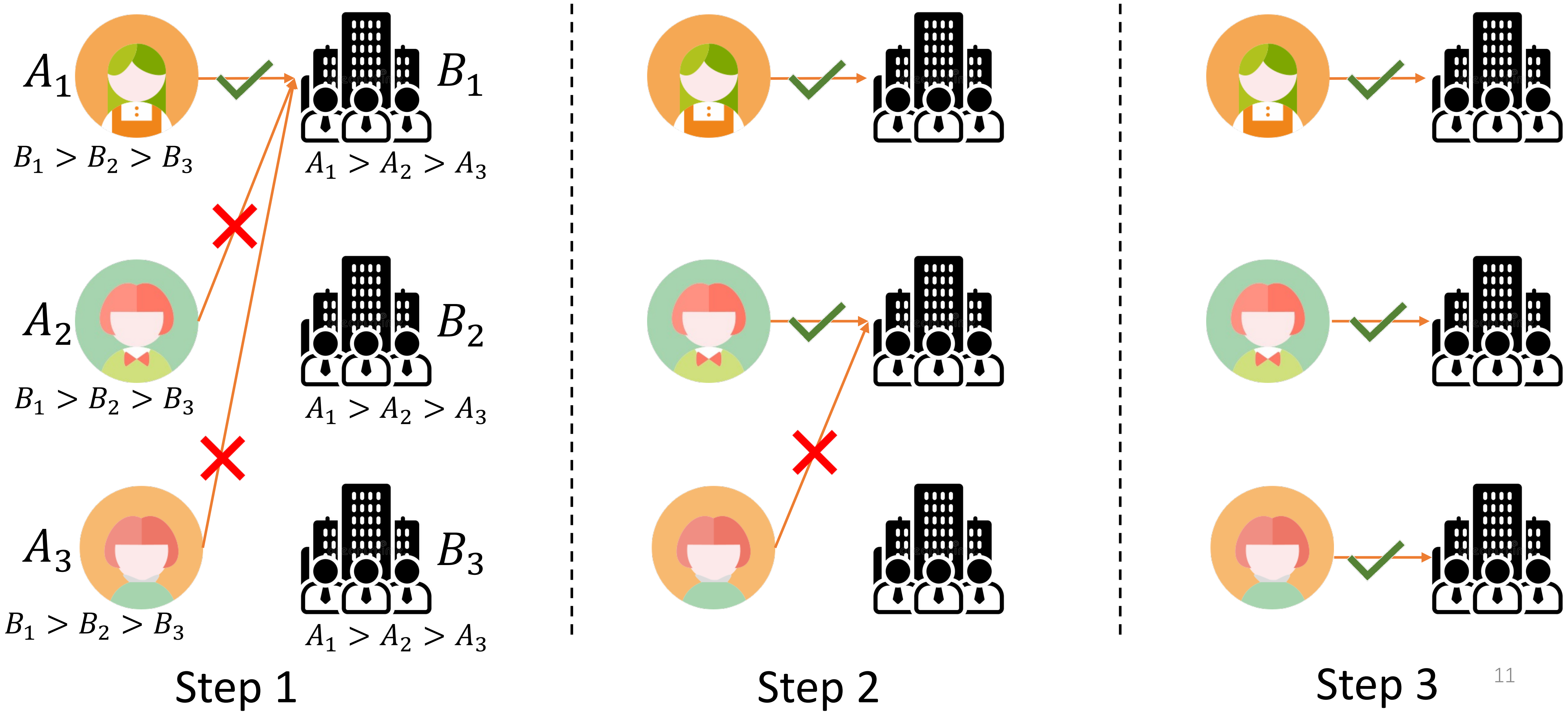
No rejection happens!

Gale-Shapley (GS) algorithm

[Gale and Shapley (1962)]

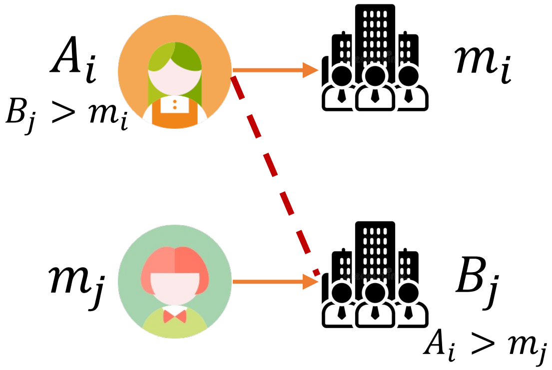
Agents on one side independently propose to agents on the other side according to their preference ranking until no rejection happens

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!

