Analysis of Thompson Sampling for Graphical Bandits Without the Graphs

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

Zizhan Zheng
Tulane University
New Orleans, LA 70118
zzheng3@tulane.edu

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

Abstract

We study multi-armed bandit problems with graph feedback, in which the decision maker is allowed to observe the neighboring actions of the chosen action, in a setting where the graph may vary over time and is never fully revealed to the decision maker. We show that when the feedback graphs are undirected, the original Thompson Sampling achieves the optimal (within logarithmic factors) regret $\tilde{O}\left(\sqrt{\beta_0(G)T}\right)$ over time horizon T, where $\beta_0(G)$ is the average independence number of the latent graphs. To the best of our knowledge, this is the first result showing that the original Thompson Sampling is optimal for graphical bandits in the undirected setting. A slightly weaker regret bound of Thompson Sampling in the directed setting is also presented. To fill this gap, we propose a variant of Thompson Sampling, that attains the optimal regret in the directed setting within a logarithmic factor. Both algorithms can be implemented efficiently and do not require the knowledge of the feedback graphs at any time.

1 INTRODUCTION

Multi-Armed Bandits (MAB) models are quintessential models for sequential decision making. In the classical MAB setting, at each time, a policy 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 T.

In this work, we consider an important variant of bandit problems, called graphical bandits, where choosing an action i not only generates a reward from action i, but also reveals observations for a subset of the remaining actions. Graphical bandits are also known as bandits with graph-structured feedback or bandits with side-observations, in which the feedback model is specified by a sequence $\{G_t\}_{t\geq 1}$ of feedback graphs. Each feedback graph G_t is a directed graph whose nodes correspond to the actions. An arc[1] (i, j) in the graph indicates that the agent observes the reward of action j if action i is chosen in that round.

Motivating examples for situations where side observations are available include viral marketing and online pricing. Consider the viral marketing problem, where a decision maker wants to find the user with the maximum influence in an online social network (e.g., Facebook) to offer a promotion (Carpentier and Valko [2016]). Each time the decision maker offers a promotion to a user, it also has an opportunity to survey the user's neighbors in the network regarding their potential interest in the same offer. 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?". Here, choosing an action in the graphical bandit problem corresponds to choosing a user in the network and side-observations across actions are captured by the links in the social network.

Consider another example in the online pricing problem, where a seller is selling goods on the Internet. In each round, the seller announces a price for the product. Then, a buyer arrives and decides whether or not to purchase the product based on its private value. A purchase takes place if and only if the announced price is no more than

---

[1] We also use the notation $i \rightarrow j$ to represent an arc from node i to node j for simplicity.

Table 1: Comparison of the existing algorithms

Algorithm	Reference	Graph	Undirected	Directed
Non-stochastic graphical bandits				
ExpBan	Mannor and Shamir [2011]	Informed	$O\left(\sqrt{\chi(G)T\log K}\right)$	
ELP	Mannor and Shamir [2011]	Informed	$O\left(\sqrt{\beta_0(G)T\log K}\right)$	$O\left(\sqrt{\chi(G)T\log K}\right)$
Exp3-SET	Alon *et al.* [2013]	Uninformed	$O\left(\sqrt{\beta_0(G)T\log K}\right)$	$O\left(\sqrt{mas(G)T\log K}\right)$
Exp3-DOM	Alon *et al.* [2013]	Informed	$O\left(\sqrt{\beta_0(G)T\log(KT)}\log K\right)$	
Exp3.G	Alon *et al.* [2015]	Uninformed	$O\left(\sqrt{\beta_0(G)T}\log(KT)\right)$	
Exp3-IX	Kocák *et al.* [2014]	Uninformed	$O\left(\sqrt{\beta_0(G)T\log K\log(KT)}\right)$	
Stochastic graphical bandits				
Cohen's Algo. 1	Cohen *et al.* [2016]	Without	$O\left(\sqrt{\beta_0(G)T\log K}\log(KT)\right)$	
IDS-N/IDSN-LP	Liu *et al.* [2018]	Informed	$O\left(\sqrt{\chi(G)T\log K}\right)$	
TS-N	Liu *et al.* [2018]	Without	$O\left(\sqrt{\chi(G)T\log K}\right)$	
TS-N	this paper	Without	$O\left(\sqrt{\beta_0(G)T\log K}\right)$	$O\left(\sqrt{mas(G)T\log K}\right)$
TS-U	this paper	Without	$O\left(\sqrt{\beta_0(G)T\log K\log(KT)}\right)$	

its private value. At the end of the round, the seller observes whether or not the buyer purchased the product at the announced price. If the buyer purchases the product, then the seller knows that the buyer would have bought the product at any lower price. Otherwise, the seller knows that the buyer would not have bought the product at any higher price. Here, actions in the graphical bandit problem corresponds to the prices that the seller can choose. The feedback graph is a directed graph over the prices that a price is connected to a lower (higher) price if and only if they are both below (above) the private value of the buyer.

