The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Information Directed Sampling for Stochastic Bandits with Graph Feedback

Fang Liu
The Ohio State University
Columbus, Ohio 43210
liu.3977@osu.edu

Swapna Buccapatnam
AT&T Labs Research
Middletown, NJ 07748
sb646f@att.com

Ness Shroff
The Ohio State University
Columbus, Ohio 43210
shroff.11@osu.edu

Abstract

We consider stochastic multi-armed bandit problems with graph feedback, where the decision maker is allowed to observe the neighboring actions of the chosen action. We allow the graph structure to vary with time and consider both deterministic and Erdős-Rényi random graph models. For such a graph feedback model, we first present a novel analysis of Thompson sampling that leads to tighter performance bound than existing work. Next, we propose new Information Directed Sampling based policies that are graph-aware in their decision making. Under the deterministic graph case, we establish a Bayesian regret bound for the proposed policies that scales with the clique cover number of the graph instead of the number of actions. Under the random graph case, we provide a Bayesian regret bound for the proposed policies that scales with the ratio of the number of actions over the expected number of observations per iteration. To the best of our knowledge, this is the first analytical result for stochastic bandits with random graph feedback. Finally, using numerical evaluations, we demonstrate that our proposed IDS policies outperform existing approaches, including adaptions of upper confidence bound, ϵ-greedy and Exp3 algorithms.

1 Introduction

Multi-Armed Bandits (MAB) have been used as quintessential models for sequential decision making. In the classical MAB setting, at each time, a decision maker must choose an action from a set of K actions with unknown probability distributions. Choosing an action i at time t reveals a random reward $X_i(t)$ drawn from the probability distribution of action i. The goal is to find policies that minimize the expected loss due to uncertainty about actions' distributions over a given time horizon.

In this work, we consider an important MAB setting, called the graph-structured feedback or the side-observation model (Mannor and Shamir 2011; Caron et al. 2012; Buccapatnam, Eryilmaz, and Shroff 2014; Tossou, Dimitrakakis, and Dubhashi 2017), where choosing an action i not only generates a reward from action i, but also reveals observations for a subset of the remaining actions.

Such a scenario occurs in social networks, sensors networks, and advertising. For example, a decision maker must choose one user at each time in an online social network (e.g., Facebook) to offer a promotion (Caron et al. 2012; Carpentier and Valko 2016). Each time the decision maker offers a promotion to a user, he also has an opportunity to survey the user's neighbors in the network regarding their potential interest in a similar offer.[1] Users are found to be more responsive to such surveys using social network information compared to generic surveys (Ugander et al. 2011), and this effect can be leveraged to construct side observations.

Consider another example, when the actions are advertisements (Mannor and Shamir 2011) - the decision maker constructs a graph of different vacation places (Hawaii, Caribbean, Paris, etc.), where links capture similarities between different places. When a customer shows interest in one of the places, he is also asked to provide his opinion about the neighboring places in the graph.

Side-observation models are also applicable to sensor networks that monitor events, where an agent must choose one (or a few) sensor to sample at each time. The reward obtained from choosing a sensor is related to its accuracy of monitoring the event. Neighboring sensors can communicate their observations to each other and this data aggregation has the desired affect of obtaining side-observations from neighboring sensors on determining which sensor(s) to select.

In the existing literature (Caron et al. 2012; Buccapatnam, Eryilmaz, and Shroff 2014; Buccapatnam et al. 2017; Tossou, Dimitrakakis, and Dubhashi 2017), researchers propose new policies for the side observation model in the stochastic bandit setting that exploit the graph structure to accelerate learning. Caron et al. (2012) and Buccapatnam, Eryilmaz, and Shroff (2014) propose extensions to upper confidence bound based policies, originally proposed for the classical MAB setting in Auer, Cesa-Bianchi, and Fischer (2002). Policies proposed by Buccapatnam, Eryilmaz, and Shroff (2014), namely ϵ_t-greedy-LP and UCB-LP, are shown to be asymptotically optimal, both in terms of the network structure and time. Tossou, Dimitrakakis, and Dub-

[1] This is possible when the online network has an additional survey feature that generates "side observations". For example, when user i is offered a promotion, her neighbors may be queried as follows: "User i was recently offered a promotion. Would you also be interested in the offer?".

hashi (2017) analyze the Bayesian regret performance of another well known classical bandit policy called Thompson Sampling (TS) (Thompson 1933) for the side-observation model. We make the following important contributions to this existing literature on graphical bandits:

- For the graph structured MAB feedback model we allow the side observation graph to vary with time. We focus on developing a problem-independent Bayesian regret bound. We provide a tighter bound for Thompson sampling, given in terms of the clique cover number of the side-observation graph than the bound presented by Tossou, Dimitrakakis, and Dubhashi (2017).
- We also propose three algorithms, all of which are based on the approach of Information Directed Sampling (IDS) developed by Russo and Van Roy (2014). We show that these algorithms enjoy the same theoretical bounds as Thompson Sampling in terms of the clique cover number of the graph. However, using numerical evaluations, we show in Section 7, that IDS based policies outperform existing policies in Buccapatnam, Eryilmaz, and Shroff (2014), that are provably asymptotically optimal both in terms of network structure and time. Hence, this raises the open question of how to determine better Bayesian regret bounds for our IDS based policies in terms of the network structure.
- In contrast with existing works, we also consider the novel setting of a time variant random graph feedback model, where side observations from neighboring actions are obtained with a probability r_t in time t.[2] We provide Bayesian regret bounds for Thompson Sampling and our proposed IDS based policies for this probabilistic model as well. We believe that our work provides the first result for stochastic bandits with random graph feedback.

