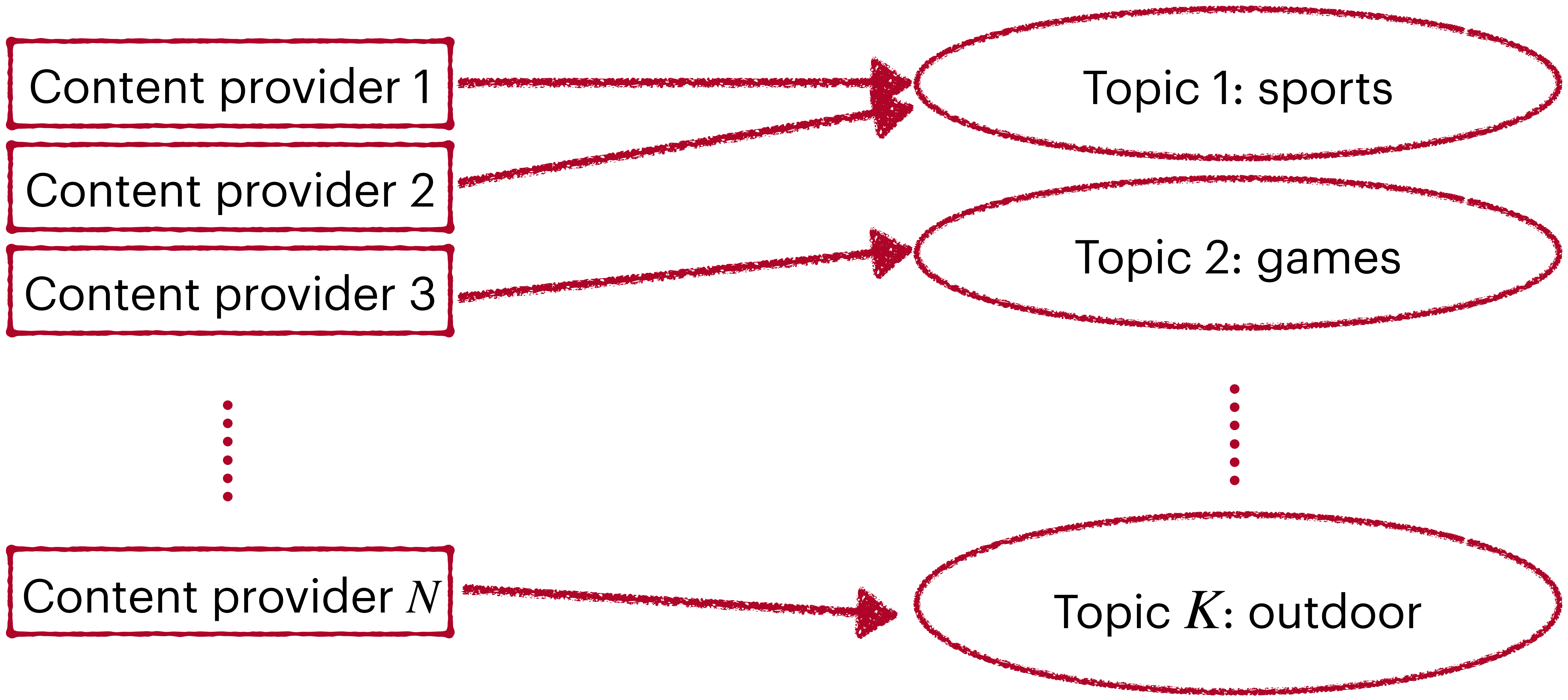
Competing for Shareable Arms in Multi-Player Multi-Armed Bandits

Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, Peng Cui

Background

- **Motivating example — competence between content providers in recommender systems (e.g., TikTok)**
  - Content providers choose to generate contents with various **topics**
  - The demand for each content topic is **constant and unknown** to content providers
  - Content providers **compete for exposure**
    - When several content providers choose the same topic, they **share the total exposure** of the topic
    - Each content provider aims to maximize his own total exposure



Problem formulation

- **A novel multi-player multi-armed bandit (MPMAB) setting with an averaging allocation model**
  - N agents (*content providers*) and K arms (*content topics*)
    - Each agent aims to maximize his own total reward after T rounds
    - Arm k has expected reward μ_k (*expected total exposure of each topic*)
  - At round $t \in [T]$
    - Agent j pulls arm $\pi_j(t)$ **selfishly** and gets reward $R_j(t)$

The reward of arm k at round t

The number of agents that pull arm k at round t

$$\mathbb{E}[\boxed{R_j(t)} \mid \boxed{X_k(t)}, \boxed{M_k(t)}] = \frac{X_k(t)}{M_k(t)}$$

Expected reward earned by an agent that pulls arm k at round t

Problem formulation

- **Target — a policy for each agent**
  - Properties when all players follow the policy
    - Convergence to the equilibrium at each round
    - Low regret for each player
  - Robust to a single agent's strategic deviation
    - Strategic deviation of a single agent can not bring significant changes to himself and other agents

Proposed method

- **Equilibrium at each round when arms' expected rewards are known**
  - The pure Nash equilibrium (PNE) exists
  - The numbers of agents that choose each arm are proportional to the arms' expected rewards
  - **More explanation of the equilibrium**
    - Content providers tend to create contents with popular topics
- **Online policy for agents**
  - Propose a novel **alternate exploration based** method to maximize the total rewards for each agent

Proposed method

- **Theoretical guarantee**
  - Properties when all players follow the policy
    - **Convergence**: Number of non-equilibrium rounds are $O(\log T)$
    - **Regret** for each player is $O(\log T)$
      - Match the lower bound under several mild assumptions
  - Strategic deviation of a single agent can not bring significant changes to himself and other agents
    - The policy is an ϵ-**Nash equilibrium** with $\epsilon = O(\log T)$
    - (β, ϵ)**-stable**
      - If an agent wants to incur a considerable loss u to another agent, then he will also suffer from a comparable loss of at least $\beta u - \epsilon$.
      - β is a constant and $\epsilon = O(\log T)$.

Conclusion

- Propose **a novel MPMAB setting with an averaging allocation model** to characterize the selfish behaviors of agents
- **Analyze the Nash equilibrium** of the problem at each round and **develop an online policy** for each agent
- Prove the following statements
  - The policy achieves a good regret guarantee when all players follow the policy
  - Any selfish player can not bring significant changes in rewards for himself and other players by deviation

Thanks for listening!

Renzhe Xu, Ph.D. candidate at Tsinghua University
Paper: https://arxiv.org/abs/2305.19158
Code: https://github.com/windxrz/smaa
Email: xrz199721@gmail.com