Problem-dependent regret bounds for online learning with feedback graphs

Bingshan Hu
University of Victoria
bingshanhu@uvic.ca

Nishant A. Mehta
University of Victoria
nmehta@uvic.ca

Jianping Pan
University of Victoria
pan@uvic.ca

Abstract

This paper addresses the stochastic multi-armed bandit problem with an undirected feedback graph. We devise a UCB-based algorithm, UCB-NE, to provide a problem-dependent regret bound that depends on a clique covering. Our algorithm obtains regret which provably scales linearly with the clique covering number. Additionally, we provide problem-dependent regret bounds for a Thompson Sampling-based algorithm, TS-N, where again the bounds are linear in the clique covering number. Finally, we present experimental results to see how UCB-NE, TS-N, and a few related algorithms perform practically.

1 INTRODUCTION

In the stochastic multi-armed bandit problem, a learning agent sequentially decides to pull an arm in each of T rounds in order to maximize its cumulative reward. Each arm emits rewards that are i.i.d. according to a fixed but unknown distribution specific to that arm, and in a given round the agent only observes the reward of the arm it pulled in that round. Naturally, the limited feedback aspect of this game creates a tension between exploration — acquiring information to better estimate the mean reward of an arm — and exploitation — pulling the arm that empirically looks the best so far.

The standard notion of regret in this setting is the *pseudo-regret* (hereafter referred to simply as "regret"), which measures the difference between the agent's expected cumulative reward and the expected cumulative reward of the arm with the highest mean reward. For simplicity of this initial exposition, we consider the case of K arms where one arm has a mean reward of μ and all other arms have a mean reward of $\mu - \Delta$ for some $\Delta > 0$. While it is known that a *problem-independent* regret bound of order $O(\sqrt{TK})$ is possible (Audibert and Bubeck, 2009), more refined, *problem-dependent* regret bounds that take into account distributional information also exist (Auer et al., 2002); (Garivier and Cappé, 2011); (Agrawal and Goyal, 2017). These bounds grow only logarithmically in T and take the form $O\left(\frac{K \log T}{\Delta}\right)$ or $O\left(\frac{\Delta K \log T}{d(\mu-\Delta,\mu)}\right)$.[1]

A number of recent works have considered the online learning with feedback graphs setting. This setting can be viewed as an extension of the multi-armed bandit setting where additional *side observations* are available when pulling an arm, as specified by a feedback graph G. When pulling an arm, one receives observations from that arm and all of its neighbors in the feedback graph. A concrete application is an online advertising/promotion system in a social network. A merchant may give a special discount to selected users to promote their items. The merchant can then observe whether the selected users like the advertised items or not. Meanwhile, the selected users are likely to recommend the advertised items to their friends via social networks. Therefore, the merchant may also get additional observations from the friends of the selected users.

Whereas the regret bounds in the standard multi-armed bandit problem are inherently linear in the number of arms, in the feedback graph setting it is possible to break this dependence, replacing K by certain graph-theoretic properties. For instance, in the case of undirected feedback graphs, Caron et al. (2012) developed an index-style algorithm, UCB-N, that replaces K by the clique covering number in the *leading term* of the regret bound (the term depending on T); however, their regret bound still has a constant term (the term not depending on T) that is linear in K. For directed feedback graphs, Cohen et al. (2016) developed an arm elimination-style al-

[1] Here, $d(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the KL divergence of a Bernoulli distribution with success probability p from a Bernoulli distribution with success probability q.

gorithm which, remarkably, replaces K by $\alpha(G)\log K$ for the leading term (and also the constant term) in a problem-dependent bound; here, $\alpha(G)$ is the independence number of feedback graph G (where directed edges are counted as undirected edges). However, as we explain in Section 6, the additional $\log K$ factor is sometimes unnecessary and the algorithm does not perform well in practice.

Thompson Sampling-based algorithms typically perform the best, and this is also the case for the online learning with feedback graphs problem. Indeed, an algorithm called TS-N (due to Liu et al. (2018a)) exhibits excellent empirical performance in the case of feedback graphs. However, whereas there are problem-dependent regret bounds for Thompson Sampling in the case of standard bandit feedback (Agrawal and Goyal, 2017; Kaufmann et al., 2012), no problem-dependent regret bounds have been shown for TS-N in the case of feedback graphs. Existing bounds, due to Liu et al. (2018a); Liu et al. (2018b), do depend on the clique covering number or $\alpha(G)$ but are only on the Bayesian regret.