Graphical bandits have been studied in both non-stochastic (adversarial) domain by Mannor and Shamir [2011]; Alon *et al.* [2013, 2015]; Kocák *et al.* [2014], and stochastic domain by Caron *et al.* [2012]; Buccapatnam *et al.* [2014, 2017]; Tossou *et al.* [2017]; Liu *et al.* [2018]. Regret bounds as a function of combinatorial properties of the feedback graphs are characterized in different settings: undirected graphs vs directed graphs, time-invariant graphs vs time-variant graphs.

However, most of the existing works mentioned above require prior knowledge of the feedback graphs for their algorithms to run. These algorithms fall into either the informed setting (where the algorithms have access to the graph structure before making decisions) or the uninformed setting (where the algorithms have access to the graph for performing their updates after decisions).

The assumption that the feedback graph is disclosed to the decision maker does not hold in many real-world applications. For example, in the viral marketing problem, the third-party decision maker is not allowed to have the knowledge of the social network in order to protect the privacy of the users. In the online pricing problem, the private value of the buyer is never revealed to the seller. Thus the feedback graph is never disclosed to the seller. This motivates us to study the graphical bandits in a setting with limited information, where the feedback graphs are *never fully revealed* to the decision maker.

In this work, we study the graphical bandits without the graphs in a general setting, where the graphs are allowed to be time-variant and directed. Moreover, the only feedback available to the decision maker at the end of each round is the out-neighborhood of the chosen action in the latent graph, along with the rewards associated with the observed actions. Generally speaking, our results show that Thompson Sampling algorithms (introduced by Thompson [1933]) can achieve a regret bound of the form $\tilde{O}\left(\sqrt{\beta_0(G)T}\right)$, where $\beta_0(G)$ is the average independence number[2] of the latent graphs, that is optimal within logarithmic factors. More specifically, we make the following contributions to graphical bandits without knowledge of the feedback graphs. (Table 1 summarizes the main results.)

---

[2]See Section 2.2 for a brief review of the combinatorial properties of graphs.

- We develop a problem-independent Bayesian regret bound for the vanilla Thompson Sampling algorithm (TS-N) for graphical bandits without the graphs. In the undirected setting, where the latent graphs are undirected, we show that TS-N obtains the optimal (within logarithmic factors) regret bound of $O\left(\sqrt{\beta_0(G)T\log K}\right)$, where $\beta_0(G)$ is the average independence number of the latent graphs (Corollary 1). Our regret bound is much sharper than the form of $O\left(\sqrt{\chi(G)T\log K}\right)$ that was shown by Liu *et al.* [2018], where $\chi(G)$ is the average clique cover number of the latent graphs, as $\beta_0(G) \leq \chi(G)$ in general. As far as we know, this is the first result showing that Thompson Sampling, without knowledge of the graph, can attain the optimal regret within logarithmic factors.

- In the directed setting, where the graphs are allowed to be directed, we show that TS-N achieves $O\left(\sqrt{mas(G)T\log K}\right)$ regret in expectation, where $mas(G)$ is the average maximal acyclic subgraph number of the latent graphs (Corollary 2). As a byproduct, our regret bounds for TS-N provide improved regret bounds for information directed sampling algorithms (IDS-N and IDSN-LP algorithms) proposed by Liu *et al.* [2018].

- We propose a variant of the Thompson Sampling algorithm, TS-U, that achieves a regret bound of $O\left(\sqrt{\beta_0(G)T\log K\log(KT)}\right)$ for both the undirected and directed setting (Corollary 3). The regret bound of TS-U is optimal within logarithmic factors, and sharper than the state-of-the-art algorithm proposed by Cohen *et al.* [2016]. Our results offer a recipe for practitioners to choose algorithms for graphical bandits without the graphs. If the latent graphs are known to be undirected, one can choose TS-N for the best regret guarantee. Otherwise, TS-U is the choice with the best guarantee.

1.1 RELATED WORK

Graphical bandits were introduced in the non-stochastic domain by Mannor and Shamir [2011]. They propose the ExpBan algorithm that works in the time-invariant and informed setting, with the regret bound depending on the clique cover number. They also propose the ELP algorithm, that replaces the uniform distribution of Exp3 algorithm (proposed by Auer *et al.* [2002b]) with a distribution that maximizes the minimum probability to observe an action. An optimal (within logarithmic factors) regret bound of ELP is shown in the undirected setting.