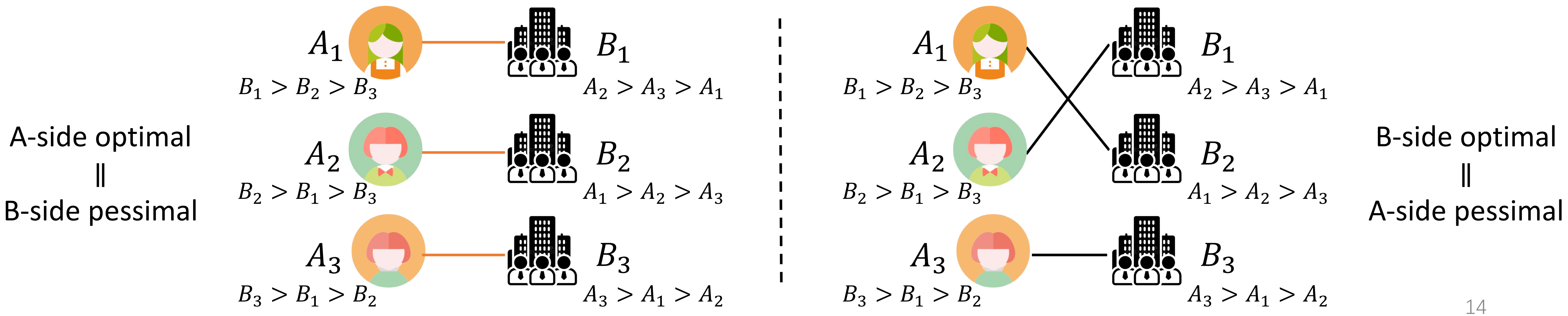
GS properties: Time complexity

- 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



But agents usually have unknown preferences in practice

Can learn them from iterative interactions !

Bandit learning in matching markets

[Liu et al., AISTATS 2020]

Michael Jordan

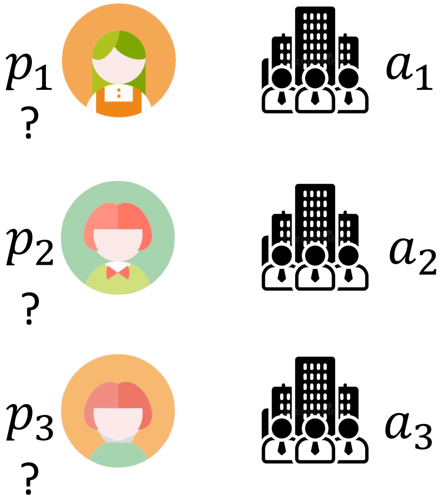
- N players: $\mathcal{N} = \{p_1, p_2, \dots, p_N\}$
- K arms: $\mathcal{K} = \{a_1, a_2, \dots, a_K\}$
- $N \le K$ to ensure players can be matched
- $\mu_{i,j} > 0$: (unknown) preference of player p_i towards arm a_j
- For each player p_i
    - $\{\mu_{i,j}\}_{j \in [K]}$ forms its preference ranking
    - 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

$$\mathbb{E}\left[X_{i,A_i(t)}(t)\right] = \mu_{i,A_i(t)}$$

    - If p_i is rejected: receive $X_{i,A_i(t)}(t) = 0$

Satisfaction over this matching experience

When would p_i be rejected?



For simplicity, assume arms know their preferences

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

Multi-armed bandits (MAB)

[Lattimore and Szepesvári, 2020]

corresponds to
N=1 player setting

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

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

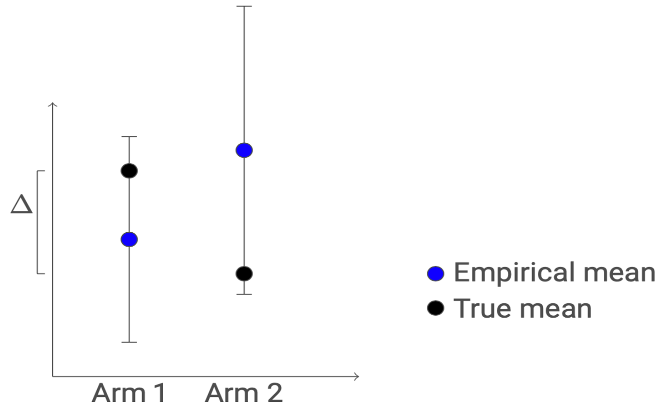
- With high probability $\geq 1-\delta$

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

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

Upper confidence bound (UCB)