Our core contributions, all for undirected feedback graphs, are as follows:

1. We devise a new upper confidence bound-based algorithm, UCB-NE, for the stochastic $\mathcal{N}$-armed bandit problem with an undirected feedback graph. We prove a problem-dependent regret bound for this algorithm which, for any clique covering, is linear in the size of the clique covering and logarithmic in the size of the cliques, both with respect to the leading and constant terms; the precise result can be found in Theorem 1. UCB-NE does not depend on a clique covering as input, instead only using degree information to construct upper confidence bounds.[2]
2. For the TS-N algorithm of Liu et al. (2018a), we give two problem-dependent regret bounds that, similar to UCB-NE, depend only linearly on the size of a clique covering and logarithmically on the size of each clique. These are the first problem-dependent regret bounds for any Thompson Sampling algorithm that improve with properties of feedback graphs. Both bounds involve a free parameter ϵ which allows a tradeoff between the leading term and the constant term, similar to previous bounds by Agrawal and Goyal (2017); Kaufmann et al. (2012). The first bound, Theorem 2, tends to optimize the leading term and hides problem-dependent constants, again similar to previous regret bounds by Agrawal and Goyal (2017); Kaufmann et al. (2012) in the standard bandit setting. This makes it difficult to assess the tradeoff between the leading and constant terms, as is needed to tune ϵ. We therefore give a second regret bound, Theorem 3, that gives an explicit form for the constant term, thereby enabling a user to suitably tune ϵ. We note that our bounds also hold for the special case of standard bandit feedback, in which case our bounds represent the first fully explicit bounds for Thompson Sampling; previous bounds did not explicitly control the constant term, which in some cases may actually be larger than the leading term.
3. We present experimental results to practically study how the regret grows for UCB-NE, TS-N, UCB-N, the arm elimination-style algorithm of Cohen et al. (2016), and another algorithm called TS-MaxN (Tossou et al., 2017).

This paper are organized as follows. Section 2 formally presents the stochastic multi-armed bandit problem with undirected feedback graphs. Section 3 discusses related work. Section 4 presents our algorithm, UCB-NE, along with a problem-dependent regret bound, and Section 5 presents problem-dependent regret bounds for TS-N. Experimental results are provided in Section 6. Finally, Section 7 concludes the paper. All proofs that do not appear in this paper are in the supplementary material.

2 STOCHASTIC MULTI-ARMED BANDITS WITH UNDIRECTED FEEDBACK GRAPHS

We consider a stochastic $\mathcal{N}$-armed bandit problem with an undirected feedback graph. The learner plays this game for T rounds. At the beginning of round t, the environment generates random rewards in $[0,1]$ for all arms independently[3] from fixed but unknown distributions. Graph $\mathcal{G} := (\mathcal{N}, \mathcal{E})$ denotes an undirected feedback graph that captures all the feedback relationships over arm set $\mathcal{N}$. An edge $i \leftrightarrow j$ in $\mathcal{E}$ means that the learner can get a side observation of arm j when pulling arm i, and vice versa. Note that pulling arm i always lets the learner observe the reward of arm i itself, i.e., $\mathcal{E}$ includes all self-loops. We assume that graph $\mathcal{G}$ does not vary over time. For each $i \in \mathcal{N}$, let set $\mathcal{N}_i$ collect arm i and all its neighbors in $\mathcal{G}$. In each round t, the learner pulls an arm $I_t \in \mathcal{N}$ and then observes the reward of each arm in $\mathcal{N}_{I_t}$. The goal of the learner is to pull arms

[2] We note in passing that Caron et al. (2012) introduced an algorithm called UCB-MaxN that also attempted to improve the constant term. However, as we explain in Section 3, the regret analysis of this algorithm may not always realize such an improvement.

[3] Actually, for UCB-NE, it is not required that the random rewards of all arms be generated independently, i.e., they can be generated from a joint distribution.

sequentially to maximize its expected cumulative reward over T rounds.

