Unimodal Thompson Sampling for Graph–Structured Arms

Stefano Paladino, Francesco Trovò, Marcello Restelli, Nicola Gatti

Dipartimento di Elettronica, Informazione e Bioingegneria

Politecnico di Milano, Milano, Italy

{stefano.paladino, francesco1.trovo, marcello.restelli, nicola.gatti}@polimi.it

Abstract

We study, to the best of our knowledge, the first Bayesian algorithm for unimodal Multi–Armed Bandit (MAB) problems with graph structure. In this setting, each arm corresponds to a node of a graph and each edge provides a relationship, unknown to the learner, between two nodes in terms of expected reward. Furthermore, for any node of the graph there is a path leading to the unique node providing the maximum expected reward, along which the expected reward is monotonically increasing. Previous results on this setting describe the behavior of frequentist MAB algorithms. In our paper, we design a Thompson Sampling–based algorithm whose asymptotic pseudo–regret matches the lower bound for the considered setting. We show that—as it happens in a wide number of scenarios—Bayesian MAB algorithms dramatically outperform frequentist ones. In particular, we provide a thorough experimental evaluation of the performance of our and state–of–the–art algorithms as the properties of the graph vary.

Introduction

Multi–Armed Bandit (MAB) algorithms (Auer, Cesa-Bianchi, and Fischer 2002) have been proven to provide effective solutions for a wide range of applications fitting the sequential decisions making scenario. In this framework, at each round over a finite horizon T, the learner selects an action (usually called *arm*) from a finite set and observes only the reward corresponding to the choice she made. The goal of a MAB algorithm is to converge to the optimal arm, i.e., the one with the highest expected reward, while minimizing the loss incurred in the learning process and, therefore, its performance is measured through its expected *pseudo–regret*, defined as the difference between the expected reward achieved by an oracle algorithm always selecting the optimal arm and the one achieved by the considered algorithm. We focus on the so–called *Unimodal* MAB (UMAB), introduced in (Combes and Proutiere 2014a), in which each arm corresponds to a node of a graph and each edge is associated with a relationship specifying which node of the edge gives the largest expected reward (providing thus a partial ordering over the arm space). Furthermore, from any node there is a path leading to the unique node with the maximum expected reward along which the expected reward is

monotonically increasing. While the graph structure may be (not necessarily) known *a priori* by the UMAB algorithm, the relationship defined over the edges is discovered during the learning. In the present paper, we propose a novel algorithm relying on the Bayesian learning approach for a generic UMAB setting.

Models presenting a graph structure have become more and more interesting in last years due to the spread of social networks. Indeed, the relationships among the entities of a social network have a natural graph structure. A practical problem in this scenario is the *targeted advertisement* problem, whose goal is to discover the part of the network that is interested in a given product. This task is heavily influenced by the graph structure, since in social networks people tend to have similar characteristics to those of their friends (i.e., neighbor nodes in the graph), therefore interests of people in a social network change smoothly and neighboring nodes in the graph look similar to each other (McPherson, Smith-Lovin, and Cook 2001; Crandall et al. 2008). More specifically, an advertiser aims at finding those users that maximize the ad expected revenue (i.e., the product between click probability and value per click), while at the same time reducing the amount of times the advertisement is presented to people not interested in its content.

Under the assumption of unimodal expected reward, the learner can move from low expected rewards to high ones just by climbing them in the graph, preventing from the need for an exploration over all the graph nodes. This assumption reduces the complexity in the search for the optimal arm, since the learning algorithm can avoid to pull the arms corresponding to some subset of non–optimal nodes, reducing thus the regret. Other applications might benefit from this structure, e.g., *recommender systems* which aims at coupling items with those users are likely to enjoy them. Similarly, the use of the unimodal graph structure might provide more meaningful recommendations without testing all the users in the social network. Finally, notice that unimodal problems with a single variable, e.g., in sequential pricing (Jia and Mannor 2011), bidding in online sponsored search auctions (Edelman and Ostrovsky 2007) and single–peak preferences economics and voting settings (Mas-Collel, Whinston, and Green 1995), are graph–structured problems in which the graph is a line.

Frequentist approaches for UMAB with graph structure are proposed in (Jia and Mannor 2011) and (Combes and Proutiere 2014a). Jia and Mannor (2011) introduce the GLSE algorithm with a regret of order $O(\sqrt{T}\log(T))$. However, GLSE performs better than classical bandit algorithms only when the number of arms is $\Theta(T)$. Combes and Proutiere (2014a) present the OSUB algorithm—based on KLUCB—achieving asymptotic regret of $O(\log(T))$ and outperforming GLSE in settings with a few arms. To the best of our knowledge, no Bayesian approach has been proposed for unimodal bandit settings, included the UMAB setting we study. However, it is well known that Bayesian MAB algorithms—the most popular is Thompson Sampling (TS)—usually suffer of same order of regret as the best frequentist one (e.g., in unstructured settings (Kaufmann, Korda, and Munos 2012)), but they outperform the frequentist methods in a wide range of problems (e.g., in bandit problems without structure (Chapelle and Li 2011) and in bandit problems with budget (Xia et al. 2015)). Furthermore, in problems with structure, the classical Thompson Sampling (not exploiting the problem structure) may outperform frequentist algorithms exploiting the problem structure. For this reason, in this paper we explore Bayesian approaches for the UMAB setting. More precisely, we provide the following original contributions:

- we design a novel Bayesian MAB algorithm, called UTS and based on the TS algorithm;

- we derive a tight upper bound over the pseudo–regret for UTS, which asymptotically matches the lower bound for the UMAB setting;

- we describe a wide experimental campaign showing better performance of UTS in applicative scenarios than those of state–of–the–art algorithms, evaluating also how the performance of the algorithms (ours and of the state of the art) varies as the graph structure properties vary.