Exploitation

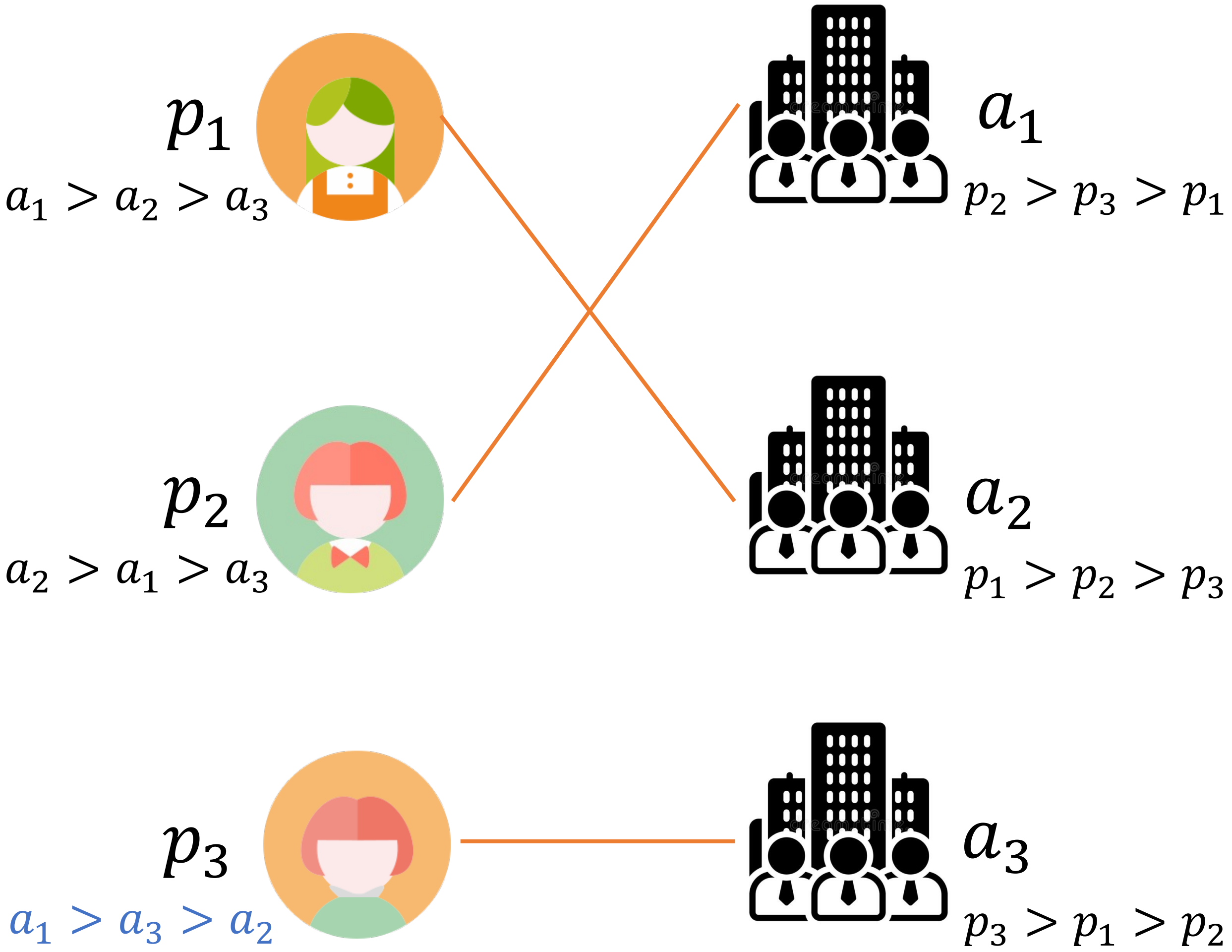
Exploration

Without knowing Δ

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

Regret type	Regret Bound	Communication type	Strategy-proofness	References
Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Centralized, known Δ	?	[Liu et al., 2020]
Player-pessimal	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Centralized	?	
	$O\left(\frac{N^5K^2\log^2 T}{\rho^{N^4}\Delta^2}\right)$	Decentralized, observed matching outcomes	No	[Liu et al., 2021]
				[Kong et al., 2022]
Unique	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Decentralized	?	[Sankararaman et al., 2021; Basu et al., 2021; Maheshwari et al., 2022]
	$O\left(\frac{N\log T}{\Delta^2}\right)$	Centralized	?	[Wang and Li, 2024; Kong et al., 2024]
Optimal stable bandits (Unique)	$\Omega\left(\frac{N\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$	Decentralized	/	[Sankararaman et al., 2021]
Player-optimal	$O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{1/\varepsilon}}\right)$	Decentralized	?	[Basu et al., 2021]
	$O\left(\frac{K\log T}{\Delta^2}\right)$	Decentralized, observed matching outcomes	No	[Kong and Li, 2023]
		Decentralized	No	[Zhang et al., 2022]
	$O\left(\frac{N^2\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$	Decentralized	Yes	[Kong et al., 2024]