Let μ_i denote the true mean of arm i's reward. We assume that the first arm is the unique best arm, i.e., $\mu_1 > \mu_i, \forall i \neq 1$. It is possible to modify the analysis if there are multiple best arms. Let $\Delta_i := \mu_1 - \mu_i$ for all $i \in \mathcal{N}$. Note that $\Delta_1 = 0$. To measure the quality of our learning algorithms, we use the (pseudo-)regret $\mathcal{R}(T)$, which is defined as

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \mu_1 - \mu_{I_t}\right] \quad . \tag{1}$$

In this work, an arbitrary clique covering $\mathcal{C}$ is used to derive our regret bound. $\mathcal{C}$ is a set of cliques such that $\bigcup_{C \in \mathcal{C}} C = \mathcal{N}$ where $C \in \mathcal{C}$ is a clique. A clique in $\mathcal{G}$ is a subset of $\mathcal{N}$ such that all nodes are neighbors with each other. Then the regret $\mathcal{R}(T)$ can be further expressed as

$$\begin{aligned} \mathcal{R}(T) &= \sum_{i \in \mathcal{N}} \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}\{I_t = i\}\right] \cdot \Delta_i \\ &\leq \sum_{C \in \mathcal{C}} \mathbb{E}[\underbrace{\sum_{t=1}^{T} \sum_{i \in C} \mathbf{1}\{I_t = i\} \cdot \Delta_i}_{R_C(T)}] \quad , \end{aligned} \tag{2}$$

where $R_C(T) := \sum_{t=1}^{T} \sum_{i \in C} \mathbf{1}\{I_t = i\} \cdot \Delta_i$ denotes the *intra-clique regret*, i.e., the regret of pulling any sub-optimal arm in clique C. Note that we only need to analyze the cliques that are not equal to $\{1\}$. For any $C \neq \{1\}$, let $\mu_C^{\max} := \max_{i \in C \setminus 1} \mu_i$, $\Delta_C^{\max} := \max_{i \in C \setminus 1} \Delta_i$, and $\Delta_C^{\min} := \min_{i \in C \setminus 1} \Delta_i$.

3 RELATED WORK

To fully exploit the feedback structure, previous works have used either a *clique covering* $\mathcal{C}$ over all the nodes in $\mathcal{G}$ or the *independence number* $\alpha(\mathcal{G})$ to derive regret bounds. The independence number of a graph is defined as the cardinality of the maximum independent set. The first regret bound of a stochastic $\mathcal{N}$-armed bandit problem with an undirected feedback graph was provided by Caron et al. (2012). The authors devised two UCB-based algorithms: UCB-N and UCB-MaxN. In UCB-N, just like in previous work (Auer et al., 2002), the learner pulls the arm with the highest upper confidence bound in each round while in UCB-MaxN, the learner first locates the arm with the highest upper confidence bound but actually pulls the arm with the highest empirical mean among the neighbors of the arm with the highest confidence bound. Caron et al. (2012) exploited properties of clique coverings to derive problem-dependent regret bounds, i.e., pulling any arm within a clique C allows the learner to obtain an observation of all the arms within C. The leading term for UCB-N is $O\left(\sum_{C \in \mathcal{C}} \frac{\Delta_C^{\max} \ln(T)}{(\Delta_C^{\min})^2}\right)$ while the constant term is $O\left(\sum_{C \in \mathcal{C}} |C|\right) = O(|\mathcal{N}|)$. Note that the learner does not need to know the feedback graph in advance for UCB-N. Using the algorithm UCB-MaxN, it seems to be possible to improve the problem-dependent constant term to $O(|\mathcal{C}|)$ asymptotically under an assumption, i.e., that the best sub-optimal arm within each clique is unique and the gap δ between this best sub-optimal arm and the second best sub-optimal arm (within the same clique) is not arbitrarily small. However, as we explain in the supplementary material, there appears to be a subtle issue with proof of the regret bound for UCB-MaxN. Our algorithm UCB-NE improves the constant term in their regret bounds by avoiding dependence on δ and provides a regret bound that holds for any clique covering. Note that in UCB-NE, the learner only needs to know the feedback graph instead of the knowledge of clique coverings.