Related work

Here, we mention the main works related to ours. Some works deal with unimodal reward functions in continuous armed bandit setting (Jia and Mannor 2011; Combes and Proutiere 2014b; Kleinberg, Slivkins, and Upfal 2008). In (Jia and Mannor 2011) a successive elimination algorithm, called LSE, is proposed achieving regret of $O(\sqrt{T}\log T)$. In this case, assumptions over the minimum local decrease and increase of the expected reward is required. Combes and Proutiere (2014b) consider stochastic bandit problems with a continuous set of arms and where the expected reward is a continuous and unimodal function of the arm. They propose the SP algorithm, based on the stochastic pentachotomy procedure to narrow the search space. Unimodal MABs on metric spaces are studied in (Kleinberg, Slivkins, and Upfal 2008).

An application–dependent solution to the recommendation systems which exploits the similarity of the graph in social network in targeted advertisement has been proposed in (Valko et al. 2014). Similar information has been considered in (Caron and Bhagat 2013) where the problem of

cold–start users (i.e., new users) is studied. Another type of structure considered in sequential games is the one of *monotonicity* of the conversion rate in the price (Trovò et al. 2015). Interestingly, the assumptions of monotonicity and unimodality are orthogonal, none of them being a special case of the other, therefore the results for monotonic setting cannot be used in unimodal bandits. In (Alon et al. 2013; Mannor and Shamir 2011), a graph structure of the arm feedback in an *adversarial* setting is studied. More precisely, they assume to have correlation over rewards and not over the expected values of arms.

Problem Formulation

A learner receives in input a finite undirected graph MAB setting $G = (A, E)$, whose vertices $A = \{a_1, \ldots, a_K\}$ with $K \in \mathbb{N}$ correspond to the arms and an edge $(a_i, a_j) \in E$ exists only if there is a direct partial order relationship between the expected rewards of arms a_i and a_j. The leaner knows *a priori* the nodes and the edges (i.e., she knows the graph), but, for each edge, she does not know *a priori* which is the node of the edge with the largest expected reward (i.e., she does not know the ordering relationship). At each round t over a time horizon of $T \in \mathbb{N}$ the learner selects an arm a_i and gains the corresponding reward $x_{i,t}$. This reward is drawn from an i.i.d. random variable $X_{i,t}$ (i.e., we consider a stochastic MAB setting) characterized by an unknown distribution $\mathcal{D}_i$ with finite known support $\Omega \subset \mathbb{R}$ (as customary in MAB settings, from now on we consider $\Omega \subseteq [0, 1]$) and by unknown expected value $\mu_i := \mathbb{E}[X_{i,t}]$. We assume that there is a single optimal arm, i.e., there exists a unique arm a_{i^*} s.t. its expected value $\mu_{i^*} = \max_i \mu_i$ and, for sake of notation, we denote μ_{i^*} with μ^*.

Here we analyze a graph bandit setting with unimodality property, defined as:

Definition 1. *A graph* unimodal MAB *(UMAB) setting* $G = (A, E)$ *is a graph bandit setting G s.t. for each sub–optimal arm $a_i, i \neq i^*$ it exists a finite path $p = (i_1 = i, \ldots, i_m = i^*)$ s.t. $\mu_{i_k} < \mu_{i_{k+1}}$ and $(a_{i_k}, a_{i_{k+1}}) \in E$ for each $k \in \{1, \ldots, m-1\}$.*

This definition assures that if one is able to identify a non–decreasing path in G of expected rewards, she be able to reach the optimum arm, without getting stuck in local optima. Note that the unimodality property implies that the graph G is connected and therefore we consider only connected graphs from here on.

A policy $\mathfrak{U}$ over a UMAB setting is a procedure able to select at each round t an arm a_{i_t} by basing on the history h_t, i.e., the sequence of past selected arms and past rewards gained. The pseudo–regret $R_T(\mathfrak{U})$ of a generic policy $\mathfrak{U}$ over a UMAB setting is defined as:

$$R_T(\mathfrak{U}) := T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_{i_t, t}\right], \qquad (1)$$

where the expected value $\mathbb{E}[\cdot]$ is taken w.r.t. the stochasticity of the gained rewards $X_{i_t, t}$ and of the policy $\mathfrak{U}$.