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*

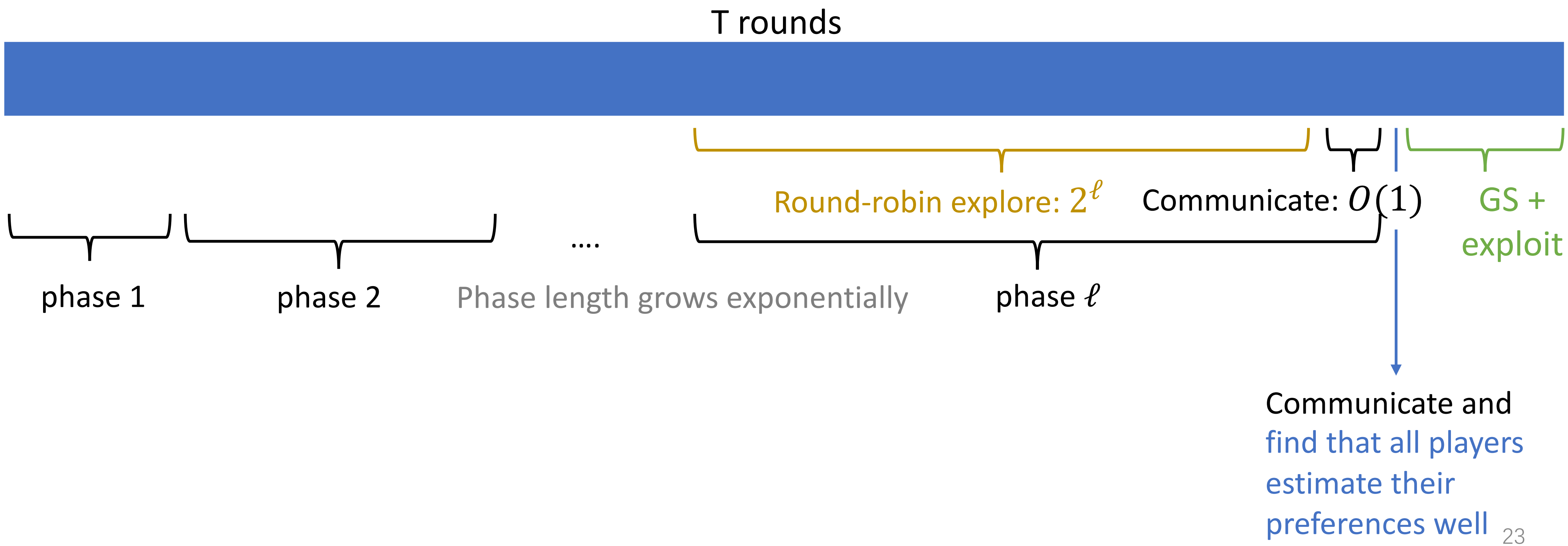
1. When p_1 and p_2 submit the correct rankings

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, SODA 2023]

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



ETGS: Regret analysis [Kong and Li, SODA 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

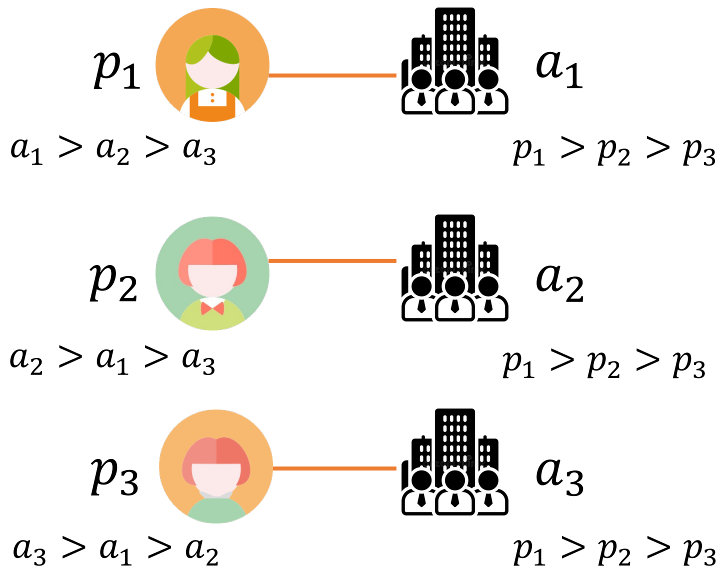
$$\overline{Reg}_i(T) = O\left(\frac{K\log T}{\Delta^2} + \log\left(\frac{K\log T}{\Delta^2}\right)\right)$$

- 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., AISTATS 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 $p_1, p_2 \ldots \ldots p_{i-1}$ than p_i
  - $T_{i,j}$: the number of times that p_i selects a_j



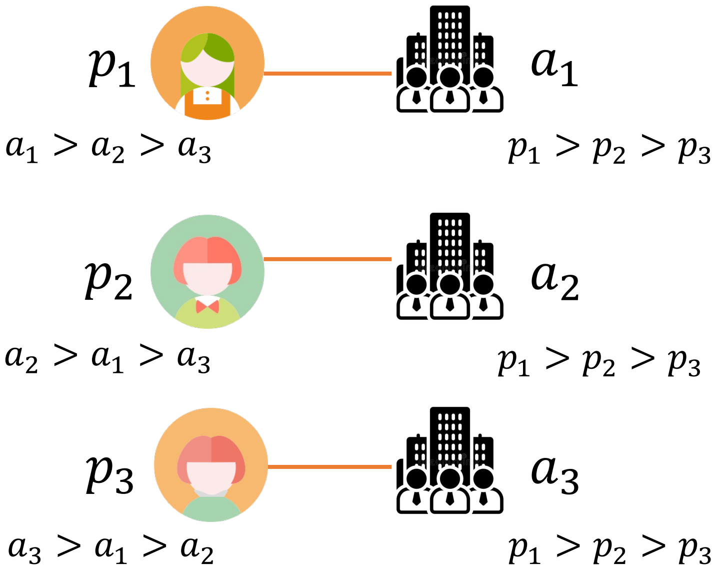
The minimum regret that p_i may suffer at any round