2 Related Work

Our work is related to Russo and Van Roy (2014), the authors of which propose a novel approach called IDS. IDS samples each action in a manner that minimizes the ratio between the squared expected single-period regret and the mutual information between optimal action and the next observation. It has been shown in Russo and Van Roy (2014) using numerical simulations that IDS outperforms TS and Upper Confidence Bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002) in the classical bandit setting. This motivated us to investigate extensions of IDS policy for the stochastic bandits with graph feedback and compare with adaptions of TS and UCB policies, which have been studied in the literature (Caron et al. 2012; Buccapatnam, Eryilmaz, and Shroff 2014; Tossou, Dimitrakakis, and Dubhashi 2017).

The Thompson Sampling algorithm as analyzed by Tossou, Dimitrakakis, and Dubhashi (2017) does not explicitly use the graph structure in each step for its operation (similar to UCB-N algorithm proposed by Caron et al. (2012)). This is attractive when such graphical information is difficult to obtain, a case also studied in Cohen, Hazan, and Koren (2016). However, in many cases such as the problem of promotions in online social networks and sensor networks, the graph structure is revealed or can be learned a priori. When such knowledge is available, Buccapatnam, Eryilmaz, and Shroff (2014) show that using graphical information to make choices in each time helps in obtaining the optimal regret both in terms of network structure and time asymptotically. This work motivated us to investigate extensions of IDS policies that exploit the knowledge of network structure in our work. Using numerical evaluations, we find that IDS policies outperform asymptotically optimal policies presented in Buccapatnam, Eryilmaz, and Shroff (2014).

Non-stochastic bandits with graph feedback have been studied by a line of work (Mannor and Shamir 2011; Alon et al. 2013; Kocák et al. 2014; Alon et al. 2014; 2015). Other related partial feedback models include label efficient bandit in Audibert and Bubeck (2010) and prediction with limited advice in Seldin et al. (2014), where side observations are limited by a budget. A summary of the bandits on graphs can be found in Valko (2016).

3 Problem Formulation

3.1 Stochastic Bandit Model

We consider a Bayesian formulation of the stochastic K-armed bandit problem in which uncertainties are modeled as random variables. At each time $t \in \mathbb{N}$, a decision maker chooses an action A_t from a finite action set $\mathcal{K} = \{1, \ldots, K\}$ and receives the corresponding random reward Y_{t,A_t}. Without loss of generality, we assume the space of possible rewards $\mathcal{Y} = [0, 1]$. Note that the results in this work can be extended to the case where reward distributions are sub-Gaussian. There is a random variable $Y_{t,a} \in \mathcal{Y}$ associated with each action $a \in \mathcal{K}$ and $t \in \mathbb{N}$. We assume that $\{Y_{t,a}, \forall a \in \mathcal{K}\}$ are independent for each time t. Let $\boldsymbol{Y}_t \triangleq (Y_{t,a})_{a\in\mathcal{K}}$ be the vector of random variables at time $t \in \mathbb{N}$. The true reward distribution p^* is a distribution over $\mathcal{Y}^K$, which is randomly drawn from the family of distributions $\mathcal{P}$ and unknown to the decision maker. Conditioned on p^*, $(\boldsymbol{Y}_t)_{t\in\mathbb{N}}$ is an independent and identically distributed sequence with each element $\boldsymbol{Y}_t$ sampled from the distribution p^*.

Let $A^* \in \arg\max_{a\in\mathcal{K}} \mathbb{E}[Y_{t,a}|p^*]$ be the true optimal action conditioned on p^*. Then the T period regret of the decision maker is the expected difference between the total rewards obtained by an oracle that always chooses the optimal action and the accumulated rewards up to time horizon T. Formally, we study the expected regret

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

where the expectation is taken over the randomness in the action sequence $(A_1, \ldots, A_T)$ and the outcomes $(\boldsymbol{Y}_t)_{t\in\mathbb{N}}$ and over the prior distribution over p^*. This notion of regret is also known as *Bayesian regret* or *Bayes risk*.

[2] This models the scenario of sensor networks where errors due to channel conditions can cause side observations to be randomly erased.

3.2 Graph Feedback Model

In this problem, we assume the existence of side observations, which is described by a graph $G_t = (\mathcal{K}, \mathcal{E}_t)$ over the action set for each time t. The graph G_t may be directed or undirected and can be dependent on time t. At each time t, the decision maker observes the reward Y_{t,A_t} for playing action A_t as well as the outcome $Y_{t,a}$ for each action $a \in \{a \in \mathcal{K} | (A_t, a) \in \mathcal{E}_t\}$. Note that it becomes the classical bandit feedback setting when the graph is empty (i.e., no edge exists) and it becomes the full-information (expert) setting when the graph is complete for all time t. In this work, we study two types of graph feedback models: *deterministic graph* and *Erdős-Rényi random graph*.

Deterministic graph. In the deterministic graph feedback model, we assume that the graph G_t is fixed before the decision is made at each time t. Let $\boldsymbol{G}_t \in \mathbb{R}^{K \times K}$ be the adjacent matrix that represents the deterministic graph feedback structure G_t. Let $\boldsymbol{G}_t(i,j)$ be the element at the i-th row and j-th column of the matrix. Then $\boldsymbol{G}_t(i,j) = 1$ if there exists an edge $(i,j) \in \mathcal{E}_t$ and $\boldsymbol{G}_t(i,j) = 0$ otherwise. Note that we assume $\boldsymbol{G}_t(i,i) = 1$ for any $i \in \mathcal{K}$.

