Information Directed Sampling for Stochastic Bandits with Graph Feedback

Fang Liu[1], Swapna Buccapatnam[2] and Ness Shroff[1]

[1]The Ohio State University

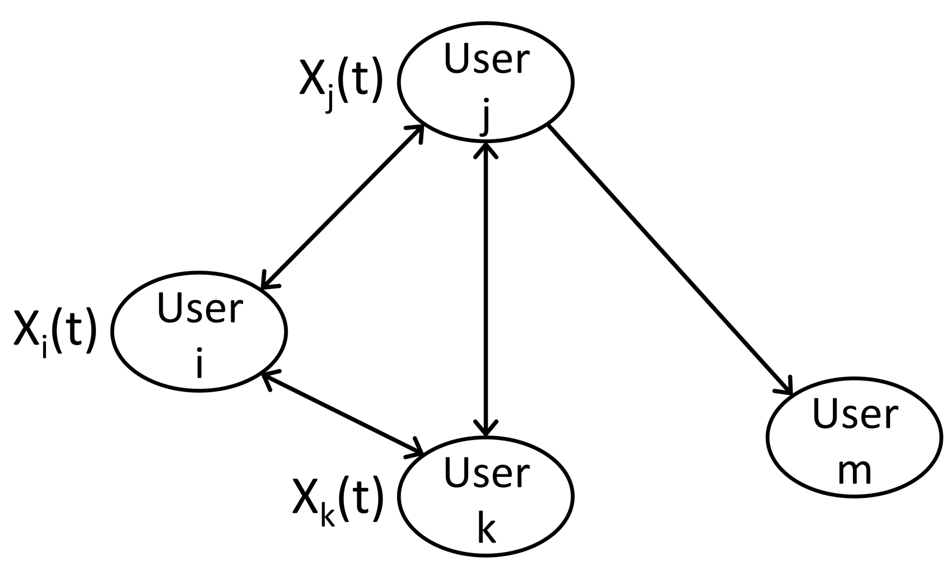
[2]AT&T Labs Research

Introduction

We study stochastic multi-armed bandits with graph feedback:
- Bandit: only obtain the reward of the chosen arm;
- Graph feedback: playing one arm may reveal other arms;
- Goal: online learning to minimize the regret due to uncertainty.

Motivation:
- Side observations are available
- Can reduce the regret
- How to use the graph information?

Applications:
- Online advertising on social media
- Recommendation systems with social connections like Yelp, Tripadvisor, …



Model

Basic setting:
- Discrete time model of horizon T, there are K arms;
- At each time t, choosing an arm A_t returns a reward Y_{t,A_t} drawn from the distribution of arm A_t;
- A^*: arm with the highest expected reward.
- Regret: expected loss compared to the oracle that plays arm A^* each time.

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,A^*} - Y_{t,A_t}\right]$$

Graph feedback $G_t=(K, E_t)$ may change over time.
- Deterministic graph: G_t is known.
- Random graph: G_t is unknown.
- Erdos−Renyi graph

Randomized policy π_t.
- Update the posterior distribution α_t of A^*
- Δ_t: instantaneous expected regret
- h_t: information gain

Algorithm

Algorithm 1 Meta-algorithm for Information Directed Sampling with Graph Feedback

Input: Time horizon T and feedback graph model $(\boldsymbol{G}_t)_{t\le T}$
for t **from** 1 **to** T **do**
  Updating statistics: compute $\boldsymbol{\alpha}_t$, $\boldsymbol{\Delta}_t$ and $\boldsymbol{h}_t$ accordingly.
  Generating policy: generate $\boldsymbol{\pi}_t$ as a function of $(\boldsymbol{\alpha}_t, \boldsymbol{\Delta}_t, \boldsymbol{h}_t, \boldsymbol{G}_t)$. (To be determined)
  Sampling: sample A_t according to $\boldsymbol{\pi}_t$, play action A_t and receive reward Y_{t,A_t}.
  Observations: observe $Y_{t,a}$ if $(A_t, a) \in \mathcal{E}_t$, where $G_t = (\mathcal{K}, \mathcal{E}_t)$ is the graph generated by $\boldsymbol{G}_t$.
end for

TS-N policy: $\boldsymbol{\pi}_t^{\text{TS-N}} = \boldsymbol{\alpha}_t$. Unaware of graph, probability matching.