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

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

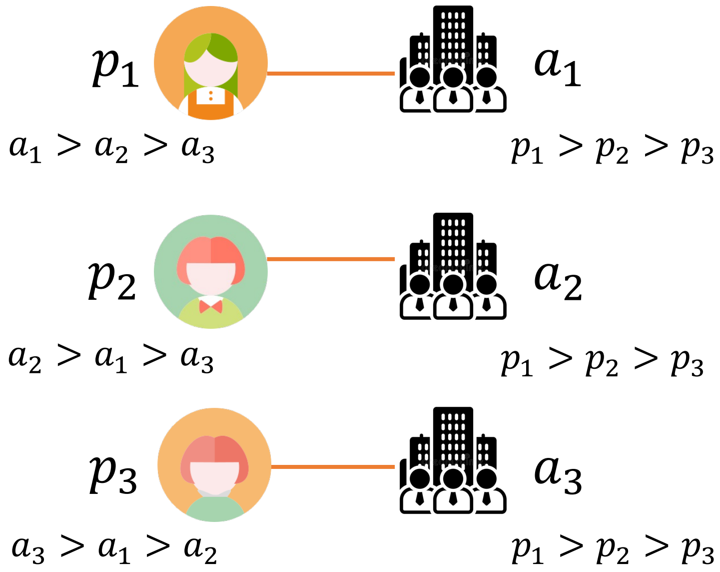
$$\Omega\left(\frac{\log T}{\Delta^2_{i',i',i}}\right) \text{ times}$$

- 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

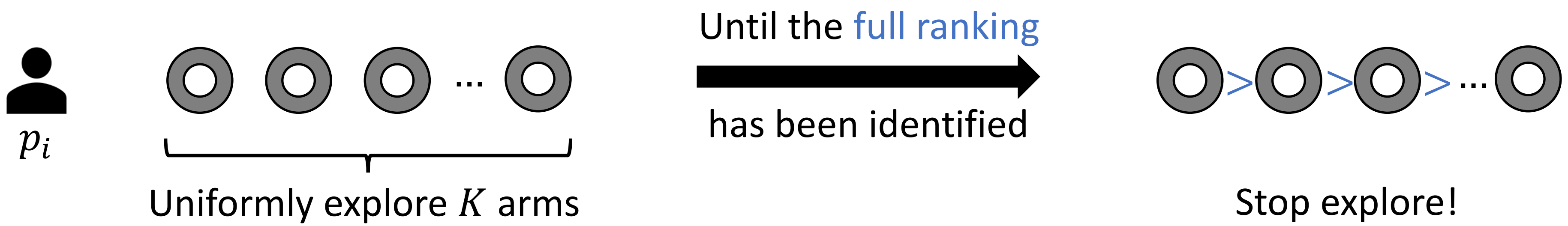
$$\overline{Reg}_i(T) \geq \Omega\left(\frac{N\log T}{\Delta^2} + \frac{K\log T}{\Delta}\right)$$



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

Sub-optimality of ETGS

- Needs to identify the full ranking among K arms

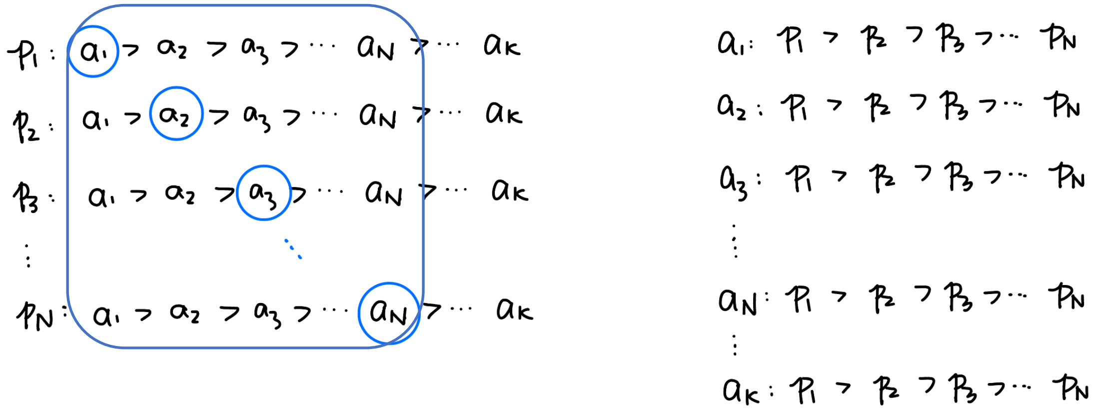


- But for the Gale-Shapley algorithm, what is the real complexity to find the optimal stable matching?
  - Whether it is necessary to determine the full ranking over K arms

Key observation of GS properties

Could the dependence of K be improved as N?