Definition 1. (Clique cover number) A *Clique* of a graph $G = (\mathcal{K}, \mathcal{E})$ is a subset $S \subseteq \mathcal{K}$ such that the sub-graph formed by S and $\mathcal{E}$ is a complete graph. A *Clique cover* of a graph $G = (\mathcal{K}, \mathcal{E})$ is a partition of $\mathcal{K}$, denoted by $\mathcal{C}$, such that S is a clique for each $S \in \mathcal{C}$. The cardinality of the smallest clique cover is called the *clique cover number*, which is denoted by $\chi(G)$.

In this work, we slightly abuse the notation of clique cover number and use $\chi(G_t)$ and $\chi(\boldsymbol{G}_t)$ interchangeably since $\boldsymbol{G}_t$ fully characterizes the graph structure G_t.

Erdős-Rényi Random Graph. In the Erdős-Rényi random graph feedback model, we assume that the graph G_t is generated from an Erdős-Rényi model with time-dependent parameter r_t after the decision is made at each time t. In other words, the decision maker can reveal the outcome $Y_{t,a}$ with probability r_t for each action $a \neq A_t$ at time t. This feedback model is also known as *probabilistically triggered arms* (Chen et al. 2016).

We generalize the adjacent matrix representation of a deterministic graph feedback model to a random graph feedback model, such that each (i,j)-th element of the matrix is the probability of observing action j via playing action i. For each time t, the adjacent matrix is denoted by $\boldsymbol{G}_t$ to unify the representation of our algorithms and analysis. Then, we have that $\boldsymbol{G}_t(i,i) = 1$ for any $i \in \mathcal{K}$ and $\boldsymbol{G}_t(i,j) = r_t$ for any $i \neq j$. Note that parameter r_t fully characterizes the random graph feedback model.

3.3 Randomized Policies

We define all random variables with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$ such that $\mathcal{F}_t \subseteq \mathcal{F}$ is the σ-algebra generated by the observation history O_{t-1}. The observation history O_t includes all decisions, rewards and side observations from time 1 to time t. For each time t, the decision maker chooses an action based on the history O_{t-1} and possibly some randomness. Any policy of the decision maker can be viewed as a *randomized policy* $\boldsymbol{\pi}$, which is an $\mathcal{F}_t$-adapted sequence $(\boldsymbol{\pi}_t)_{t \in \mathbb{N}}$. For each time t, the decision maker chooses an action randomly according to $\boldsymbol{\pi}_t(\cdot) = \mathbb{P}(A_t = \cdot | \mathcal{F}_t)$, which is a probability distribution over $\mathcal{K}$. Let $\mathbb{E}[R(T, \boldsymbol{\pi})]$ be the Bayesian regret defined by (1) when the decisions $(A_1, \ldots, A_T)$ are chosen according to $\boldsymbol{\pi}$.

Uncertainty about p^* induces uncertainty about the true optimal action A^*, which is described by a prior distribution $\boldsymbol{\alpha}_1$ of A^*. Let $\boldsymbol{\alpha}_t$ be the posterior distribution of A^* given the history O_{t-1}, i.e., $\boldsymbol{\alpha}_t(\cdot) = \mathbb{P}(A^* = \cdot | \mathcal{F}_t)$. Then, $\boldsymbol{\alpha}_{t+1}$ can be updated by Bayes rule given $\boldsymbol{\alpha}_t$, decision A_t, reward Y_{t,A_t} and side observations. The *Shannon entropy* of $\boldsymbol{\alpha}_t$ is defined as $H(\boldsymbol{\alpha}_t) \triangleq -\sum_{i \in \mathcal{K}} \boldsymbol{\alpha}_t(i) \log(\boldsymbol{\alpha}_t(i))$. We slightly abuse the notation of $\boldsymbol{\pi}_t$ and $\boldsymbol{\alpha}_t$ such that they represent distributions (or functions) over finite set $\mathcal{K}$ as well as vectors in a simplex $\mathcal{S} \subset \mathbb{R}^K$. Note that $\mathcal{S} = \{\boldsymbol{\pi} \in \mathbb{R}^K | \sum_{i=1}^K \boldsymbol{\pi}(i) = 1, \boldsymbol{\pi}(i) \geq 0, \forall i \in \mathcal{K}\}$.

Let $\boldsymbol{\Delta}_t$ be the instantaneous regret vector such that the i-th coordinate, $\boldsymbol{\Delta}_t(i) \triangleq \mathbb{E}[Y_{t,A^*} - Y_{t,i} | \mathcal{F}_t]$, is the expected regret of playing action i at time t. Let $\boldsymbol{g}_t$ be the information gain vector such that the i-th coordinate, $\boldsymbol{g}_t(i) = \mathbb{E}[H(\boldsymbol{\alpha}_t) - H(\boldsymbol{\alpha}_{t+1}) | \mathcal{F}_t, A_t = i]$, is the expected information gain of playing action i at time t. Note that the information gain of playing action i consists of that of observing the reward $Y_{t,i}$ and possibly some side observations. We define the information gain of observing action a (i.e., $Y_{t,a}$) as $\boldsymbol{h}_t(a) \triangleq I_t(A^*; Y_{t,a})$, which is the *mutual information* under the posterior distribution between random variables A^* and $Y_{t,a}$. Let $D(\cdot||\cdot)$ be the *Kullback-Leibler* divergence between two distributions[3]. By the definition of mutual information, we have that $I_t(A^*; Y_{t,a}) \triangleq$

$$D(\mathbb{P}((A^*, Y_{t,a}) \in \cdot | \mathcal{F}_t) || \mathbb{P}(A^* \in \cdot | \mathcal{F}_t) \mathbb{P}(Y_{t,a} \in \cdot | \mathcal{F}_t)). \tag{2}$$

The following proposition reveals the relationship between vector $\boldsymbol{g}_t$ and $\boldsymbol{h}_t$.