However, the regret bound of the ELP depends on the clique cover number in the directed setting. These results are improved by Alon *et al.* [2013]. They show that the vanilla Exp3 algorithm without mixing uniform distribution (Exp3-SET) achieves the same (but improved in the directed setting) regret bound as ELP, even in the uninformed setting. In the informed setting, they propose the Exp3-DOM algorithm, which is a variant of Exp3 algorithm with mixing uniform distribution over the dominating set of the feedback graph, achieves $\tilde{O}\left(\sqrt{\beta_0(G)T}\right)$ regret. This regret bound is further attained by Exp3.G (Alon *et al.* [2015]) and Exp3-IX (Kocák *et al.* [2014]) in the uninformed setting. The Exp3.G algorithm is a variant of Exp3-DOM where it replaces the dominating set with the universal set. The Exp3-IX algorithm uses a novel implicit exploration idea. However, these algorithms still require the knowledge of the feedback graphs for performing updates after the decisions in the uninformed setting.

Graphical bandits have also been considered in the stochastic domain by Caron *et al.* [2012], who propose a natural variant of upper confidence bounds (introduced by Auer *et al.* [2002a]) algorithm (UCB-N) and provide a problem-dependent regret guarantee depending on the clique cover number. This result is improved by Buccapatnam *et al.* [2014] in the informed and time-invariant setting. Policies proposed by Buccapatnam *et al.* [2014], namely ϵ_t-greedy-LP and UCB-LP, are shown to be asymptotically optimal, both in terms of the graph structure and time.

However, all of the afore-mentioned algorithms do not apply when the feedback graphs vary over the time and are never fully disclosed. Recently, researchers have developed new algorithms for graphical bandits in the setting with limited information, where the feedback graphs are time-variant, directed and never revealed to the decision maker. Cohen *et al.* [2016] propose an elimination-based algorithm that achieves the $\tilde{O}\left(\sqrt{\beta_0(G)T}\right)$ regret bound. Tossou *et al.* [2017] analyze the Bayesian regret performance of Thompson Sampling for the graphical bandits and provide a regret bound depending on the maximal clique cover number of the latent graphs. This result is improved to a regret bound depending on the average clique cover number by Liu *et al.* [2018]. In this work, we provide sharper regret bounds for the vanilla Thompson Sampling (TS-N) and propose a variant of Thompson Sampling (TS-U) that obtains a better (within logarithmic factor) regret bound than the algorithm developed by Cohen *et al.* [2016].

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. Graphical

bandits with Erdős-Rényi random graphs are studied by Kocák *et al.* [2016a]; Chen *et al.* [2016]; Liu *et al.* [2018]. Graphical bandits with noisy observations are studied by Kocák *et al.* [2016b]; Wu *et al.* [2015]. A survey of the graphical bandits refers to Valko [2016].

2 PROBLEM FORMULATION

2.1 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

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

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*.

2.2 GRAPH FEEDBACK MODEL

In this problem, we assume the existence of side observations, which are 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. Note that the graph G_t is *never fully revealed* to the decision maker.

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 $\chi(G)$.

Definition 2. (Independence number) An *independent set* of a graph $G = (\mathcal{K}, \mathcal{E})$ is a subset $S \subseteq \mathcal{K}$ such that no two $i, j \in \mathcal{K}$ are connected by an edge in $\mathcal{E}$. The cardinality of a largest independent set is the *independence number* of G, denoted by $\beta_0(G)$.

Note that the independence number of a directed graph is equivalent to that of the undirected graph by ignoring arc orientation. We can also lift the notion of independence number of an undirected graph to directed graph through the notion of maximum acyclic subgraphs.

Definition 3. (Maximum acyclic subgraphs) An *acyclic subgraph* of $G = (\mathcal{K}, \mathcal{E})$ is any graph $G' = (\mathcal{K}', \mathcal{E}')$ such that $\mathcal{K}' \subseteq \mathcal{K}$, and $\mathcal{E}' = \mathcal{E} \cap (\mathcal{K}' \times \mathcal{K}')$, with no directed cycles. The cardinality of the largest such $\mathcal{K}'$ is the *maximum acyclic subgraphs* number, denoted by $mas(G)$.

Note that $mas(G) \geq \beta_0(G)$ in general. The equality holds when the graph G is undirected. In this work, we slightly abuse the notation of the above graph numbers and use $\chi(G_t)$ and $\chi(\boldsymbol{G}_t)$ interchangeably since $\boldsymbol{G}_t$ fully characterizes the graph structure G_t.

2.3 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 notion of $\boldsymbol{\pi}_t$ and $\boldsymbol{\alpha}_t$ such that they represent distributions (or functions) over the 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 $I_t(A^*; Y_{t,a}) \triangleq$

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

At each time t, a randomized policy updates $\boldsymbol{\alpha}_t$, and makes a decision according to a sampling distribution $\boldsymbol{\pi}_t$. For any randomized policy, we define the *information ratio* (Russo and Van Roy [2016]) of sampling distribution $\boldsymbol{\pi}_t$ at time t as

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

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.