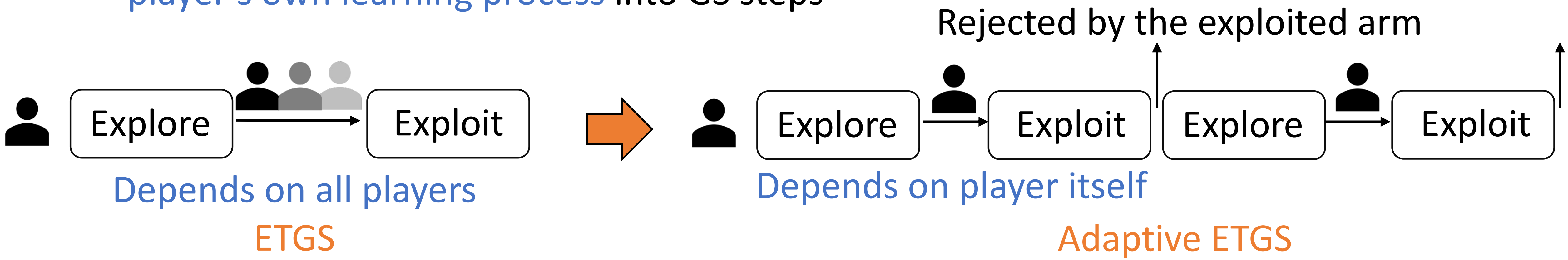
- The optimal stable arm must be the first N-ranked
  - The player moves to the next arm only if this arm is occupied by another player
  - N players at most occupy N arms



- The GS algorithm proceeds for at most N^2 steps
  - Those N arms can reject each of N players for at most once

Improvement: Adaptive ETGS [KL, AAAI 2024]

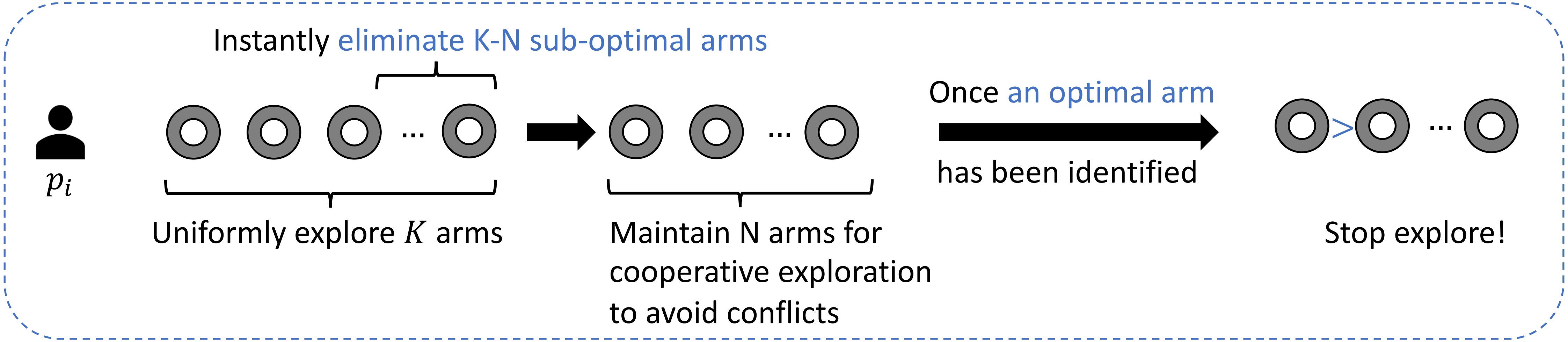
- Idea: Instead of starting GS + exploitation with all players' agreement, integrating each player's own learning process into GS steps



- Players cooperatively explore arms in a round-robin manner
- Once a player identifies the most preferred one, starts exploiting this arm
- If the exploited arm is occupied by a higher-priority player (the arm "rejects" the player)
  - Explore the next most preferred arm (enter the next step of GS)

Adaptive ETGS: Regret [KWL, NeurIPS 2024]

- Arrangement of the exploration process



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

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

  - Each player explore N arms round-robin to determine next best arm (N in total)

Δ is the minimum preference gap between any arms that have higher ranking than the arm after the player-optimal stable one.