Let us define the neighborhood of arm a_i as $N(i) := \{j | (a_i, a_j) \in E\}$, i.e., the set of each index j of the arm a_j

connected to the arm a_i by an edge $(a_i, a_j) \in E$. It has been shown in (Combes and Proutiere 2014a) that the problem of learning in a UMAB setting presents a lower bound over the regret $R_T(\mathfrak{U})$ of the following form:

Theorem 1. *Let $\mathfrak{U}$ be a uniformly good policy, i.e., a policy s.t. $R_T(\mathfrak{U}) = o(T^c)$ for each $c > 0$. Given a UMAB setting $G = (A, E)$ we have:*

$$\liminf_{T \to \infty} \frac{R_T(\mathfrak{U})}{\log(T)} = \sum_{i \in N(i^*)} \frac{\mu^* - \mu_i}{KL(\mu_i, \mu^*)}, \qquad (2)$$

where $KL(p, q) = p \log\left(\frac{p}{q}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right)$, i.e., the Kullaback–Leibler divergence of two Bernoulli distributions with means p and q, respectively.

This result is similar to the one provided in (Lai and Robbins 1985), with the only difference that the summation is restricted to the arms laying in the neighborhood of the optimal arm $N(i^*)$ and reduces to it when the optimal arm is connected to all the others (i.e., $N(i^*) \equiv \{1, \ldots, K\}$) or the graph is completely connected (i.e., $N(i) \equiv \{1, \ldots, K\}, \forall i$). We would like to point out that by relying on the assumption of having a single maximum of the expected rewards, we also assure that the optimal arm neighborhood $N(i^*)$ is uniquely defined and, thus, the lower bound inequality in Equation 2 is well defined.

The UTS algorithm

We describe the UTS algorithm and we show that its regret is asymptotically optimal, i.e., it asymptotically matches the lower bound of Theorem 1. The algorithm is an extension of the Thompson Sampling (Thompson 1933) that exploits the graph structure and the unimodal property of the UMAB setting. Basically, the rationale of the algorithm is to apply a simple variation of the TS algorithm to only the arms associated with the nodes that compose the neighborhood of the arm with the highest empirical mean reward, called *leader*.

The UTS pseudo–code

The pseudo–code of the UTS algorithm is presented in Algorithm 1. The algorithm receives in input the graph structure G, the time horizon T, and a Bayesian prior π_i for each expected reward μ_i. At each round t, the algorithm computes the empirical expected reward for each arm (Line 3):

$$\hat{\mu}_{i,t} := \begin{cases} \dfrac{S_{i_t}}{T_{i,t}} & \text{if } T_{i,t} > 0 \\ 0 & \text{otherwise} \end{cases},$$

where $S_{i,t} = \sum_{h=1}^{t-1} X_{i,h} \mathbf{1}\{\mathfrak{U}(h) = a_i\}$ is the cumulative reward of arm a_i up to round t and $T_{i,t} = \sum_{h=1}^{t-1} \mathbf{1}\{\mathfrak{U}(h) = a_i\}$ is the number of times the arm a_i has been pulled up to round t.[1] After that, UTS selects the arm denoted as the leader $a_{l(t)}$ for round t, i.e., the one having the maximum empirical expected reward:

$$a_{l(t)} = \arg\max_{a_i \in A} \hat{\mu}_{i,t}. \qquad (3)$$

Algorithm 1 UTS

1: **Input:** UMAB setting $G = (V, E)$, Horizon T, Priors $\{\pi_i\}_{i=1}^K$
2: **for** $t \in \{1, \ldots, T\}$ **do**
3:     Compute $\hat{\mu}_{i,t}$ for each $i \in \{1, \ldots, K\}$
4:     Find the leader $a_{l(t)}$
5:     **if** $L_{l(t),t} \bmod |N^+(l(t))| = 0$ **then**
6:         Collect reward $x_{l(t),t}$
7:     **else**
8:         Draw $\theta_{i,t}$ from $\pi_{i,t}$ for each $i \in N^+(l(t))$
9:         Collect reward $x_{i_t,t}$ where $i_t = \arg\max_i \theta_{i,t}$

Once the leader has been chosen, we restrict the selection procedure to it and its neighborhood, considering only arms with indexes in $N^+(l(t)) := N(l(t)) \cup \{l(t)\}$. Denote with $L_{i,t} := \sum_{h=1}^{t-1} \mathbf{1}\{l(h) = i\}$ the number of times the arm a_i has been selected as leader before round t. If $L_{l(t),t}$ is a multiple of $|N^+(l(t))|$, then the leader is pulled and reward $x_{l(t),t}$ is gained (Line 6).[2] Otherwise, the TS algorithm is performed over arms a_i s.t. $i \in N^+(l(t))$ (Lines 8–9).

Basically, under the assumption of having a prior π_i, we can compute the posterior distribution $\pi_{i,t}$ for μ_i after t rounds, using the information gathered from the rounds in which a_i has been pulled. We denote with $\theta_{i,t}$ a sample drawn from $\pi_{i,t}$, called *Thompson sample*. For instance, for Bernoulli rewards and by assuming uniform priors we have that $\pi_{i,t} = Beta(1 + S_{i,t}, 1 + T_{i,t} - S_{i,t})$, where $Beta(\alpha, \beta)$ is the beta distribution with parameters α and β. Finally, the UTS algorithm pulls the arm with the largest Thompson sample $\theta_{i,n}$ and collects the corresponding reward $x_{i_t,t}$. See (Kaufmann, Korda, and Munos 2012) for further details.