Cohen et al. (2016) devised an elimination-based algorithm[4] to exploit a directed feedback graph. Note that an undirected feedback graph can treated as a special directed feedback graph. They gave a problem-dependent regret bound that scales with the independence number $\alpha(\mathcal{G})$. Their regret bound is $O\left(\sum_{v \in V'} \frac{\ln(T)}{\Delta_v}\right)$, where V' is the set of $O\left(\alpha(\mathcal{G}) \ln(|\mathcal{N}|)\right)$ arms with the smallest gaps. Although the independence number $\alpha(\mathcal{G})$ is always no greater than the clique covering number, due to the multiplicative interaction with $\ln(|\mathcal{N}|)$, their regret bound may not be always better than one which scales with the clique covering number. Also, although this elimination-based algorithm has good theoretical guarantees, it does not work well practically as shown by Liu et al. (2018b) and further confirmed by our experiments in Section 6. Additionally, the learner needs to know the time horizon T in advance. Otherwise, the learner needs to resort a "doubling trick" shown in (Auer and Ortner, 2010).

Liu et al. (2018a) and Liu et al. (2018b) devised a Thompson Sampling-based algorithm, TS-N, to exploit an undirected feedback graph. They gave regret bounds scaling with clique covering number ($O\left(\sqrt{|\mathcal{C}|T\ln(|\mathcal{N}|)}\right)$) and independence number ($O\left(\sqrt{\alpha(\mathcal{G})T\ln(|\mathcal{N}|)}\right)$). However, they used *Bayesian*

[4] Their algorithm admits regret bounds even if $\mathcal{G}$ varies over time.

regret instead of the *pseudo-regret* to measure the quality of the algorithm, and their regret bounds are problem-independent. We derive problem-dependent regret bounds for TS-N that depend on a clique covering.

4 UCB-NE

In this section, we introduce UCB-NE and provide a problem-dependent regret bound.

4.1 ALGORITHM

Algorithm 1 presents the UCB-NE ('E' stands for extra exploration). Let $O_i(t)$ be the number of observations of arm i until the end of round t and $\hat{\mu}_{i,O_i(t)}$ be the empirical mean of arm i until the end of round t. Let $\bar{\mu}_i(t) := \hat{\mu}_{i,O_i(t-1)} + \sqrt{\frac{2\ln\left(|\mathcal{N}_i|^{\frac{1}{4}}\cdot t\right)}{O_i(t-1)}}$ be the upper confidence bound of arm i at round t. Note that the second term in the upper confidence bound is enlarged as compared to the standard value of $\sqrt{\frac{2\ln(t)}{O_i(t-1)}}$. This enlargement makes the algorithm explore more and, in the regret analysis, enables us to get rid of the factor that makes the constant term scale linearly with the clique size. More specifically, the extra exploration allows the constant term from each clique to be divided by something no smaller than the clique size. In every round t, the learner pulls the arm with the highest upper confidence bound, i.e., $I_t \leftarrow \arg\max_{i\in\mathcal{N}} \bar{\mu}_i(t)$. Then at the end of round t, all the neighboring arms of the pulled arm including itself, i.e., all $i \in \mathcal{N}_{I_t}$ will be observed and the corresponding $O_i(t)$ and $\hat{\mu}_{i,O_i(t)}$ will be updated. Let $X_i(t) \in [0,1]$ be the random reward for arm i at round t. Although UCB-NE does not depend on a clique covering as input, the algorithm needs the knowledge of graph structure as the degree information for each arm is used to construct the upper confidence bound.

Algorithm 1 UCB-NE

```
Set O_i ← 0, μ̂_{i,O_i} ← 0, ∀i ∈ 𝒩
for t = 1 : T do
    Set μ̄_i(t) = μ̂_{i,O_i} + sqrt(2 ln(|𝒩_i|^{1/4} · t) / O_i), ∀i ∈ 𝒩
    Pull arm I_t ← arg max_{i∈𝒩} μ̄_i(t)
    for i ∈ 𝒩_{I_t} do
        Set O_i ← O_i + 1; Observe X_i(t)
        Set μ̂_{i,O_i} ← (μ̂_{i,O_i} · (O_i − 1) + X_i(t)) / O_i
    end for
end for
```

4.2 REGRET ANALYSIS

Let $N_C := \max_{i\in C}\left\{|\mathcal{N}_i|^{\frac{1}{4}}\right\}$.