Key Observation: Stable position is enough

- Lemma. When $\{\bar{m}_i : i\}$ are ranked correct, GS will return $\bar{m}$
- Proof Sketch.
  1. Any reordering of arms before stable arm can be divided into swaps
  2. A swap in ranking results in a swap in rejection
  3. After the rejection, the GS states are the same, thus will return the same $\bar{m}$

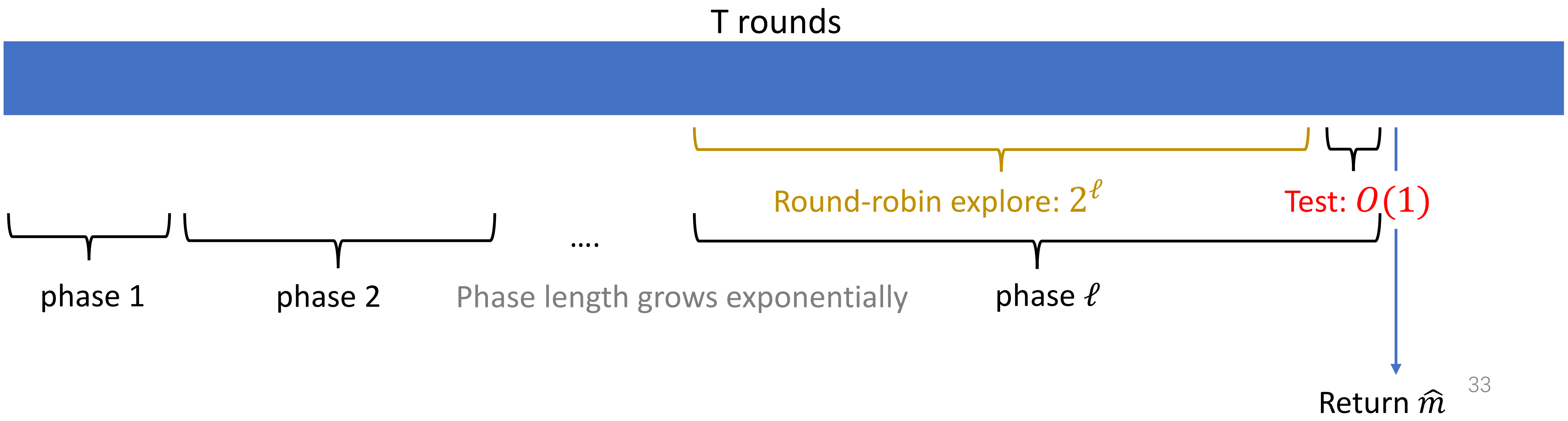
μ_i: $k > k' > \bar{m}_i > \cdots$

$\hat{\mu}_i$

- No need to learn all top rankings. Only need to learn the stable boundary

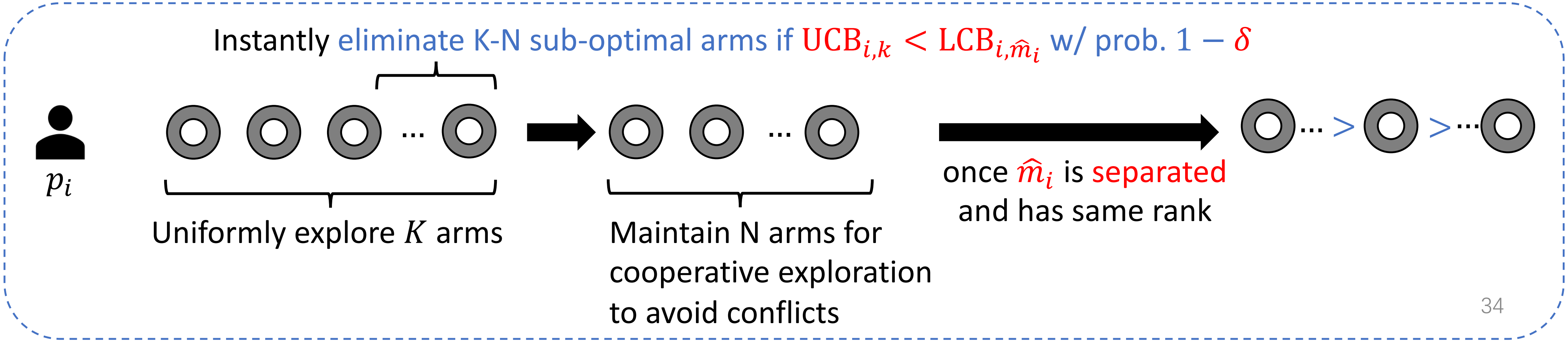
Hypothesis test: Guess the stable boundary

1. Modify ETGS to adaptively test whether $\{\overline{m}_i : i\}$ exists w/ moderate prob.
    - Run GS on $\{\hat{\mu}_{i,k}\}$ to get $\widehat{m}$
    - Check whether $\widehat{m}$ is separated i.e. $\mathrm{LCB}_{i,k} > \mathrm{UCB}_{i,\widehat{m}_i} \vee \mathrm{LCB}_{\widehat{m}_i} > \mathrm{UCB}_{i,k}$ w.p. $1-p$
    - $\Longrightarrow \mathbb{P}(\widehat{m} = \overline{m}) > 1-p$ and the exploration ends in $O\left(\frac{K \log 1/p}{\Delta^2}\right)$ rounds