IDS-N policy: $\min_{\boldsymbol{\pi}_t \in \mathcal{S}} (\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t)^2 / (\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t)$. Minimize the information ratio.

IDSN-LP policy: $\min_{\boldsymbol{\pi}_t \in \mathcal{S}} \boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t$ s.t. $\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t \ge \boldsymbol{\alpha}_t^T \boldsymbol{G}_t \boldsymbol{h}_t$. Minimize the regret.

IDS-LP policy: $\min_{\boldsymbol{\pi}_t \in \mathcal{S}} \boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t$ s.t. $\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t \ge \boldsymbol{\alpha}_t^T \boldsymbol{h}_t$. Minimize the regret.

Analysis

Theorem 1. *For any (deterministic or random) graph feedback, the Bayesian regret of IDS-LP is*

$$\mathbb{E}[R(T, \boldsymbol{\pi}^{\text{IDS-LP}})] \le \sqrt{\frac{K}{2} T H(\boldsymbol{\alpha}_1)}.$$

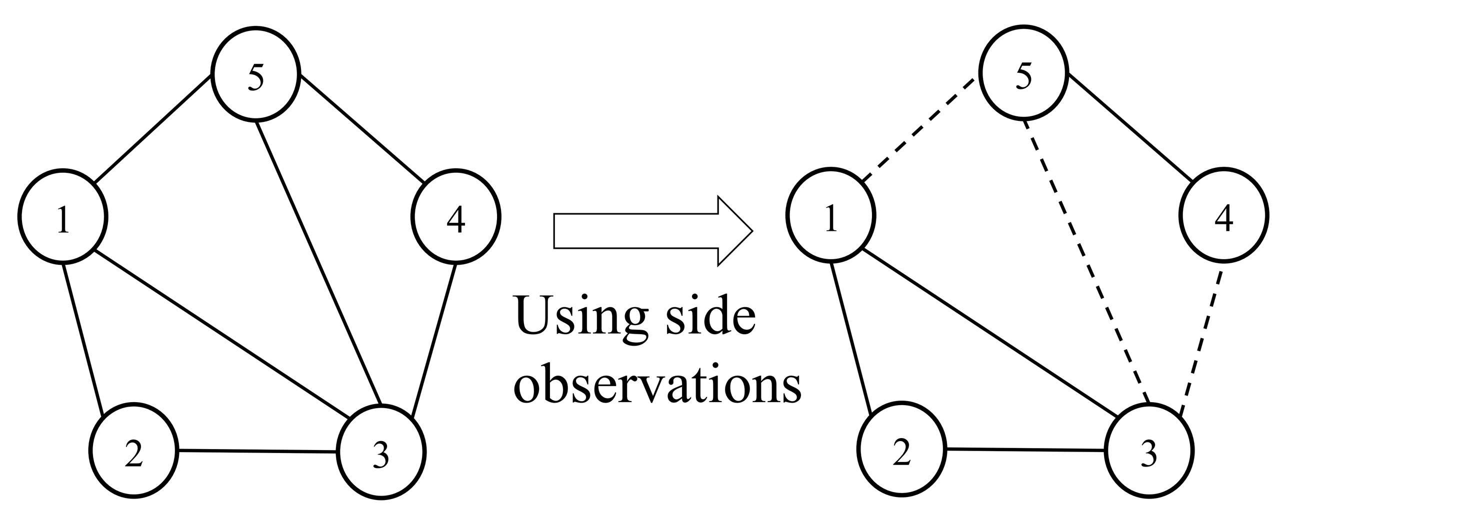
Theorem 2. *For any deterministic graph feedback* $(\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \ldots)$*, the Bayesian regrets of TS-N, IDS-N and IDSN-LP are upper-bounded by*

$$\sqrt{\sum_{t=1}^{T} \frac{\chi(\boldsymbol{G}_t)}{2} H(\boldsymbol{\alpha}_1)}.$$