---

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

Algorithm 1 TS-N algorithm

Input: time horizon T
  for t **from** 1 **to** T **do**
    Updating statistics: compute $\boldsymbol{\alpha}_t$ accordingly.
    Generating policy: $\boldsymbol{\pi}_t = \boldsymbol{\alpha}_t$.
    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 latent graph.
  end for

2.4 THOMPSON SAMPLING

Thompson Sampling algorithm simply samples actions according to the posterior probability that they are optimal. In particular, actions are chosen randomly at time t according to the sampling distribution $\boldsymbol{\pi}_t = \boldsymbol{\alpha}_t$. This conceptually elegant policy can be efficiently implemented. Consider the case where $\mathcal{P} = \{p_\theta\}_{\theta \in \Theta}$ is some parametric family of distributions. The true reward distribution p^* is indexed by $\theta^* \in \Theta$ in the sense that $p^* = p_{\theta^*}$. Practical implementations of Thompson Sampling consist of two steps. First, an index $\hat{\theta}_t \sim \mathbb{P}(\theta^* \in \cdot|\mathcal{F}_t)$ is sampled from the posterior distribution. Then, the algorithm selects the action $A_t = \arg\max_{a \in \mathcal{K}} \mathbb{E}\left[Y_{t,a}|\theta^* = \hat{\theta}_t\right]$ that would be optimal if the sampled parameter were the true parameter. Given the observation of playing A_t, the posterior distribution is updated by Bayes' rule.

3 VANILLA THOMPSON SAMPLING

In this section, we show that a vanilla Thompson Sampling algorithm, TS-N as shown in Algorithm 1, obtains optimal regret (within a logarithmic factor) in the undirected setting. A slightly weaker regret bound in the directed setting is also presented.

TS-N is the Thompson Sampling algorithm for graphical bandits such that $\boldsymbol{\pi}_t = \boldsymbol{\alpha}_t$, where $\boldsymbol{\alpha}_t$ is updated based on all the observations available, without additional modifications. It naturally keeps the information ratio $\Psi_t(\boldsymbol{\pi}_t)$ bounded as well as balances between having low expected instantaneous regret (a.k.a. exploitation) and obtaining knowledge about the optimal action (a.k.a. exploration). 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., $H(\boldsymbol{\alpha}_1)$. First, we bound the information ratio of

TS-N in terms of the key quantity

$$Q_t(\boldsymbol{\pi}_t) = \sum_{i \in \mathcal{K}} \frac{\boldsymbol{\pi}_t(i)}{\sum_{j:j \overset{t}{\to} i} \boldsymbol{\pi}_t(j)}. \quad (4)$$

Note that $j \overset{t}{\to} i$ represents an arc (j, i) in graph G_t.

Proposition 1. *If $\boldsymbol{\pi}_t = \boldsymbol{\alpha}_t$, then the information ratio satisfies $\Psi_t(\boldsymbol{\pi}_t) \leq \frac{1}{2}Q_t(\boldsymbol{\pi}_t)$ almost surely.*

Proposition 1 is a tight bound for the information ratio of the vanilla Thompson Sampling (Thompson [1933]). If the graph is empty (i.e., there is no edges in the graph), then the quantity $Q_t(\boldsymbol{\pi}_t)$ equals to K. If the graph is complete, then the quantity $Q_t(\boldsymbol{\pi}_t)$ equals to 1. These recover the information ratio bounds shown by Russo and Van Roy [2016]. Also, the quantity $Q_t(\boldsymbol{\pi}_t)$ is upper bounded by clique cover number $\chi(G_t)$ as one can separate the sum by cliques and dropping the weights out of the clique. This recovers the information ratio bound shown by Liu *et al.* [2018]. Proposition 1 allows us to show a tighter regret bound of Thompson Sampling for graphical bandits. Next, we bound the regret of TS-N in terms of the quantity $Q_t(\boldsymbol{\pi}_t)$.

Theorem 1. *The regret of TS-N satisfies*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] \leq \sqrt{\frac{1}{2} \sum_{t=1}^{T} \mathbb{E}[Q_t(\boldsymbol{\alpha}_t)]H(\boldsymbol{\alpha}_1)}. \quad (5)$$

Note that the entropy is bounded, i.e., $H(\boldsymbol{\alpha}_1) \leq \log K$. It has been shown by Mannor and Shamir [2011]; Alon *et al.* [2014] that the quantity $Q_t(\boldsymbol{\pi}_t)$ is related to the graph numbers irrespective of the choice of the distribution $\boldsymbol{\pi}_t$. The following graph-theoretic result shows that the quantity $Q_t(\boldsymbol{\pi}_t)$ is bounded by the independence number of the latent graph if the graph is undirected.