Hypothesis test: Confirmation

1. Modify ETGS to adaptively test whether $\{\bar{m}_i : i\}$ exists w/ moderate prob.
2. Boost $\mathbb{P}(\hat{m} = \bar{m}) > 1 - p$ to $\mathbb{P}(\hat{m} = \bar{m}) > 1 - \delta$
   - Run *AdaptiveETGS-Explore* on the $\hat{m}$'s positions till $\hat{m}$ is separated and confirmed
   - If $\hat{m}$ has the same rank, $\hat{m} = \bar{m}$, this causes $O\left(\frac{N\log T}{\Delta^2} + \frac{K\log T}{\Delta}\right)$ (main order) regret
   - If $\hat{m}$'s rank change, $\hat{m} \neq \bar{m}$, it results in $O\left(p \cdot \frac{K\log T}{\Delta^2}\right)$ regret explore w/ wrong $\hat{m}$



Hypothesis test: Exploitation

1. Modify ETGS to adaptively test whether $\{\bar{m}_i : i\}$ exists w/ moderate prob.
2. Boost $\mathbb{P}(\hat{m} = \bar{m}) > 1 - p$ to $\mathbb{P}(\hat{m} = \bar{m}) > 1 - \delta$
   - Run *AdaptiveETGS-Explore* on the $\hat{m}$'s positions till $\hat{m}$ is separated and confirmed
   - When $\hat{m} = \bar{m}$, this causes $O\left(\frac{N \log T}{\Delta^2} + \frac{K \log T}{\Delta}\right)$ - main order regret
   - When $\hat{m} \neq \bar{m}$, some m' is returned in $O\left(\frac{K \log T}{\Delta^2}\right)$ rounds, in $O\left(p \frac{K \log T}{\Delta^2}\right)$ regret
3. Exploitation
   - When $\hat{m}$ is returned in *Boost*, exploit $\hat{m}$; When $\hat{m}$ is not returned, ETGS

- Regret= $O\left(\frac{K \log 1/p}{\Delta^2} + \frac{N \log T}{\Delta^2} + \frac{K \log T}{\Delta} + p \frac{K \log T}{\Delta^2} + p \frac{K \log T}{\Delta^2}\right)$
- $= O\left(\frac{N \log T}{\Delta^2} + \frac{K \log \log T}{\Delta^2} + \frac{K \log T}{\Delta}\right)$ by optimizing $p = \frac{1}{\log T}$

Regret type	Regret Bound	Communication type	Strategy-proofness	References
Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Centralized, known Δ	?	[Liu et al., 2020]
Player-pessimal	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Centralized	?	[Liu et al., 2020]
Player-pessimal	$O\left(\frac{N^5K^2\log^2 T}{\rho^{N^4}\Delta^2}\right)$	Decentralized, observed matching outcomes	No	[Liu et al., 2021] [Kong et al., 2022]
Unique	$O\left(\frac{NK\log T}{\Delta^2}\right)$	Decentralized	?	[Sankararaman et al., 2021; Basu et al., 2021; Maheshwari et al., 2022]
Unique	$O\left(\frac{N\log T}{\Delta^2}\right)$	Centralized	?	[Wang and Li, 2024; Kong et al., 2024]
Optimal stable bandits (Unique)	$\Omega\left(\frac{N\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$	Decentralized	/	[Sankararaman et al., 2021]
Player-optimal	$O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{1/\varepsilon}}\right)$	Decentralized	?	[Basu et al., 2021]
Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Decentralized, observed matching outcomes	No	[Kong and Li, 2023]
Player-optimal	$O\left(\frac{K\log T}{\Delta^2}\right)$	Decentralized	No	[Zhang et al., 2022]
Player-optimal	$O\left(\frac{N^2\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$	Decentralized	Yes	[Kong et al., 2024]
Player-optimal	$\Theta\left(\frac{N\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)$	Centralized	No	[WL, 2025]

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

Open problems

- How about $O(\log\log T)$?
- How to guarantee strategy-proofness when players have more freedom?
  - The player needs to determine not only which ‘optimal arm’ to report
- How to generalize the setting and what is the optimal regret in these settings?
  - How to deal with players’ indifferent preferences?
  - How to utilize the contextual information to accelerate the learning efficiency?
  - How to handle asynchronous agents?
- Maximum matching VS. Stable matching?

Thanks!
&
Questions?

Shuai Li

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

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.
- Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.
- 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 2:

- 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.

References 3:

- Wang, Zilong and Li, Shuai. "Optimal Analysis for Bandit Learning in Matching Markets with Serial Dictatorship." Theoretical Computer Science (TCS), 2024.
- Kong, Fang, Zilong Wang and Shuai Li. "Improved Analysis for Bandit Learning in Matching Markets." 38th Conference on Neural Information Processing Systems (NeurIPS). 2024.
- Kong, Fang, et al. "Bandit Learning in Matching Markets with Indifference." The 13rd International Conference on Learning Representations (ICLR), 2025.
- Zhang, Yirui and Fang, Zhixuan. "Decentralized Competing Bandits in Many-to-One Matching Markets." International Conference on Autonomous Agents and Multiagent Systems, Extended Abstract, 2024.