Theorem 1. *The regret $\mathcal{R}(T)$ of UCB-NE is at most*

$$\inf_{\mathcal{C}}\left\{\sum_{C\in\mathcal{C},C\neq\{1\}}\mathbb{E}\left[R_C(T)\right]\right\} \leq \inf_{\mathcal{C}}\sum_{\substack{C\in\mathcal{C}\\C\neq\{1\}}}\left(\frac{8\Delta_C^{\max}\ln(N_C\cdot T)}{(\Delta_C^{\min})^2}+\left(1+\frac{\pi^2}{3}\right)\Delta_C^{\max}\right).$$

Several remarks are in order. First, we discuss the case where no side observations are available, i.e., a standard stochastic multi-armed bandit problem. We can take a trivial clique covering $\mathcal{C} = \{\{i\}, \forall i \in \mathcal{N}\}$ to recover the regret bound of this classic setting. From $\mathcal{C} = \{\{i\}, \forall i \in \mathcal{N}\}$ we have $\Delta_C^{\min} = \Delta_C^{\max} = \Delta_i$ and $N_C = |\mathcal{N}_i| = 1$ for all $C \neq \{1\}$. Then our regret bound is the same as the one for UCB1 in (Auer et al., 2002). Next, we discuss the difference between UCB-N in (Caron et al., 2012) and UCB-NE if side observations are available. Given the same feedback graph, the leading term of UCB-NE and UCB-N is the same. With respect to the constant term, UCB-N is $O(|C|)$ while UCB-NE improves to $O\left(\frac{\ln(N_C)}{\Delta}\right)$ when the clique size is large. However, when taking the trivial clique covering $\mathcal{C} = \{\{i\}, \forall i \in \mathcal{N}\}$, UCB-N boils down to the same regret bound as UCB1 while UCB-NE needs to pay an additional price of $\frac{2\ln(|\mathcal{N}_i|)}{\Delta_i}$ for each sub-optimal arm i.

Similar to the analysis of UCB-N, to obtain our regret bound, we also bound the total number of times that the learner pulls any sub-optimal arm within each clique. For each clique C, the regret can be decomposed into two regimes, the under-sampled regime and the sufficiently sampled regime. Specifically, we say that a clique C is in the under-sampled regime if the total number of times that the learner has pulled any arm in C is less than a threshold $L_C := \left\lceil \frac{8\ln(N_C\cdot T)}{(\Delta_C^{\min})^2}\right\rceil$, where we recall that $N_C = \max_{i\in C}\left\{|\mathcal{N}_i|^{\frac{1}{4}}\right\}$. For the rounds when clique C is in the under-sampled regime, the total regret is at most $L_C \cdot \Delta_C^{\max}$ while for the rounds when clique C is in the sufficiently sampled regime, we use a concentration inequality to bound the total regret contribution from this regime by a constant not depending on the clique size. Note that the term N_C appearing in L_C typically would not be present in a standard UCB analysis or the analysis of UCB-N. We use this term because, as explained earlier, UCB-NE's upper confidence bounds have an extra exploration term $|\mathcal{N}_i|^{1/4}$ that is upper bounded by N_C.

5 TS-N

In this section, we introduce the algorithm TS-N of Liu et al. (2018a) and provide problem-dependent regret bounds. Unlike the previous section, we now restrict to the case of Bernoulli rewards.

5.1 ALGORITHM

Algorithm 2 presents TS-N in detail. Unlike the previous section, $O_i(t)$ denotes the number of times that arm i has been observed until the end of round $t-1$. $Q_i(t)$ denotes the number of times that the learner gets reward equal to 1 among the $O_i(t)$ observations, i.e., the number of times that the Bernoulli trial succeeds until the end of round $t-1$. For each arm $i \in \mathcal{N}$, let $\theta_i(t)$ denote a random value independently generated from posterior distribution $Beta(Q_i(t)+1, O_i(t)-Q_i(t)+1)$ at round t, where $Beta(\alpha,\beta)$ denotes a beta distribution with parameter α,β. At the end of round t, all the neighboring arms of the pulled arm including itself will be observed and the parameters of the corresponding beta distributions will be updated. Let $X_i(t) \in \{0,1\}$ be the random reward for arm i at round t. Note that TS-N does not depend on a clique covering as input nor does the learner need knowledge of the feedback graph.