Remark 1. *Assuming that the UTS algorithm receives in input the whole graph G is unnecessary. The algorithm just requires an oracle that, at each round t, is able to return the neighborhood $N(l(t))$ of the arm which is currently the leader $a_{l(t)}$. This is crucial in all the applications in which the graph is discovered by means of a series of queries and the queries have a non–negligible cost (e.g., in social networks a query might be computationally costly). Finally, we remark that the frequentist counterpart of our algorithm (i.e., the OSUB algorithm) requires the computation of the maximum node degree $\gamma := \max_i |N(i)|$, thus requiring at least an initial analysis of the entire graph G.*

Finite–time analysis of UTS

Theorem 2. *Given a UMAB setting $G = (A, E)$, the expected pseudo–regret of the UTS algorithm satisfies, for every $\varepsilon > 0$:*

$$\mathcal{R}_T(UTS) \leq (1 + \varepsilon) \sum_{i \in N(i^*)} \frac{\mu^* - \mu_i}{KL(\mu_i, \mu^*)} [\log(T) + \log\log(T)] + \tilde{C},$$

where $\tilde{C} > 0$ is a constant depending on ε, the number of arms K and the expected rewards $\{\mu_1, \ldots, \mu_K\}$.

---

[1] We here denote with $\mathbf{1}\{\cdot\}$ the indicator function.

[2] We here denote with $|\cdot|$ the cardinality operator.

Sketch of proof. (The complete version of the proof is reported in (Paladino et al. 2016).) At first, we remark that a straightforward application of the proof provided for OSUB is not possible in the case of UTS. Indeed, the use of frequentist upper bounds over the expected reward in OSUB implies that in finite time and with high probability the bounds are ordered as the expected values. Since we are using a Bayesian algorithm, we would require the same assurance over the Thompson samples $\theta_{i,t}$, but we do not have a direct bound over $\mathbb{P}(\theta_{i,t} > \theta_{i',t})$ where $a_{i'}$ is the optimal arm in the neighborhood $N^+(i)$. This fact requires to follow a completely different strategy when we analyze the case in which the leader is not the optimal arm.

The regret of the UTS algorithm $R_T(\text{UTS})$ can be divided in two parts: the one obtained during those rounds in which the optimal arm a^* is the leader, called $\mathcal{R}_1$, and the summation of the regrets in the rounds in which the leader is the arm $a_i \neq a^*$, called $\mathcal{R}_i$. $\mathcal{R}_1$ is obtained when i^* is the leader, thus, the UTS algorithms behaves like Thompson Sampling restricted to the optimal arm and its neighborhood $N^+(i^*)$, and the regret upper bound is the one derived in (Kaufmann, Korda, and Munos 2012) for the TS algorithm.

$\mathcal{R}_i$ is upper bounded by the expected number of rounds the arm a_i has been selected as leader $\mathbb{E}[L_{i,T}]$ over the horizon T. Let us consider $\hat{L}_{i,T}$ defined as the number of rounds spent with a_i as leader when restricting the problem to its neighborhood $N^+(i)$. $\mathbb{E}[\hat{L}_{i,T}]$ is an upper bound over $\mathbb{E}[L_{i,T}]$, since there is nonzero probability that the UTS algorithm moves in another neighborhood. Since $i \neq i^*$ and the setting is unimodal, there exists an optimal arm $a_{i'}, i' \neq i$ among those in the neighborhood $N(i)$ s.t. $\mu_{i'} = \max_{i|a_i \in N(i)} \mu_i$ and $\hat{\mu}_{i,t} \geq \hat{\mu}_{i'}$. Thus:

$$\mathcal{R}_i \leq \mathbb{E}[\hat{L}_{i,T}] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\{\hat{\mu}_{i,t} = \max_{a_j \in N^+(i)} \hat{\mu}_{j,t}\}\right]$$

$$= \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t} \geq \max_{a_j \in N^+(i)} \hat{\mu}_{j,t}\right) \leq \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t} \geq \hat{\mu}_{i',t}\right)$$

$$= \sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t} - \mu_i - \frac{\Delta_i}{2} - \hat{\mu}_{i',t} + \mu_{i'} - \frac{\Delta_i}{2} \geq 0\right)$$

$$\leq \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_{i,t} - \mu_i - \frac{\Delta_i}{2} \geq 0\right)}_{\mathcal{R}_{i1}} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(\hat{\mu}_{i',t} - \mu_{i'} + \frac{\Delta_i}{2} \leq 0\right)}_{\mathcal{R}_{i2}},$$

where $\Delta_i = \max_{i'|a_i \in N(i)} \mu_{i'} - \mu_i$ is the expected loss incurred in choosing a_i instead of its best adjacent one $a_{i'}$.

$\mathcal{R}_{i1}$ can be upper bounded by a constant by relying on conditional probability definition and the Hoeffding inequality (Hoeffding 1963). Specifically, we rely on the fact that the leader is chosen at least $\left\lfloor \frac{L_{l(t),t}}{|N^+(l(t))|} \right\rfloor$ times. Upper bounding $\mathcal{R}_{i2}$ by a constant term requires the use of Proposition 1 in (Kaufmann, Korda, and Munos 2012), which limits the expected number of times the optimal arm is pulled less than t^b times by TS, where $b \in (0,1)$ is a constant, and

Table 1: Results concerning $R_\%(\mathfrak{U}, \text{OSUB})$ in the setting with $K = 17$ and $K = 129$ and a line graph.