Proposition 1. *Under the (deterministic or random) graph feedback $\boldsymbol{G}_t$, we have $\boldsymbol{g}_t \geq \boldsymbol{G}_t \boldsymbol{h}_t$.*

Intuitively, Proposition 1 shows that the information gain of observing the reward and some side observations is at least the sum of the information gain of each individual observation. A formal proof is provided in the technical report (Liu, Buccapatnam, and Shroff 2017).

At each time t, a randomized policy updates $\boldsymbol{\alpha}_t$, $\boldsymbol{\Delta}_t$ and $\boldsymbol{h}_t$ and makes a decision according to a sampling distribution $\boldsymbol{\pi}_t$.

4 Algorithms

For any randomized policy, we define the *information ratio* of sampling distribution $\boldsymbol{\pi}_t$ at time t as

$$\Psi_t(\boldsymbol{\pi}_t) \triangleq (\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t)^2 / (\boldsymbol{\pi}_t^T \boldsymbol{g}_t). \tag{3}$$

[3] If P is absolutely continuous with respect to Q, then $D(P||Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P w.r.t. Q.

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

Require: Time horizon T and feedback graph model $(\boldsymbol{G}_t)_{t\leq T}$
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

Note that $\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t$ is the expected instantaneous regret of the sampling distribution $\boldsymbol{\pi}_t$, and $\boldsymbol{\pi}_t^T \boldsymbol{g}_t$ is the expected information gain of the sampling distribution $\boldsymbol{\pi}_t$. So the information ratio $\Psi_t(\boldsymbol{\pi}_t)$ measures the "energy" cost (which is the square of the expected instantaneous regret) per bit of information acquired.

The key idea of the IDS based policy is keeping the information ratio bounded in order to balance between having low expected instantaneous regret (a.k.a. exploitation) and obtaining knowledge about the optimal action (a.k.a. exploration). In other words, if the information ratio is bounded, then the expected regret is bounded in terms of the maximum amount of information one could expect to acquire, which is at most the entropy of the prior distribution of A^*, i.e., $H(\boldsymbol{\alpha}_1)$. As we show in Section 5, we can find upper bounds for the information ratios of the policies we provide here.

In practice, the information gain vector $\boldsymbol{g}_t$ is quite complicated to compute even assuming a Bernoulli distribution model for each action. However, computing the information gain of observing each individual action, i.e., $\boldsymbol{h}_t$, is much easier since it is only the mutual information of two random variables. By Proposition 1, we have that $\Psi_t(\boldsymbol{\pi}_t) \leq (\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t)^2/(\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t)$. So we can design our IDS based policies according to $\boldsymbol{h}_t$ and $\boldsymbol{G}_t$ instead of $\boldsymbol{g}_t$. We provide a meta-algorithm for IDS based policies in Algorithm 1. What remains is to design $\boldsymbol{\pi}_t$ as a function of $\boldsymbol{\alpha}_t$, $\boldsymbol{\Delta}_t$, $\boldsymbol{h}_t$ and $\boldsymbol{G}_t$. Note that one can replace $\boldsymbol{G}_t \boldsymbol{h}_t$ by $\boldsymbol{g}_t$ in the IDS based algorithms and the regret results in Section 5 still hold.

TS-N policy is a natural adaption of Thompson Sampling under the graph feedback. It replaces the generating policy step in Algorithm 1 by

$$\boldsymbol{\pi}_t^{\text{TS-N}} = \boldsymbol{\alpha}_t. \tag{4}$$

The TS-N policy ignores the graph structure information $\boldsymbol{G}_t$, and sample the action according to the posterior distribution of A^*.

IDS-N policy replaces the generating policy step in Algorithm 1 by $\boldsymbol{\pi}_t^{\text{IDS-N}}$, which is the solution of the following optimization problem P_1.

$$P_1 : \min_{\boldsymbol{\pi}_t \in \mathcal{S}} (\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t)^2/(\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t). \tag{5}$$

The IDS-N policy greedily minimizes the information ratio (upper bound) at each time.

IDSN-LP policy replaces the generating policy step in Algorithm 1 by $\boldsymbol{\pi}_t^{\text{IDSN-LP}}$, which is the solution of the following linear programming problem P_2.

$$P_2 : \min_{\boldsymbol{\pi}_t \in \mathcal{S}} \boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t \quad \text{s.t. } \boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t \geq \boldsymbol{\alpha}_t^T \boldsymbol{G}_t \boldsymbol{h}_t. \tag{6}$$

The IDSN-LP policy greedily minimizes the expected instantaneous regret at each time with the constraint that the information gain is at least the one obtained by TS-N policy.

IDS-LP policy replaces the generating policy step in Algorithm 1 by $\boldsymbol{\pi}_t^{\text{IDS-LP}}$, which is the solution of the following linear programming problem P_3.

$$P_3 : \min_{\boldsymbol{\pi}_t \in \mathcal{S}} \boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t \quad \text{s.t. } \boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t \geq \boldsymbol{\alpha}_t^T \boldsymbol{h}_t. \tag{7}$$

The IDS-LP policy greedily minimizes the expected instantaneous regret at each time with the constraint that the information gain is at least the one obtained by TS policy without graph feedback. IDS-LP policy reduces the extent of exploration compared to IDSN-LP policy. Intuitively, it greedily exploits the current knowledge of the optimal action with controlled exploration. Though we can not find better regret bound for IDS-LP than IDSN-LP, IDS-N and TS-N, IDS-LP outperforms the others in numerical results as shown in Section 7.