Algorithm 2 TS-N (Liu et al., 2018a)

Set $O_i \leftarrow 0, Q_i \leftarrow 0, \forall i \in \mathcal{N}$
for $t = 1 : T$ **do**
  Sample $\theta_i(t)$ from $Beta(Q_i+1, O_i-Q_i+1), \forall i \in \mathcal{N}$
  Pull arm $I_t \leftarrow \arg\max_{i\in\mathcal{N}} \theta_i(t)$
  for $i \in \mathcal{N}_{I_t}$ **do**
    Set $O_i \leftarrow O_i + 1$; Observe $X_i(t)$
    Set $Q_i \leftarrow Q_i + X_i(t)$
  end for
end for

5.2 REGRET ANALYSIS

Let $N_C := \max_{i\in C} |\mathcal{N}_i|$ and, for $a,b \in [0,1]$, define $d(a,b) := a\ln(\frac{a}{b}) + (1-a)\ln(\frac{1-a}{1-b})$ to be the Kullback-Leibler (KL) divergence of a Bernoulli distribution with success probability a from a Bernoulli distribution with success probability b.

Theorem 2. *The regret $\mathcal{R}(T)$ of TS-N is at most*

$$\inf_{\mathcal{C}} \left\{ \sum_{C\in\mathcal{C}, C\neq\{1\}} \mathbb{E}\left[R_C(T)\right] \right\}$$
$$\leq \inf_{\mathcal{C}} \left\{ \sum_{C\in\mathcal{C}, C\neq\{1\}} \frac{\Delta_C^{\max}(1+\epsilon_C)\ln(T)}{d(\mu_C^{\max},\mu_1)} + O\left(\frac{\ln(N_C)+1}{(\epsilon_C)^2}\right) \right\},$$

where ϵ_C can be any value in $\left(0, \min\left\{\frac{d(\mu_C^{\max},\mu_1)}{d(m_C,\mu_1)} - 1, 1\right\}\right)$ and $m_C \in (\mu_C^{\max}, \mu_1)$ is a unique clique-specific problem-dependent constant. The Big-O notation in the constant term hides problem-dependent constants.

Let us make a few remarks about this theorem. First, we discuss the case where there is no feedback graph, i.e., a standard stochastic multi-armed bandit problem. We compare our regret bound with Theorem 1 in (Agrawal and Goyal, 2017). We can take a trivial clique covering $\mathcal{C} = \{\{i\}, \forall i \in \mathcal{N}\}$ to represent the case where there is no feedback graph. Then we have $\mu_C^{\max} = \mu_i$ and $N_C = |\mathcal{N}_i| = 1$ for all $C \neq \{1\}$, and our regret bound boils down to Theorem 1 in (Agrawal and Goyal, 2017) with the only difference of the choice of ϵ_C. In (Agrawal and Goyal, 2017), they have freedom to choose any $\epsilon_C \in (0,1)$ while we may not have that freedom. During the proof of our Theorem 2, more precisely, in Lemma 1, we present the range of ϵ_C in our regret bound. We use ϵ_C to control the problem-dependent constant term to make it scale logarithmically with the clique size. It is important to note that ϵ_C does not depend on the number of arms within clique C. Instead, ϵ_C only depends on μ_1 and $\mu_C^{\max}$ (the mean reward of the best sub-optimal arm in clique C). Next, we discuss the difference between TS-N and UCB-NE. With respect to the leading term, TS-N is better than UCB-NE while for the constant term, TS-N may be worse than UCB-NE. However, the constant terms for TS-N and UCB-NE both scale logarithmically with the clique size instead of linearly.

With respect to the leading term, Theorem 2 provides a good theoretical guarantee while for the constant term, it hides many problem-dependent constants. The hidden terms can be found in the proof. Also, there is a limitation of the choice of ϵ_C for each clique C. Therefore, we provide another theorem for which any $\epsilon \in (0,1)$ is allowed and the constant terms can be expressed explicitly. The exposure of the previously-hidden constant term enables us to achieve a good tradeoff between the leading term and constant term by tuning ϵ properly.