	K	
	17	129
KLUCB	3.08 ± 0.05	6.51 ± 0.07
TS	1.34 ± 0.07	2.68 ± 0.05
UTS	$\mathbf{0.52 \pm 0.07}$	$\mathbf{0.76 \pm 0.15}$

the use of a technique already used on $\mathcal{R}_{i1}$. Summing up the regret over $i \neq i^*$ and considering the three obtained bounds concludes the proof. $\square$

Experimental Evaluation

In this section, we compare the empirical performance of the proposed algorithm UTS with the performance of a number of algorithms. We study the performance of the state–of–the–art algorithm OSUB (Combes and Proutiere 2014a) to evaluate the improvement due to the employment of Bayesian approaches w.r.t. frequentist approaches. Furthermore, we study the performance of TS (Thompson 1933) to evaluate the improvement in Bayesian approaches due to the exploitation of the problem structure. For completeness, we study also the performance of KLUCB (Garivier and Cappé 2011), being a frequentist algorithm that is optimal for Bernoulli distributions.

Figures of merit Given a policy $\mathfrak{U}$, we evaluate the average and 95%–confidence intervals of the following figures of merit:

- the pseudo–regret $R_T(\mathfrak{U})$ as defined in Equation 1; the lower $R_T(\mathfrak{U})$ the better the performance;

- the regret ratio $R_\%(\mathfrak{U}_1, \mathfrak{U}_2) = \frac{R_T(\mathfrak{U}_1)}{R_T(\mathfrak{U}_2)}$ showing the ratio between the total regret of policy $\mathfrak{U}_1$ after T rounds and the one obtained with $\mathfrak{U}_2$; the lower $R_\%(\mathfrak{U}_1, \mathfrak{U}_2)$ the larger the relative improvement of $\mathfrak{U}_1$ w.r.t. $\mathfrak{U}_2$.

Line graphs We initially consider the same experimental settings, composed of line graphs, that are studied in (Combes and Proutiere 2014a). They consider graphs with $K \in \{17, 129\}$ arms, where the arms are ordered on a line from the arm with smallest index to the arm with the largest index and with Bernoulli rewards whose averages have a triangular shape with the maximum on the arm in the middle of the line. More precisely, the minimum average is 0.1, associated with arms a_1 and a_{17} when $K = 17$ and with arms a_1 and a_{129} with $K = 129$, while the maximum average reward is $\mu^* = 0.9$, associated with arm a_9 when $K = 17$ and with arm a_{65} with $K = 129$. The averages decrease linearly from the maximum one to the minimum one.

For both the experiments, we average the regret over 100 independent trials of length $T = 10^5$. We report $R_t(\mathfrak{U})$ for each policy $\mathfrak{U}$ as t varies in Fig. 1(a), for $K = 17$, and in Fig. 1(b), for $K = 129$. The UTS algorithm outperforms all the other algorithms along the whole time horizon, providing a significant improvement in terms of regret w.r.t. the
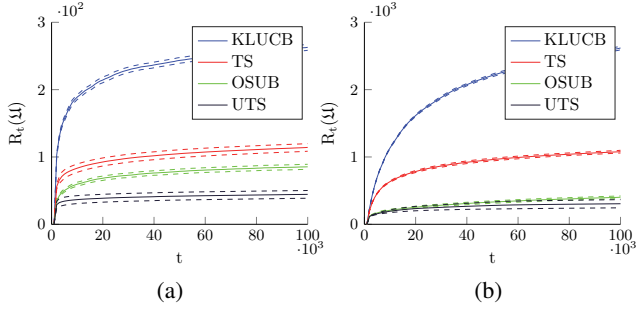
Figure 1: Results for the pseudo–regret $R_t(\mathfrak{U})$ in line graphs settings with $K = 17$ (a) and $K = 129$ (b) as defined in (Combes and Proutiere 2014a).

state–of–the–art algorithms. In order to have a more precise evaluation of the reduction of the regret w.r.t. OSUB algorithm, we report $R_\%(\mathfrak{U}, \text{OSUB})$ in Tab. 1. As also confirmed below by a more exhaustive series of experiments, in line graphs the relative improvement of performance due to UTS w.r.t. OSUB reduces as the number of arms increases, while the relative improvement of performance due to UTS w.r.t. TS increases as the number of arms increases.

Erdős-Rényi graphs To provide a thorough experimental evaluation of the considered algorithms in settings in which the space of arms has a graph structure, we generate graphs using the model proposed by Erdős and Rényi (1959), which allows us to simulate graph structures more complex than a simple line. An Erdős-Rényi graph is generated by connecting nodes randomly: each edge is included in the graph with probability p, independently from existing edges. We consider connected graphs with $K \in \{5, 10, 20, 50, 100, 1000\}$ and with probability $p \in \{1, \frac{1}{2}, \frac{\log(K)}{K}, \ell\}$, where $p = 1$ corresponds to have a fully connected graph and therefore the graph structure is useless, $p = \frac{1}{2}$ corresponds to have a number of edges that increases linearly in the number of nodes, $p = \frac{\log(K)}{K}$ corresponds to have a few edges w.r.t. the nodes, and we use $p = \ell$ to denote line graphs (these line graphs differ from those used for the experimental evaluation discussed above for the reward function, as discussed in what follows). We use different values of p in order to see how the performance of UTS changes w.r.t. the number of edges in the graph; we remark that such an analysis is unexplored in the literature so far. The optimal arm is chosen randomly among the existing arms and its reward is given by a Bernoulli distribution with expected value 0.9. The rewards of the suboptimal arms are given by Bernoulli distributions with expected value depending on their distance from the optimal one. More precisely, let d_i^* be the shortest path from the i–th arm to the optimal arm and let:

$$d_{\max}^* = \max_{i \in \{1,\dots,K\}} d_i^*$$

be the maximum shortest path of the graph. The expected reward of the i–th arm is:

$$\mu_i = 0.9 - d_i^* \frac{(0.9 - 0.1)}{d_{\max}^*},$$

Table 2: Results concerning $R_T(\mathfrak{U})$ ($T = 10^5$) in the setting with Erdős-Rényi graphs.

		p			
		1	1/2	$\log(K)/K$	ℓ
5	KLUCB	34 ± 0.4	50 ± 1.5	52 ± 3.7	56 ± 2.2
	TS	18 ± 0.2	23 ± 0.6	24 ± 1.3	25 ± 0.7
	OSUB	34 ± 0.3	32 ± 7.2	35 ± 5.8	31 ± 4.1
	UTS	$\mathbf{17 \pm 0.1}$	$\mathbf{15 \pm 2.4}$	$\mathbf{16 \pm 2.2}$	$\mathbf{14 \pm 1.3}$
10	KLUCB	77 ± 0.5	107 ± 5.5	127 ± 11.2	159 ± 7.0
	TS	40 ± 0.2	50 ± 2.0	56 ± 3.8	67 ± 2.5
	OSUB	77 ± 0.3	76 ± 8.1	57 ± 5.6	70 ± 8.1
	UTS	$\mathbf{39 \pm 0.2}$	$\mathbf{35 \pm 3.2}$	$\mathbf{27 \pm 2.1}$	$\mathbf{34 \pm 2.4}$
20	KLUCB	163 ± 0.7	217 ± 6.2	262 ± 16.2	386 ± 21.3
	TS	84 ± 0.5	102 ± 2.3	117 ± 5.7	157 ± 6.9
	OSUB	163 ± 0.8	148 ± 14.9	86 ± 14.6	124 ± 11.7
	UTS	$\mathbf{83 \pm 0.3}$	$\mathbf{70 \pm 6.0}$	$\mathbf{44 \pm 4.8}$	$\mathbf{65 \pm 8.8}$
50	KLUCB	420 ± 0.7	560 ± 15.0	686 ± 30.5	1132 ± 49.2
	TS	217 ± 0.5	262 ± 4.4	303 ± 10.0	454 ± 19.9
	OSUB	420 ± 1.0	382 ± 35.6	162 ± 13.9	240 ± 15.8
	UTS	$\mathbf{216 \pm 0.7}$	$\mathbf{182 \pm 14.2}$	$\mathbf{89 \pm 5.5}$	$\mathbf{156 \pm 30.1}$
100	KLUCB	846 ± 2.0	1134 ± 17.8	1313 ± 59.7	2327 ± 63.5
	TS	$\mathbf{436 \pm 1.1}$	528 ± 4.9	586 ± 18.4	973 ± 31.8
	OSUB	846 ± 2.7	786 ± 39.0	226 ± 27.1	369 ± 10.7
	UTS	$\mathbf{437 \pm 0.5}$	$\mathbf{372 \pm 15.2}$	$\mathbf{141 \pm 9.1}$	$\mathbf{290 \pm 42.3}$
1000	KLUCB	8505 ± 12.2	11247 ± 60.1	12024 ± 464.7	10640 ± 291.5
	TS	$\mathbf{4391 \pm 3.4}$	5262 ± 23.0	5478 ± 151.3	6554 ± 115.2
	OSUB	8493 ± 13.6	7761 ± 153.4	1151 ± 45.0	$\mathbf{1165 \pm 20.7}$
	UTS	$\mathbf{4388 \pm 5.2}$	$\mathbf{3718 \pm 62.9}$	$\mathbf{1000 \pm 14.2}$	1165 ± 41.8

i.e., the arm with $d_{\max}^*$ has a value equal to 0.1 and the expected rewards of the arms along the path from it to the optimal arm are evenly spaced between 0.1 and 0.9. We generate 10 different graphs for each combination of K and p and we run 100 independent trials of length $T = 10^5$ for each graph. We average the regret over the results of the 10 graphs.

In Tab. 2, we report $R_T(\mathfrak{U})$ for each combination of policy $\mathfrak{U}$, K, and p. It can be observed that the UTS algorithm outperforms all the other algorithms, providing in every case the smallest regret except for $K = 1000$ and $p = \ell$. Below we discuss how the relative performance of the algorithms vary as the values of the parameters K and p vary.