5 Regret Analysis

In this section, we first present a known general bound for any randomized policy and provide the regret upper bound results of the proposed policies for the deterministic and random graph feedback. The regret analysis relies on the following bound, which is shown in Russo and Roy (2014).

Lemma 1. (General Bound from Russo and Roy (2014)) *For any policy $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3, \ldots)$ and time horizon $T \in \mathbb{N}$,*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] \leq \sqrt{\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\pi}}[\Psi_t(\boldsymbol{\pi}_t)] H(\boldsymbol{\alpha}_1)}. \tag{8}$$

Lemma 1 shows that we only need to bound expected information ratio $\mathbb{E}_{\boldsymbol{\pi}}[\Psi_t(\boldsymbol{\pi}_t)]$ to obtain an upper bound for a randomized policy. The next result follows from the fact that the information ratio of IDS-LP policy can be bounded by $K/2$.

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

$$\mathbb{E}[R(T, \boldsymbol{\pi}^{\textit{IDS-LP}})] \leq \sqrt{\frac{K}{2} T H(\boldsymbol{\alpha}_1)}. \tag{9}$$

The key idea of the proof is comparing the information ratio of IDS-LP to that of TS with bandit feedback. The detailed proof of Theorem 1 can be found in the technical report (Liu, Buccapatnam, and Shroff 2017). The next proposition shows a general bound for information ratios of TS-N, IDS-N and IDSN-LP policies.

Proposition 2. *For any (deterministic or random) graph feedback $\boldsymbol{G}_t$, we have that $\Psi_t(\boldsymbol{\pi}_t^{\text{TS-N}})$, $\Psi_t(\boldsymbol{\pi}_t^{\text{IDS-N}})$ and $\Psi_t(\boldsymbol{\pi}_t^{\text{IDSN-LP}})$ are upper-bounded by $\psi_t \triangleq \frac{\left(\boldsymbol{\Delta}_t^T \boldsymbol{\alpha}_t\right)^2}{(\boldsymbol{G}_t \boldsymbol{h}_t)^T \boldsymbol{\alpha}_t}$.*

The proof of Proposition 2 can be found in the technical report (Liu, Buccapatnam, and Shroff 2017). Combining this result with Lemma 1, we can obtain unified regret result for TS-N, IDS-N and IDSN-LP by bounding the ratio ψ_t. Now, we are ready to present the regret results separately for the deterministic and the random graph feedback.

5.1 Deterministic Graph

The following result shows the unified regret upper bound of TS-N, IDS-N and IDSN-LP under the deterministic graph feedback. The detailed proof is presented in the technical report (Liu, Buccapatnam, and Shroff 2017).

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*

$$\sqrt{\sum_{t=1}^{T} \frac{\chi(\boldsymbol{G}_t)}{2} H(\boldsymbol{\alpha}_1)}. \tag{10}$$

Recently, a similar result for TS-N has been shown to be $\sqrt{\max_t \frac{\chi(\boldsymbol{G}_t)}{2} T H(\boldsymbol{\alpha}_1)}$ in Tossou, Dimitrakakis, and Dubhashi (2017). Apparently, Theorem 2 provides a tighter bound. We have the following result when the graph is also time-invariant.

Corollary 1. *For any time-invariant and deterministic graph feedback $\boldsymbol{G}$ (i.e., $\boldsymbol{G}_t = \boldsymbol{G}\ \forall t \in \mathbb{N}$), the Bayesian regrets of TS-N, IDS-N and IDSN-LP are upper-bounded by*

$$\sqrt{\frac{\chi(\boldsymbol{G})}{2} T H(\boldsymbol{\alpha}_1)}. \tag{11}$$

Corollary 1 shows that TS-N, IDS-N and IDSN-LP can benefit from the side observations. In other words, the above problem-independent regret upper bound scales with the clique cover number of the graph instead of the number of actions (it is known that the regret bound of TS without side observations is (9)). A similar result has been disclosed by Caron et al. (2012) in the form of problem-dependent upper bound. They show that UCB-N scales with the clique cover number compared to UCB without side observations.

An information theoretic lower bound on the problem-independent regret has been shown in Mannor and Shamir (2011) to scale with the independence number[4]. In general, the independence number is less than or equal to the clique cover number. However, the equality holds for a large class of graphs, such as star graphs and perfect graphs. In other words, our policies are order-optimal for a large class of graphs.

[4]Independence number is the largest number of nodes without edges between them.

5.2 Erdős-Rényi Random Graph

The following result shows the unified regret upper bound of TS-N, IDS-N and IDSN-LP under the Erdős-Rényi random graph feedback. The detailed proof is presented in the technical report (Liu, Buccapatnam, and Shroff 2017).

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*

$$\sqrt{\sum_{t=1}^{T} \frac{K}{2(Kr_t + 1 - r_t)} H(\boldsymbol{\alpha}_1)}. \tag{12}$$

As far as we know, this is the first result for stochastic bandit with random graph feedback. An analogous result has been shown for the non-stochastic bandit, for which Kocák, Neu, and Valko (2016) proposed Exp3-Res[5] policy with guarantee of $O\left(\sqrt{\sum_{t=1}^{T} \frac{1}{r_t} \log K}\right)$ if $r_t \geq \frac{\log T}{2K-2}$ holds for all t. Theorem 3 recovers the same guarantee without restriction on r_t since $H(\boldsymbol{\alpha}_1) \leq \log K$. We have the following result when r_t is time-invariant.