Theorem 3. *For any random graph feedback* $(r_1, r_2, r_3, \ldots)$*, the Bayesian regrets of TS-N, IDS-N and IDSN-LP are upper-bounded by*

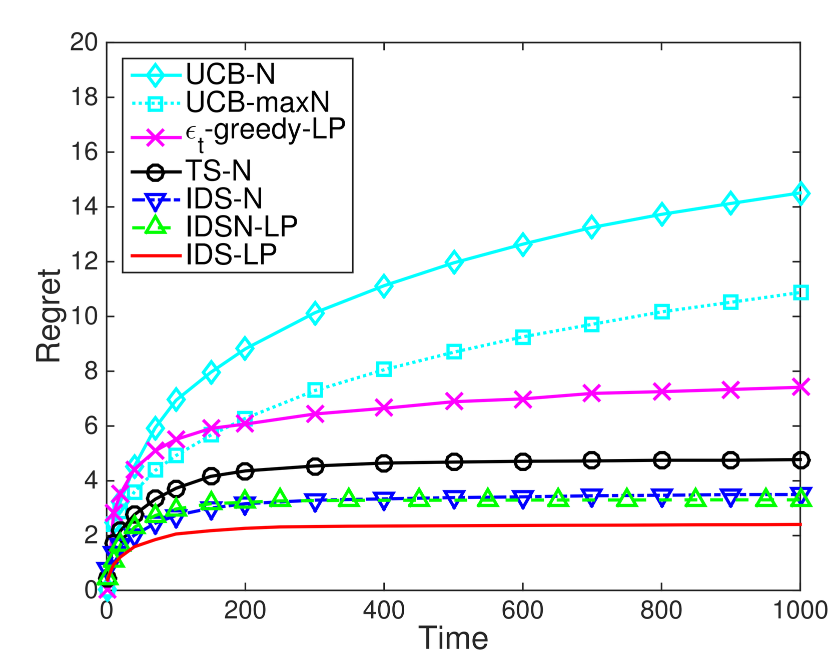
$$\sqrt{\sum_{t=1}^{T} \frac{K}{2(Kr_t + 1 - r_t)} H(\boldsymbol{\alpha}_1)}.$$

Clique cover number, $\chi(G)$, is the smallest cardinality of clique partition.