Consider the case with $p = 1$. The performance of UTS and TS are approximately equal and the same holds for the performance of OSUB and KLUCB. This is due to the fact that the neighborhood of each node is composed by all the arms, the graphs being fully connected, and therefore UTS and OSUB cannot take any advantage from the structure of the problem. We notice, however, that UTS and TS have not the same behavior and that UTS always performs slightly better than TS. It can be observed in Fig. 2 with $K = 5$ and $p = 1$ that the relative improvement is mainly at the beginning of the time horizon and that it goes to zero as K increases (the same holds for OSUB w.r.t. KLUCB). The reason behind this behavior is that UTS reduces the exploration performed by TS in the first rounds, forcing the algorithm to pull the leader—chosen as the arm maximizing the empirical mean—for a larger number of rounds.
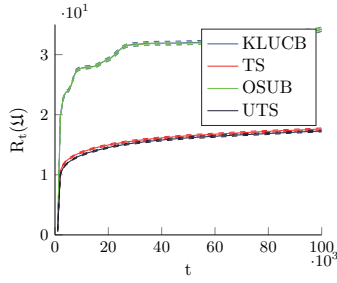
Figure 2: Results for the pseudo–regret $R_t(\mathfrak{U})$ in the setting with $K = 5$ and $p = 1$.

(a)   (b)
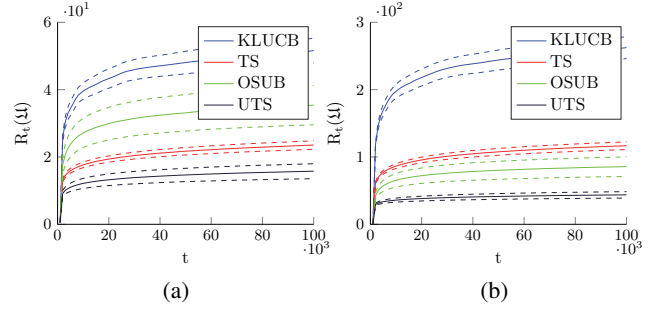
Figure 3: Results for the pseudo–regret $R_t(\mathfrak{U})$ in the setting with $K = 5$ (a) and $K = 20$ (b) and $p = \frac{\log(K)}{K}$.

Consider the case with $p = \frac{1}{2}$. In the considered experimental setting, the relative performance of the algorithms does not depend on K. The ordering, from the best to the worst, over the performance of the algorithms is: UTS, TS, OSUB, and finally KLUCB. Surprisingly, even the dependency of the following ratios on K is negligible: $R_\%(\text{UTS}, \text{TS}) = 0.68 \pm 0.03$, $R_\%(\text{UTS}, \text{OSUB}) = 0.47 \pm 0.01$, and $R_\%(\text{OSUB}, \text{KLUCB}) = 0.68 \pm 0.03$. This shows that the relative improvement due to UTS is constant w.r.t. TS and OSUB as K varies. These results raise the question whether the relative performance of OSUB and TS would be the same, except for the numerical values, for every p constant w.r.t. K. To answer to this question, we consider the case in which $p = 0.1$, corresponding to the case in which the number of edges is linear in K, but it is smaller than the case with $p = \frac{1}{2}$. The results in terms of $R_T(\mathfrak{U})$, reported in Table 3 show that OSUB outperforms TS for $K \geq 10$, suggesting that, when p is constant in K, OSUB may or may not outperform TS depending on the specific pair (p, K).

Consider the case with $p = \frac{\log(K)}{K}$. The ordering over the performance of the algorithms changes as K varies. More precisely, while UTS keeps to be the best algorithm for every K and KLUCB the worst algorithm for every K, the ordering between TS and OSUB changes. When $K \leq 10$ TS performs better than OSUB, instead when $K \geq 20$ OSUB outperforms TS, see Fig. 3. This is due to the fact that, with a small number of arms, exploiting the graph structure is not sufficient for a frequentist algorithm to outperform the performance of TS, while with many arms exploiting the graph structure even with a frequentist algorithm is much better than employing a general-purpose Bayesian algorithm. The ratio $R_\%(\text{UTS}, \text{TS})$ monotonically decreases as K increases, from 0.66 when $K = 5$ to 0.19 when $K = 1000$, suggesting that exploiting the graph structure provides ad-

Table 3: Results concerning $R_T(\mathfrak{U})$ ($T = 10^5$) in the setting with Erdős-Rényi graphs and $p = 0.1$.

			K			
	5	10	20	50	100	1000
TS	**25**	66	162	278	519	4564
OSUB	29	**64**	**126**	**144**	**266**	**2358**

vantages as K increases. Instead, the ratio $R_\%(\text{UTS}, \text{OSUB})$ monotonically increases as K increases, from 0.45 when $K = 5$ to 0.94 when $K = 1000$, suggesting that the improvement provided by employing Bayesian approaches reduces as K increases as observed above in line graphs.