Corollary 2. *For any time-invariant and random graph feedback r (i.e., $r_t = r\ \forall t \in \mathbb{N}$), the Bayesian regrets of TS-N, IDS-N and IDSN-LP are upper-bounded by*

$$\sqrt{\frac{K}{2(Kr + 1 - r)} T H(\boldsymbol{\alpha}_1)}. \tag{13}$$

Corollary 2 shows that the benefit from side observations can be measured by the expected number of observations per time step, i.e., $(K-1)r+1$. In other words, the above regret upper bound scales with the ratio of the number of actions and the expected number of observations. When $r = 1$, this ratio equals to 1, which yields the regret result for stochastic bandit with full information (Russo and Van Roy 2016). When $r = 0$, this ratio equals to K, which yields the regret result for stochastic bandit with bandit feedback (Russo and Van Roy 2016). An analogous result can be found as Corollary 3 in Alon et al. (2013) for the non-stochastic bandits.

6 Computation

In this section, we provide computational methods for updating statistics and discuss the complexity issues of the algorithms.

6.1 Computational Methods for Updating Statistics

Algorithm 1 offers an abstract design principle with the availability of the statistics (i.e., $\boldsymbol{\alpha}_t$, $\boldsymbol{h}_t$ and $\boldsymbol{\Delta}_t$). However, additional work is required to design efficient computational methods to update these statistics for specific problems. In general, the challenge of updating statistics is to compute and represent a posterior distribution given observations, which is also faced with Thompson Sampling.

[5]Note that they assume that r_t is not available to Exp3-Res. However, it is still reasonable to compare since TS-N is not aware of r_t as well.

When the posterior distribution is complex, one can often generate samples from this distribution using Markov Chain Monte Carlo (MCMC) algorithms, enabling efficient implementation of IDS. A detailed discussion of applying MCMC methods for implementing randomized policy can be found in Scott (2010). However, when the posterior distribution has a closed form or the conjugate prior is well studied, the posterior distributions can be efficiently computed and stored, as are the cases of Beta-Bernoulli bandits and Gaussian bandits (Wu, György, and Szepesvári 2015).

In the numerical experiment, we implement Algorithm 2 in Russo and Van Roy (2014) to represent the posterior distribution and compute the statistics[6] for Beta-Bernoulli bandits. The key idea is that the Beta distribution is a conjugate prior for the Bernoulli distribution. Specifically, given the prior that the expectation θ_i is drawn from Beta(β_i^1, β_i^2), the posterior distribution of observing $Y_i \sim$ Bernoulli(θ_i) is Beta$(\beta_i^1 + Y_i, \beta_i^2 + 1 - Y_i)$. So the posterior distribution can be updated and represented easily. Then what remains is to calculate the statistics $\boldsymbol{\alpha}_t$, $\boldsymbol{h}_t$ and $\boldsymbol{\Delta}_t$ given the posterior distributions. More details of the calculations can be found in Russo and Van Roy (2014). As stated in Russo and Van Roy (2014), practical implementation of updating statistics involves integrals, which can be evaluated at a discrete grid of points within interval $[0, 1]$. The computational cost of updating statistics is $O(K^2 n)$ where n is the number of points used in the discretization of $[0, 1]$.

6.2 Complexity of Optimization Problems involved in IDS based Policies

The following result shows that problem P_1 is a convex optimization problem and has a structure in the optimal solution.

Proposition 3. *The function* $\Psi_t : \boldsymbol{\pi}_t \rightarrow (\boldsymbol{\pi}_t^T \boldsymbol{\Delta}_t)^2 / (\boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t)$ *is convex on* $\{\boldsymbol{\pi}_t \in \mathcal{S} | \boldsymbol{\pi}_t^T \boldsymbol{G}_t \boldsymbol{h}_t > 0\}$. *Moreover, there is an optimal solution* $\boldsymbol{\pi}_t^*$ *to problem* P_1 *such that* $|\{i : \boldsymbol{\pi}_t^*(i) > 0\}| \leq 2$.

The proof is an adaption of the proof of Proposition 1 in Russo and Van Roy (2014) by replacing $\boldsymbol{g}_t$ by $\boldsymbol{G}_t \boldsymbol{h}_t$. Proposition 3 shows that problem P_1 is a convex optimization problem, which can be solved by a standard convex optimization solver. What's more, there exists an optimal solution with support size of at most 2. One can search all the pairs of actions and find the optimal solution by brute force. For each pair, it remains to solve a convex optimization problem with one parameter by closed form. So the computational complexity is $O(K^2)$.

Problems P_2 and P_3 are linear programming problems, which can be solved efficiently in polynomial time by standard methods. Moreover, the following result shows that they can be solved much faster. The proof is presented in the technical report (Liu, Buccapatnam, and Shroff 2017).

Proposition 4. *The optimization problems* P_2 *and* P_3 *can be solved in* $O(K)$ *iterations.*

In sum, the computational complexity of the proposed IDS based policies (including TS-N) is $O(K^2 n)$ per iteration, where n is the number of points used in the discretization of $[0, 1]$. Note that the complexity of UCB based policies is $O(K)$. IDS based policies can improve the regret performance with reasonable computation cost.

7 Numerical Results

This section presents numerical results from experiments that evaluate the effectiveness of IDS based policies in comparison to alternative algorithms. We consider the classical Beta-Bernoulli bandit problem with independent actions. The reward of each action i is a Bernoulli(θ_i) random variable and θ_i is independently drawn from Beta$(1, 1)$. In the experiment, we set $K = 5$ and $T = 1000$. All the regret results are averaged over 1000 trials.