Lemma 1. (Lemma 3 in Mannor and Shamir [2011]) *Let $G = (\mathcal{K}, \mathcal{E})$ be an undirected graph. For any distribution $\boldsymbol{\pi}$ over $\mathcal{K}$,*

$$\sum_{i \in \mathcal{K}} \frac{\boldsymbol{\pi}(i)}{\sum_{j:j \overset{t}{\to} i} \boldsymbol{\pi}(j)} \leq \beta_0(G). \quad (6)$$

The following regret result of TS-N follows immediately from Theorem 1 and Lemma 1.

Corollary 1. *In the undirected setting, the regret of TS-N satisfies*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] \leq \sqrt{\frac{1}{2} \sum_{t=1}^{T} \beta_0(G_t)H(\boldsymbol{\alpha}_1)}. \quad (7)$$

As far as we know, this is the best regret bound for graphical bandits without the graphs. First, an information-theoretic lower bound of graphical bandits has been

shown by Mannor and Shamir [2011]; Alon *et al.* [2014] to be $\Omega\left(\sqrt{\beta_0(G)T}\right)$. So Corollary 1 shows that TS-N obtains the optimal regret (within a logarithmic factor) in the undirected setting. Moreover, the bound proven in Corollary 1 is tighter than the $O\left(\sqrt{\chi(G)TH(\boldsymbol{\alpha}_1)}\right)$ bound of TS-N shown by Liu *et al.* [2018] as $\beta_0(G) \leq \chi(G)$. At last, Corollary 1 shows that TS-N enjoys better regret bound than Cohen's Algorithm 1 developed by Cohen *et al.* [2016] in the undirected setting, both of which do not require the knowledge of the feedback graphs.

We now turn to the directed setting. The following graph-theoretic result shows that the quantity $Q_t(\boldsymbol{\pi}_t)$ is upper-bounded by the maximum acyclic subgraph number if the graph is directed.

Lemma 2. (Lemma 10 in Alon *et al.* [2014]) *Let $G = (\mathcal{K}, \mathcal{E})$ be a directed graph. For any distribution $\boldsymbol{\pi}$ over $\mathcal{K}$,*

$$\sum_{i \in \mathcal{K}} \frac{\boldsymbol{\pi}(i)}{\sum_{j:j \overset{t}{\to} i} \boldsymbol{\pi}(j)} \leq mas(G). \quad (8)$$

The following regret result of TS-N follows immediately from Theorem 1 and Lemma 2.

Corollary 2. *In the directed setting, the regret of TS-N satisfies*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] \leq \sqrt{\frac{1}{2} \sum_{t=1}^{T} mas(G_t)H(\boldsymbol{\alpha}_1)}. \quad (9)$$

The bound proven in Corollary 2 is tighter than the $O\left(\sqrt{\chi(G)TH(\boldsymbol{\alpha}_1)}\right)$ bound of TS-N shown by Liu *et al.* [2018] since $mas(G) \leq \chi(G)$. However, there is a gap between the lower bound and the upper bound shown in Corollary 2. Though $\beta_0(G) = mas(G)$ when the graph is undirected, the gap between them can be large in general directed graphs. For example, consider a directed graph $G_0 = (\mathcal{K}, \mathcal{E})$ such that arc $(i, j) \in \mathcal{E}$ if and only if $i \leq j$. It is clear that $\beta_0(G_0) = 1$ and $mas(G_0) = K$. This leads us to consider a more sophisticated randomized policy. In the next section, we show that a modified Thompson Sampling algorithm results in an optimal (within a logarithmic factor) regret bound in the general setting.

Remark 1. *The bounds proven in Corollary 1 and 2 hold for IDS-N and IDSN-LP developed by Liu* et al. *[2018] since the information ratio of IDS-N and IDSN-LP are bounded by the information ratio of TS-N almost surely. So our results also provide tighter bounds for IDS-N and IDSN-LP algorithms. Note that IDS-N and IDSN-LP algorithms require prior knowledge of the feedback graphs.*

Algorithm 2 TS-U algorithm
___
Input: time horizon T and parameter $\epsilon \in [0, 1]$
  for t **from** 1 **to** T **do**
    Updating statistics: compute $\boldsymbol{\alpha}_t$ accordingly.
    Generating policy: $\boldsymbol{\pi}_t = (1 - \epsilon)\boldsymbol{\alpha}_t + \epsilon/K$.
    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 latent graph.
  end for
___

4 THOMPSON SAMPLING WITH EXPLORATION

In this section, we show that a mixture of Thompson Sampling and uniform sampling, TS-U as shown in Algorithm 2, obtains the optimal regret within a logarithmic factor in the directed setting.

TS-U algorithm is a variant of Thompson Sampling with explicit exploration that allows the algorithm to explore some suboptimal actions with large out-degrees. The collected side observations from the uniform sampling allows us to capture the latent graph information, thus yielding a regret bound in terms of the independence number.