Consider the case with $p = \ell$. As in the case discussed above, OSUB is outperformed by TS for a small number of arms ($K \leq 10$), while it outperforms TS for many arms ($K \geq 20$). The reason is the same above. Similarly, the ratio $R_\%(\text{UTS}, \text{TS})$ monotonically decreases as K increases, from 0.58 when $K = 5$ to 0.18 when $K = 1000$, and the ratio $R_\%(\text{UTS}, \text{OSUB})$ monotonically increases as K increases, from 0.45 when $K = 5$ to 1.00 when $K = 1000$. This confirms that the performance of UTS and the one of OSUB asymptotically match as K increases when $p = \ell$ (as well as $p = \frac{\log(K)}{K}$). In order to investigate the reasons behind such a behavior, we produce an additional experiment with the line graphs of Combes and Proutiere (2014a) except that the maximum expected reward is set to 0.108 when $K = 17$ and 0.165 when $K = 129$ (thus, given any edge with terminals i and $i + 1$, we have $|\mu_i - \mu_{i+1}| = 0.001$). What we observe (details of these experiments and those described below are in (Paladino et al. 2016)) is that, on average, OSUB outperforms UTS at $T = 10^5$ suggesting that, when it is necessary to repeatedly distinguish between three arms that have very similar expected rewards, frequentist methods may outperform the Bayesian ones. This is no longer true when T is much larger, e.g., $T = 10^7$, where UTS outperforms OSUB (interestingly, differently from what happens in the other topologies, in line graphs with very small $|\mu_i - \mu_{i+1}|$, the average $R_T(\text{UTS})$ and $R_T(\text{OSUB})$ cross a number of times during the time horizon). Futhermore, we evaluate how the relative performance of OSUB w.r.t. UTS varies for $|\mu_i - \mu_{i+1}| \in \{0.001, 0.002, 0.005\}$, observing it improves as $|\mu_i - \mu_{i+1}|$ decreases. Finally, we evaluate whether this behavior emerges also in Erdős-Rényi graphs in which $p = \frac{c}{K}$ where c is a constant (we use $p = \frac{5}{K}, \frac{10}{K}$) and we observe that UTS outperforms OSUB, suggesting that line graphs with very small $|\mu_i - \mu_{i+1}|$ are pathological instances for UTS.

Conclusions and Future Work

In this paper, we focus on the Unimodal Multi–Armed Bandit problem with graph structure in which each arm corresponds to a node of a graph and each edge is associated with a relationship in terms of expected reward between its arms. We propose, to the best of our knowledge, the first Bayesian algorithm for the UMAB setting, called UTS, which is based on the well–known Thompson Sampling algorithm. We derive a tight upper bound for UTS that asymptotically matches the lower bound for the UMAB setting, providing a non-trivial derivation of the bound. Furthermore, we present a thorough experimental analysis showing that our algorithm outperforms the state–of–the–art methods.

In future, we will evaluate the performance of the algorithms considered in this paper with other classes of graphs, e.g., Barabási–Albert and lattices. Future development of this work may consider an analysis of the proposed algorithm in the case of time–varying environments, i.e., the expected reward of each arm varies over time, assuming that the unimodal structure is preserved. Another interesting study may consider the case of a continuous decision space.

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 (NIPS)*, 1610–1618.

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

Caron, S., and Bhagat, S. 2013. Mixing bandits: A recipe for improved cold–start recommendations in a social network. In *Proceedings of the Workshop on Social Network Mining and Analysis (SNA-KDD)*.

Chapelle, O., and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2249–2257.

Combes, R., and Proutiere, A. 2014a. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 521–529.

Combes, R., and Proutiere, A. 2014b. Unimodal bandits without smoothness. *arXiv preprint arXiv:1406.7447*.

Crandall, D.; Cosley, D.; Huttenlocher, D.; Kleinberg, J.; and Suri, S. 2008. Feedback effects between similarity and social influence in online communities. In *Proceedings of the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, 160–168.

Edelman, B., and Ostrovsky, M. 2007. Strategic bidder behavior in sponsored search auctions. *Decision Support Systems* 43(1):192–198.

Erdős, P., and Rényi, A. 1959. On random graphs I. *Publicationes Mathematicae (Debrecen)* 6:290–297.

Garivier, A., and Cappé, O. 2011. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the Conference on Learning Theory (COLT)*, 359–376.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30.

Jia, Y. Y., and Mannor, S. 2011. Unimodal bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 41–48.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: An asymptotically optimal finite–time analysis. In *Proceedings of the Algorithmic Learing Theory (ALT)*, 199–213.

Kleinberg, R.; Slivkins, A.; and Upfal, E. 2008. Multi–armed bandits in metric spaces. In *Proceedings of the Symposium on Theory Of Computing (STOC)*, 681–690.

Lai, T. L., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4–22.

Mannor, S., and Shamir, O. 2011. From bandits to experts: On the value of side–observations. In *Advances in Neural Information Processing Systems (NIPS)*, 684–692.

Mas-Collel, A.; Whinston, M. D.; and Green, J. R. 1995. *Micreconomic Theory*. Oxford University Press.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

Paladino, S.; Trovò, F.; Restelli, M.; and Gatti, N. 2016. Unimodal thompson sampling for graph-structured arms. *arXiv preprint arXiv:1611.05724*.

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

Trovò, F.; Paladino, S.; Restelli, M.; and Gatti, N. 2015. Multi–armed bandit for pricing. In *Proceedings of the European Workshop on Reinforcement Learning (EWRL)*.

Valko, M.; Munos, R.; Kveton, B.; and Kocak, T. 2014. Spectral bandits for smooth graph functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 46–54.

Xia, Y.; Li, H.; Qin, T.; Yu, N.; and Liu, T.-Y. 2015. Thompson sampling for budgeted multi–armed bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.