Figure 1 presents the cumulative regret results under the deterministic graph feedback. For the time-invariant case, we use a graph with 2 cliques, presented in the technical report (Liu, Buccapatnam, and Shroff 2017). For the time-variant case, the sequence of graphs is generated by the Erdős-Rényi model[7]. We compare our policies to three other algorithms that are proposed for the stochastic bandit with deterministic graph feedback. Caron et al. (2012) proposed UCB-N and UCB-maxN that closely follow the UCB policy and use side observations for better reward estimates (UCB-N) or choose one of the neighboring nodes with a better empirical estimate (UCB-maxN). It has been shown that the regret of UCB-N and UCB-maxN scale with the clique cover number in the time-invariant case. Buccapatnam, Eryilmaz, and Shroff (2014) improved the results in Caron et al. (2012) with LP-based algorithms (ϵ_t-greedy-LP and UCB-LP[8]) and guarantees scaling with the domination number[9] in the time-invariant case. We find that TS-N policy outperforms these three algorithms, which is consistent with the empirical observation in the bandit feedback setting (Chapelle and Li 2011). In addition, IDS-N, IDSN-LP and IDS-LP outperform TS-N policy in both cases. These improvements stem from the exploitation of graph structure in IDS based policies, which raises an open question of determining better regret bounds for our IDS based policies in terms of graph structure.

Figure 2 presents the cumulative regret results under the Erdős-Rényi random graph feedback. For the time-invariant case, we fix the parameter $r = 0.25$. For the time-variant case, the parameter r_t is independently drawn from the uniform distribution over the interval $[0, 1]$. We compare our policies to UCB-N[10] and two other algorithms (Exp3-SET

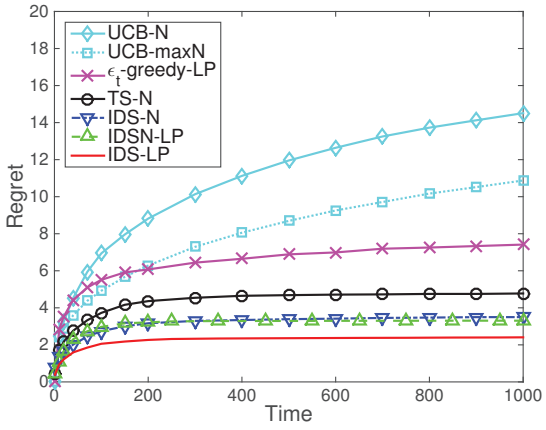
[6] Note that the vector $\boldsymbol{g}$ calculated in Russo and Van Roy (2014) is the vector $\boldsymbol{h}$ in stochastic bandits with graph feedback.

[7] It is different from the Erdős-Rényi random graph feedback since the graph is revealed to the decision maker before the decision making

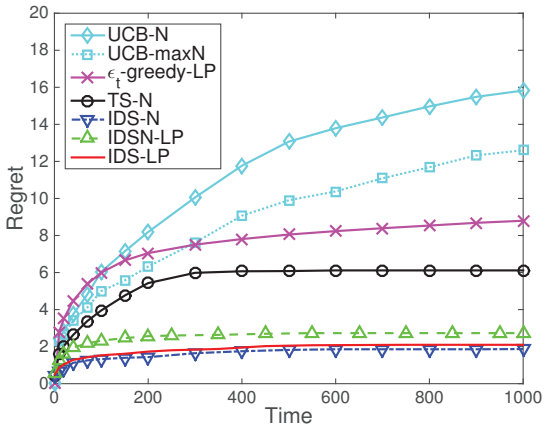
[8] The result of UCB-LP is omitted from Figure 1 because it can not be adapted to the time-variant case. Its regret result is similar to that of ϵ_t-greedy-LP in the time-invariant case.

[9] Domination number is the smallest cardinality of a dominating set, such that any node not in this set is adjacent to at least a member of this set.

[10] UCB-N is unaware of the graph structure. So it works under the random graph feedback while UCB-maxN and ϵ_t-greedy-LP do not.