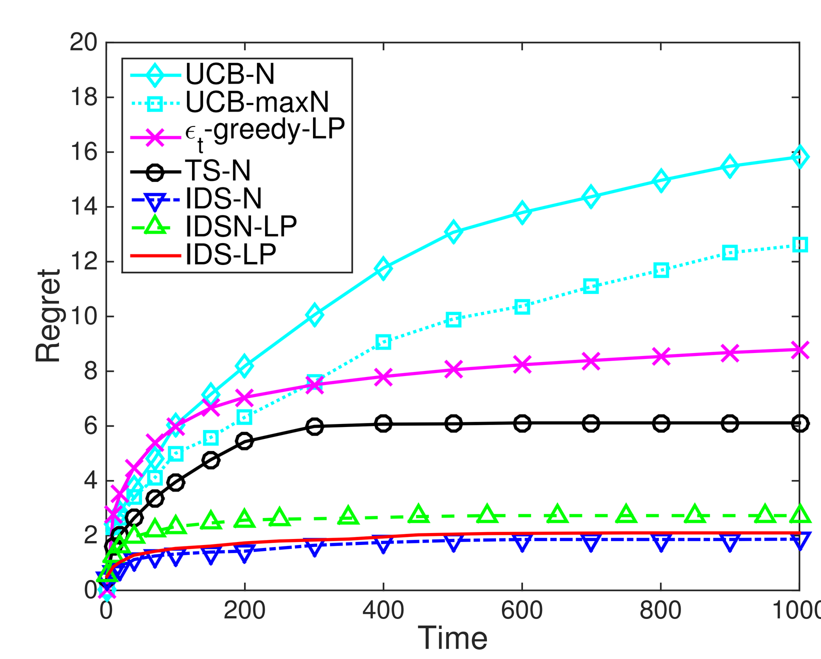


Evaluation

Deterministic graph

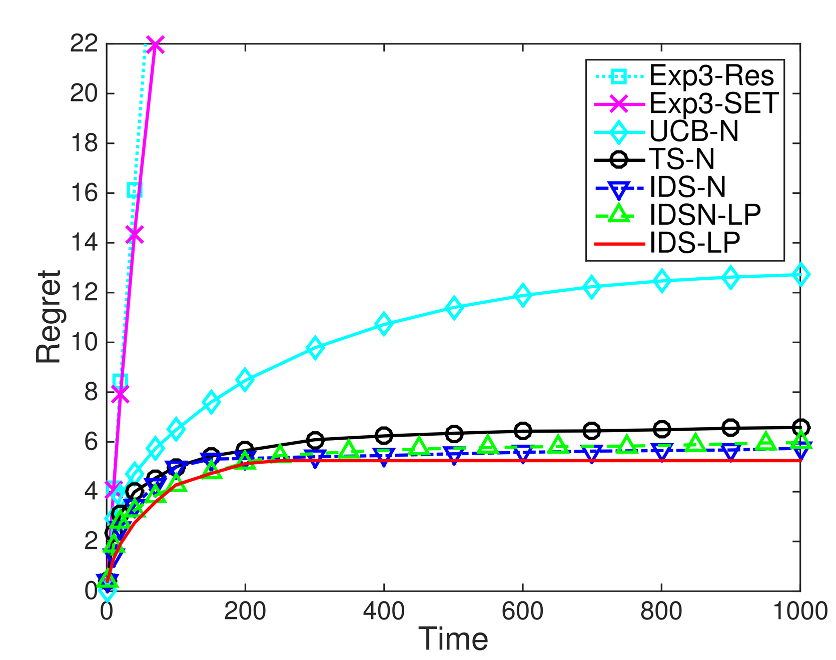


Time-invariant graphs

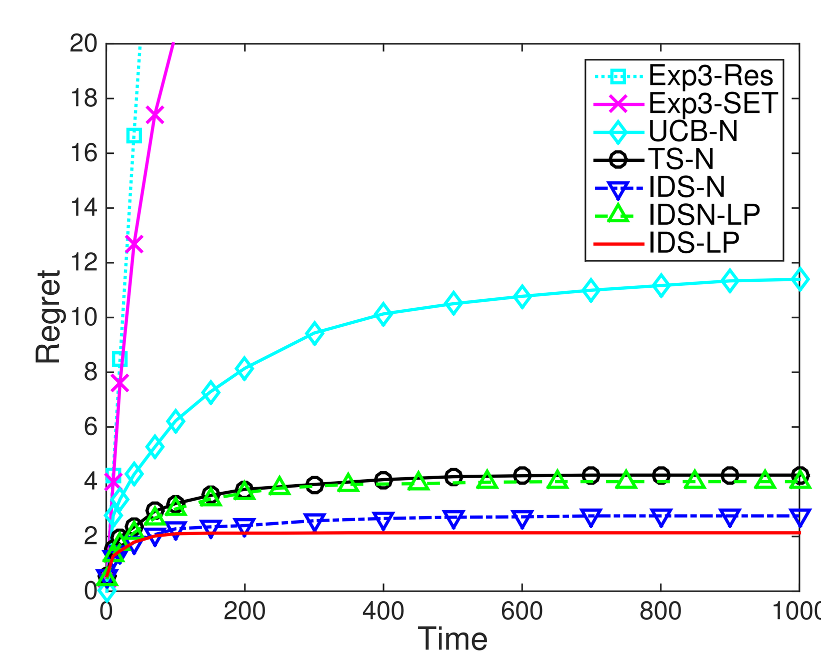


Time-variant graphs

Erdos Renyi random graph



Time-invariant r=0.25



Time-variant r_t