As shown in Algorithm 2, TS-U is a randomized policy such that $\boldsymbol{\pi}_t = (1 - \epsilon)\boldsymbol{\alpha}_t + \epsilon/K$ for some parameter $\epsilon \in [0, 1]$. The implementation and computation of TS-U is quite efficient. At each time t, TS-U algorithm plays vanilla Thompson Sampling with probability $1 - \epsilon$ and plays uniform sampling with probability ϵ. While the expected information gain diminishes, the expected instantaneous regret is bounded away from zero due to uniform sampling. Thus, the classical analysis of bounding the information ratio $\Psi_t(\boldsymbol{\pi}_t)$ does not work any more. Fortunately, the linear form of $\boldsymbol{\pi}_t$ allows us to bound the regret of TS-U by the regret due to uniform sampling plus the regret from Thompson Sampling, as shown in Theorem 2. Indeed, our techniques work for any variant of Thompson Sampling that is a linear combination of $\boldsymbol{\alpha}_t$ and some other distributions.

Theorem 2. *The regret of TS-U satisfies*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] \leq \epsilon T + \sqrt{\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[Q_t(\boldsymbol{\pi}_t)]H(\boldsymbol{\alpha}_1)}. \quad (10)$$

Theorem 2 shows that the regret of TS-U consists of two parts, the regret from the uniform sampling and the regret from the Thompson Sampling. Note that the regret from the Thompson Sampling takes into account the information gain from the uniform sampling as the term

$Q_t(\boldsymbol{\pi}_t)$ depends on $\boldsymbol{\pi}_t$ rather than $\boldsymbol{\alpha}_t$. This allows us to use the following graph-theoretic result to bound the term $Q_t(\boldsymbol{\pi}_t)$, thus the regret of TS-U, by the independence number of the latent graph.

Lemma 3. (Lemma 5 in Alon *et al.* [2015]) *Let* $G = (\mathcal{K}, \mathcal{E})$ *be a directed graph. For any distribution* $\boldsymbol{\pi}$ *over* $\mathcal{K}$ *such that* $\boldsymbol{\pi}(i) \geq \eta$ *for all* $i \in \mathcal{K}$ *for some constant* $0 < \eta < 0.5$. *Then*

$$\sum_{i \in \mathcal{K}} \frac{\boldsymbol{\pi}(i)}{\sum_{j:j \xrightarrow{t} i} \boldsymbol{\pi}(j)} \leq 4\beta_0(G)\log\left(\frac{4K}{\beta_0(G)\eta}\right). \quad (11)$$

Lemma 3 shows that the quantity $Q_t(\boldsymbol{\pi}_t)$ can be bounded by the independence number of the directed graph if $\boldsymbol{\pi}_t$ is bounded away from zero. The uniform sampling part of TS-U allows the sampling distribution to satisfy this condition. By Theorem 2 and taking $\eta = \epsilon/K$ in Lemma 3, we have the following result.

Corollary 3. *If* $\epsilon = 1/\sqrt{T}$, *then the regret of TS-U satisfies*

$$\mathbb{E}[R(T, \boldsymbol{\pi})] = O\left(\sqrt{\log(KT)\sum_{t=1}^{T}\beta_0(G_t)H(\boldsymbol{\alpha}_1)}\right). \quad (12)$$

Comparing the regret bound in Corollary 3 to the lower bound, $\Omega(\sqrt{\beta_0(G)T})$, the TS-U algorithm obtains the optimal regret within a logarithmic factor in the general setting. Moreover, Corollary 3 shows that TS-U enjoys a sharper (by a logarithmic factor) regret bound than Cohen's Algorithm 1 developed by Cohen *et al.* [2016] in the directed setting, both of which do not require the knowledge of the feedback graphs. As far as we know, this is the best-known regret bound for graphical bandits without the graphs. Finally, note that a comparison between Corollary 1 and Corollary 3 reveals that a symmetric observation system (i.e., undirected feedback graphs) enjoys better regret than an asymmetric observation system as the regret bound of TS-N is sharper by a logarithmic factor than the bound in Corollary 3 in the undirected setting.