Theorem 3. *For any $\epsilon \in (0,1)$, the regret $\mathcal{R}(T)$ of TS-N*

is at most

$$\inf_{\mathcal{C}} \left\{ \sum_{C\in\mathcal{C}, C\neq\{1\}} \mathbb{E}\left[R_C(T)\right] \right\}$$
$$\leq \inf_{\mathcal{C}} \left\{ \begin{array}{l} \sum\limits_{C\in\mathcal{C}, C\neq\{1\}} \frac{(3+\lambda_C)^2\Delta_C^{\max}\ln(T)}{2(1-\epsilon)^2(\Delta_C^{\min})^2} \\ + \frac{(3+\lambda_C)^2\Delta_C^{\max}(\ln(N_C)+1)}{2(1-\epsilon)^2(\Delta_C^{\min})^2} + O\left(\frac{\Delta_C^{\max}}{\epsilon^4(\Delta_C^{\min})^4}\right) \end{array} \right\},$$

where $\lambda_C := \log\left(\frac{\mu_1-\epsilon\Delta_C^{\min}}{\mu_C^{\max}}\right)$.

In the supplementary material, we show that instead of paying $O\left(\frac{\Delta_C^{\max}}{\epsilon^4(\Delta_C^{\min})^4}\right)$, an alternative is to pay $O\left(\frac{\Delta_C^{\max}\ln(T\cdot\epsilon\Delta_C^{\min})}{\epsilon^2(\Delta_C^{\min})^2}\right)$.

Notation and definitions: Before presenting the analysis, we first introduce some important notation and definitions. Let $T_i(t) := \sum_{s=1}^{t-1} \mathbf{1}\{I_s = i\}$ be the total number of times that arm i has been pulled until the end of round $t-1$ and $T_C(t)$ be the total number of times that the learner pulls any arm in clique C until the end of round $t-1$, i.e., $T_C(t) := \sum_{s=1}^{t-1} \mathbf{1}\{\exists j \in C \; s.t. \; I_s = j\}$. Different from UCB-NE, in TS-N, $\hat{\mu}_i(t) = \frac{Q_i(t)}{O_i(t)+1}$ is defined as the empirical mean of arm i at round t. $\mathcal{F}_t$ collects all the history information until the end of round t sequentially, which is $\mathcal{F}_t = \{I_s, X_i(s), \forall i \in \mathcal{N}_{I_s}, s = 1, 2, \cdots, t\}$. Define $\mathcal{F}_0 = \{\}$, and note that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_{T-1}$ always holds. For each arm i, $O_i(t)$, $Q_i(t)$, and $\hat{\mu}_i(t)$ are determined by $\mathcal{F}_{t-1}$. Also, the distribution that generates $\theta_i(t)$ is determined by $\mathcal{F}_{t-1}$.

To prove Theorem 2, we first do a regret decomposition. L_C is a clique-specific positive integer that will be chosen later, and tuning L_C needs some novel techniques.

$$\begin{aligned} R_C(T) &= \sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i\} \cdot \Delta_i \\ &= \underbrace{\sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i, T_C(t) < L_C\} \cdot \Delta_i}_{\leq L_C \cdot \Delta_C^{\max}} \\ &\quad + \underbrace{\sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i, T_C(t) \geq L_C\} \cdot \Delta_i}_{\Psi} \end{aligned} \tag{3}$$

The first term in (3) is upper bounded by $L_C \cdot \Delta_C^{\max}$ by bounding the indicator function directly. We show how to choose L_C properly via Lemma 1 and the discussions following it.

Lemma 1. *For clique C, we can always find $x_C \in (\mu_C^{\max}, \mu_1)$, $y_C \in (\mu_C^{\max}, \mu_1)$, and a sufficiently small $0 < \epsilon_C < 1$ such that the following hold simultaneously: (i) $\mu_C^{\max} < x_C < y_C < \mu_1$; (ii) $d(x_C, \mu_1) = \frac{1}{1+\epsilon_C} \cdot d(\mu_C^{\max}, \mu_1)$; (iii) $d(x_C, y_C) = \frac{1}{1+\epsilon_C} \cdot d(x_C, \mu_1)$; (iv) $d(x_C, y_C) \geq d(x_C, \mu_C^{\max})$.*

After fixing x_C, y_C, and ϵ_C that satisfy all the conditions in Lemma 1, set $L_C := \frac{\ln((N_C)^{\eta_C}\cdot T)}{d(x_C, y_C)} + 2$, where $N_C = \max_{i\in C} |\mathcal{N}_i|$ and $\eta_C := \frac{d(x_C, y_C)}{d(x_C, \mu_C^{\max})} \geq 1$ (condition (iv) in Lemma 1).