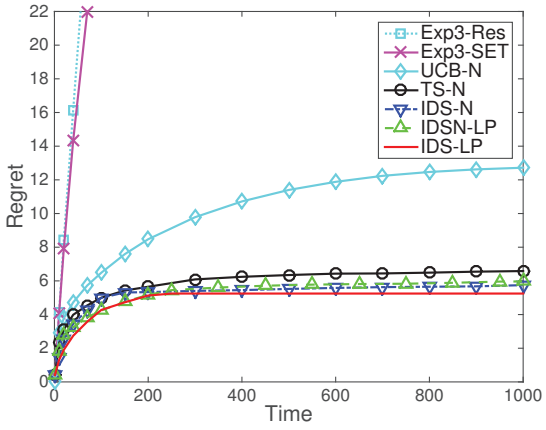


(a) Time-invariant graphs

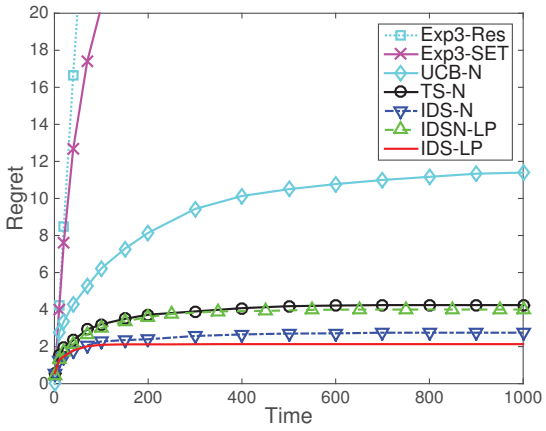


(b) Time-variant graphs

Figure 1: Regrets under the deterministic graph feedback



(a) Time-invariant $r = 0.25$



(b) Time-variant r_t

Figure 2: Regrets under the Erdős-Rényi random graph feedback

in Alon et al. (2013) and Exp3-Res) designed for the non-stochastic bandit with random graph feedback. The average regrets of Exp3-SET and Exp3-Res are dramatically larger than that of IDS based policies. For this reason, parts of Exp3-SET and Exp3-Res are omitted from Figure 2. Although, Exp3-SET and Exp3-Res have similar problem-independent upper bounds of regret, our policies utilize the stochastic model and outperform these counterparts. In addition, our IDS based policies outperform TS-N and UCB-N as well. It is interesting that IDS-LP policy performs well in both experiments though it has an upper bound that scales with the number of actions. The reason is that IDS-LP is greedy in minimizing the expected instantaneous regret, however, with guaranteed extent of exploration.

8 Conclusion

We have proposed Information Directed Sampling based policies and presented Thompson Sampling for stochastic multi-armed bandits with both deterministic and random graph feedback. We establish a unified Bayesian regret bound, that scales with the clique cover number of the graph, for TS-N, IDS-N and IDSN-LP policies under the deterministic graph case. We also present the first known theoretical guarantee, that scales with the ratio of number of actions over the expected number of observations per iteration, for TS-N, IDS-N and IDSN-LP policies under the random graph case. These results allow us to uncover the gain of partial feedback between the bandit feedback and full information feedback. Finally, we demonstrate state of art performance in numerical experiments.

This work raises the following open questions. It would be interesting to find a problem-independent regret bound that scales with the independence number of the graph instead of the clique cover number for IDS-N and IDSN-LP policies under the deterministic graph case. We believe that such improvement against TS-N can be established by exploiting the graph structure in IDS-N and IDSN-LP policies, as shown in Figure 1. Another interesting problem is to find a tighter bound for IDS-LP policy. Intuitively, IDS-LP policy can have low regret due to its greedy nature. Further, it would be an interesting extension to our work to consider

a preferential attachment random graph and other growing graphs with time to model the growth process in social networks with time.

Acknowledgment

This work has been supported in part by grants from the Army Research Office W911NF-14-1-0368 and MURI W911NF-12-1-0385, and grants from the Office of Naval Research N00014-17-1-2417 and N00014-15-1-2166.

References

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; and Mansour, Y. 2013. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, 1610–1618.

Alon, N.; Cesa-Bianchi, N.; Gentile, C.; Mannor, S.; Mansour, Y.; and Shamir, O. 2014. Nonstochastic multi-armed bandits with graph-structured feedback. *arXiv preprint arXiv:1409.8428*.

Alon, N.; Cesa-Bianchi, N.; Dekel, O.; and Koren, T. 2015. Online learning with feedback graphs: Beyond bandits. In *COLT*, 23–35.

Audibert, J.-Y., and Bubeck, S. 2010. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research* 11(Oct):2785–2836.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

Buccapatnam, S.; Liu, F.; Eryilmaz, A.; and Shroff, N. B. 2017. Reward maximization under uncertainty: Leveraging side-observations on networks. *arXiv preprint arXiv:1704.07943*.

Buccapatnam, S.; Eryilmaz, A.; and Shroff, N. B. 2014. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.* 42(1):289–300.

Caron, S.; Kveton, B.; Lelarge, M.; and Bhagat, S. 2012. Leveraging side observations in stochastic bandits. In *UAI*, 142–151. AUAI Press.

Carpentier, A., and Valko, M. 2016. Revealing graph bandits for maximizing local influence. In *International Conference on Artificial Intelligence and Statistics*, 10–18.

Chapelle, O., and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

Chen, W.; Wang, Y.; Yuan, Y.; and Wang, Q. 2016. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research* 17(50):1–33.

Cohen, A.; Hazan, T.; and Koren, T. 2016. Online learning with feedback graphs without the graphs. *CoRR* abs/1605.07018.

Kocák, T.; Neu, G.; Valko, M.; and Munos, R. 2014. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, 613–621.

Kocák, T.; Neu, G.; and Valko, M. 2016. Online learning with erdős-rényi side-observation graphs. In *Uncertainty in Artificial Intelligence*.

Liu, F.; Buccapatnam, S.; and Shroff, N. 2017. Information directed sampling for stochastic bandits with graph feedback. *arXiv preprint arXiv:1711.03198*.

Mannor, S., and Shamir, O. 2011. From bandits to experts: On the value of side-observations. In *NIPS*, 684–692.

Russo, D., and Roy, B. V. 2014. Learning to optimize via information directed sampling. *CoRR* abs/1403.5556.

Russo, D., and Van Roy, B. 2014. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, 1583–1591.

Russo, D., and Van Roy, B. 2016. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research* 17(68):1–30.

Scott, S. L. 2010. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26(6):639–658.

Seldin, Y.; Bartlett, P.; Crammer, K.; and Abbasi-Yadkori, Y. 2014. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, 280–287.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Tossou, A.; Dimitrakakis, C.; and Dubhashi, D. 2017. Thompson sampling for stochastic bandits with graph feedback. In *AAAI Conference on Artificial Intelligence*.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Valko, M. 2016. *Bandits on graphs and structures*. Ph.D. Dissertation, École normale supérieure de Cachan-ENS Cachan.

Wu, Y.; György, A.; and Szepesvári, C. 2015. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, 1360–1368.