Remark 2. *We present TS-U algorithm with fixed exploration rate* ϵ *for simplicity. It is easy to verify that the regret result of TS-U still holds if one uses some appropriate decreasing exploration rate sequence* $\{\epsilon_t\}$. *For example, when* $\epsilon_t = 1/t$, *Theorem 2 holds by replacing the term* ϵT *with* $\log T$. *Then the regret result of the corresponding TS-U algorithm follows. In practice, we recommend practitioners to use decreasing exploration rates.*
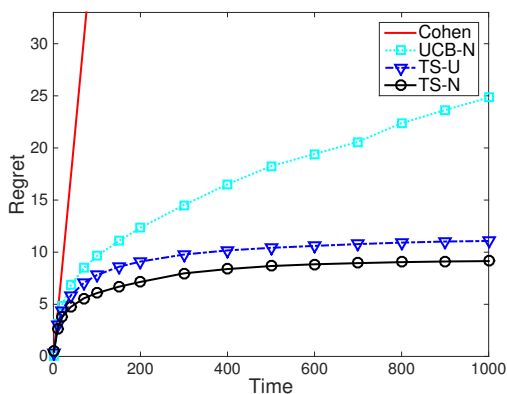
Remark 3. *In the informed setting (i.e., when the feedback graph are revealed to the decision maker before the*
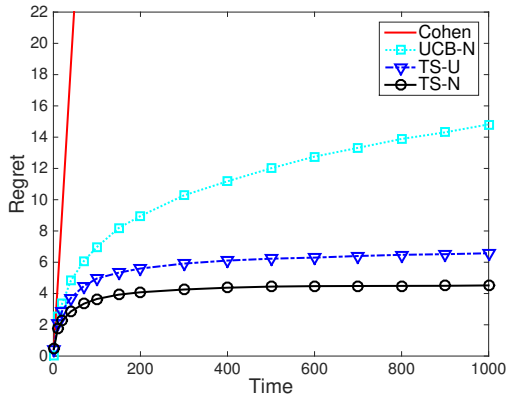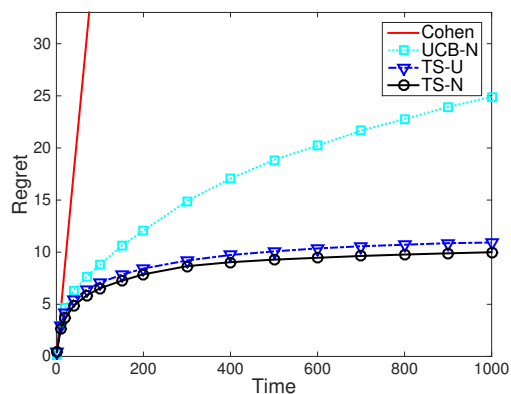
(a) Time-invariant graphs
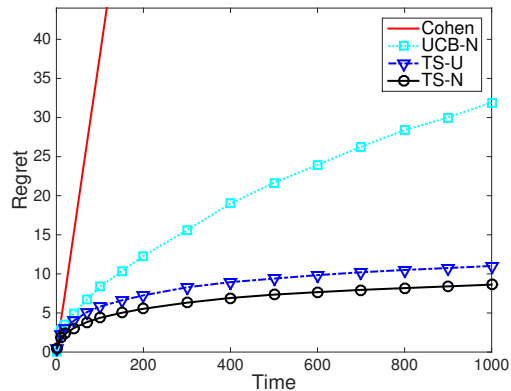
(a) Time-invariant graphs

(b) Time-variant graphs

(b) Time-variant graphs

Figure 1: Regret comparison under undirected graph feedback

Figure 2: Regret comparison under directed graph feedback

decisions), one may propose a variant of TS-U such that it restricts the exploration set to the dominating set of the feedback graph. In other words, one may replace uniform sampling over all the actions with uniform sampling over only the dominating set. The regret bound in Corollary 3 still holds by a variant of Lemma 3 shown in Alon et al. *[2014].*

5 NUMERICAL RESULTS

This section presents numerical results from experiments that evaluate the effectiveness of Thompson Sampling based policies in comparison to UCB-N and Cohen's algorithm. We consider the classical Beta-Bernoulli bandit problem with independent actions. The reward of each action i is a Bernoulli(μ_i) random variable and μ_i is independently drawn from Beta$(1, 1)$. The implementations of TS-N and TS-U are shown in Algorithms 3 and 4. In the experiment, we set $K = 5$, $T = 1000$ and $\epsilon_t = 1/t$ as suggested by Remark 2. All the regret results

are averaged over 1000 trials.

Figure 1 presents the cumulative regret results under undirected graph feedback. For the time-invariant case shown in Figure 1a, we use a graph with 2 cliques, presented in Figure 3. Figure 2 presents the cumulative regret results under directed graph feedback. For the time-invariant case shown in Figure 2a, we use the graph $G = (\mathcal{K}, \mathcal{E})$ such that $(i, j) \in \mathcal{E}$ if and only $i \leq j$, presented in Figure 4. For the time-variant cases shown in Figures 1b and 2b, the sequences of graphs are generated by the Erdős-Rényi model with parameter p_t[4] drawn from the uniform distribution over $[0, 0.2]$.

We find that TS-N and TS-U outperform the alternative algorithms, which is consistent with the empirical observation in the bandit feedback setting (Chapelle and Li [2011]). However, TS-N has better empirical performance in the tested settings even though we have

---

[4]For each time t and each pair (i, j), $(i, j) \in \mathcal{E}_t$ with probability p_t

Algorithm 3 TS-N (Bernoulli case)

Input: time horizon T
  For each arm i, set $S_i = 1$ and $F_i = 1$
  for t **from** 1 **to** T **do**
    For each arm i, sample θ_i from Beta(S_i, F_i).
    Play action $A_t = \arg\max_{i \in \mathcal{K}} \theta_i$.
    for all $a \in \mathcal{K}$ such that $(A_t, a) \in \mathcal{E}_t$ **do**
      $S_a = S_a + Y_{t,a}$ and $F_a = F_a + 1 - Y_{t,a}$.
    end for
  end for

Algorithm 4 TS-U (Bernoulli case)

Input: time horizon T and $\{\epsilon_t\}_{t \geq 1}$
  For each arm i, set $S_i = 1$ and $F_i = 1$
  for t **from** 1 **to** T **do**
    Sample β_t from uniform distribution over $[0, 1]$.
    if $\beta_t < \epsilon_t$ **then**
      Play action A_t drawn uniformly from $\mathcal{K}$.
    else
      For each arm i, sample θ_i from Beta(S_i, F_i).
      Play action $A_t = \arg\max_{i \in \mathcal{K}} \theta_i$.
    end if
    for all $a \in \mathcal{K}$ such that $(A_t, a) \in \mathcal{E}_t$ **do**
      $S_a = S_a + Y_{t,a}$ and $F_a = F_a + 1 - Y_{t,a}$.
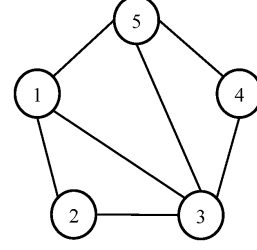   **end for**
  end for

Figure 3: Graph structure for the experiment under time-invariant and undirected graph
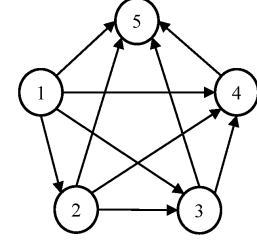
Figure 4: Graph structure for the experiment under time-invariant and directed graph

proven a better regret bound for TS-U under directed feedback graphs. The average regrets of Cohen's algorithm are dramatically larger than that of Thompson Sampling based policies. For this reason, parts of Cohen's algorithm are omitted from Figures 1 and 2.

6 CONCLUSION

We have provided regret analysis of Thompson Sampling for graphical bandits without knowing the feedback graphs at any time. We show that the regret of TS-N is bounded by $O\left(\sqrt{mas(G)T \log K}\right)$ in the general setting. In the undirected setting, $mas(G) = \beta_0(G)$, and the resulting regret bound is optimal up to a logarithmic factor. As far as we know, this is the first result that shows that Thompson Sampling, even without the knowledge of the graph, can attain the optimal regret in the graphical bandits. As a byproduct, our analysis for TS-N provide improved regret bounds for information directed sampling algorithms (IDS-N and IDSN-LP algorithms) proposed by Liu *et al.* [2018] in the informed setting.

We have proposed a variant of Thompson Sampling, TS-U, that mixes Thompson Sampling with

uniform sampling. This modification allows the algorithm to capture the graph structure and obtain $O\left(\sqrt{\beta_0(G)T \log K \log(KT)}\right)$ regret bound in the directed setting, which is optimal within a logarithmic factor. Our results offer a recipe for practitioners to choose algorithms for graphical bandits without knowledge of the graphs. If the latent graphs are known to be undirected, one can choose TS-N for the best regret guarantee. Otherwise, TS-U is the choice with the best guarantee in the general (directed) setting.

References

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems*, pages 1610–1618, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Non-stochastic multi-armed bandits with graph-structured feedback. *arXiv preprint arXiv:1409.8428*, 2014.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *COLT*, pages 23–35, 2015.

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

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. *SIGMETRICS Perform. Eval. Rev.*, 42(1):289–300, June 2014.

Swapna Buccapatnam, Fang Liu, Atilla Eryilmaz, and Ness B Shroff. Reward maximization under uncertainty: Leveraging side-observations on networks. *arXiv preprint arXiv:1704.07943*, 2017.

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

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

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

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

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. *CoRR*, abs/1605.07018, 2016.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, pages 613–621, 2014.

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with erdős-rényi side-observation graphs. In *Uncertainty in Artificial Intelligence*, 2016.

Tomás Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *AISTATS*, pages 1186–1194, 2016.

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *AAAI*, 2018.

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

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

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287, 2014.

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

Aristide Tossou, Christos Dimitrakakis, and Devdatt Dubhashi. Thompson sampling for stochastic bandits with graph feedback. In *AAAI Conference on Artificial Intelligence*, 2017.

Michal Valko. *Bandits on graphs and structures*. PhD thesis, École normale supérieure de Cachan-ENS Cachan, 2016.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.