Several remarks are in order for Lemma 1 and the choice of L_C. Regarding the choice of x_i, y_i, and ϵ in the standard Thompson Sampling analysis in (Agrawal and Goyal, 2017), ϵ can be any value in $(0,1)$. They chose to fix $\epsilon \in (0,1)$ first, and then chose $x_i \in (\mu_i, \mu_1)$ such that $d(x_i, \mu_1) = \frac{d(\mu_i, \mu_1)}{1+\epsilon}$ and $y_i \in (x_i, \mu_1)$ such that $d(x_i, y_i) = \frac{d(x_i, \mu_1)}{1+\epsilon}$. However, in this paper, if we exactly reuse the ideas in (Agrawal and Goyal, 2017) to choose x_C and y_C, i.e., fixing $\epsilon_C \in (0,1)$ first and then choosing x_C and y_C only satisfying conditions (i), (ii), and (iii) in Lemma 1, and then set $L_C = \frac{\ln(T)}{d(x_C, y_C)} + 2$, to the best of our knowledge, for each clique C, we can only derive a problem-dependent regret bound for which the constant term scales with the clique size instead of logarithmically scaling with the clique size. To have a regret bound for which the constant term scales logarithmically with the clique size in a finite time horizon setting, we may sacrifice some freedom of the choice of ϵ_C. However, ϵ_C can always be chosen as small as desired.

The second term Ψ in (3) can be further decomposed into Ψ_1, Ψ_2, and Ψ_3 by introducing events $E_C^{\mu}(t) := \left\{\max_{i\in C\setminus 1} \hat{\mu}_i(t) \leq x_C\right\}$ and $E_C^{\theta}(t) := \left\{\max_{i\in C\setminus 1} \theta_i(t) \leq y_C\right\}$, which is shown in (4).

$$\begin{aligned} \Psi &= \underbrace{\sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i, \overline{E_C^{\mu}(t)}, T_C(t) \geq L_C\} \cdot \Delta_i}_{\Psi_1} \\ &\quad + \underbrace{\sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i, E_C^{\mu}(t), \overline{E_C^{\theta}(t)}, T_C(t) \geq L_C\} \cdot \Delta_i}_{\Psi_2} \\ &\quad + \underbrace{\sum_{t=1}^{T} \sum_{i\in C} \mathbf{1}\{I_t = i, E_C^{\mu}(t), E_C^{\theta}(t), T_C(t) \geq L_C\} \cdot \Delta_i}_{\Psi_3} \end{aligned} \tag{4}$$

After the aforementioned further regret decomposition,

we show[5] (see Lemma 4) that $\mathbb{E}[\Psi_1] \leq \frac{\Delta_C^{\max}}{d(x_C,\mu_C^{\max})}$. The key step in proving this result (see Lemma 3) is to show that after a fixed arm $i \in C \setminus 1$ has been observed enough times, i.e., $O_i(t) \geq L_C$, it is a rare event that its empirical mean $\hat{\mu}_i(t)$ is greater than x_C. Next, we upper bound $\mathbb{E}[\Psi_2]$ by $\Delta_C^{\max}$, which is accomplished by Lemma 6. This lemma relies on a result, Lemma 5, which states that after a fixed arm $i \in C \setminus 1$ has been observed enough times, i.e., $O_i(t) \geq L_C$, and its empirical mean $\hat{\mu}_i(t)$ is close enough to its true mean, i.e., $\hat{\mu}_i(t) \leq x_C$, it is a rare event that its posterior sampling value $\theta_i(t)$ is greater than y_C. Lemma 5 crucially relies on condition (iv) of Lemma 1, i.e. that $d(x_C,y_C) \geq d(x_C,\mu_C^{\max})$, without which we do not know if it is possible to obtain our desired bound in Lemma 5. This control is important, as Lemma 6 is proved roughly by taking a union bound over all the arms in C, of which there are at most $|C| \leq N_C$. Finally, we show that $\mathbb{E}[\Psi_3]$ is $O(1)$ in the sense that it does not grow with T; here, the Big-O notation hides problem-dependent constants. We do this via Lemma 8, which is roughly analogous to Lemmas 2.9 and 2.10 of Agrawal and Goyal (2017). We mention in passing that Lemma 8 relies on another result, Lemma 7, which is analogous to Lemma 2.8 of Agrawal and Goyal (2017).

Proof of Theorem 2. As we are analyzing the regret for clique C, for ease of presentation, we drop the subscript C in ϵ_C. Let $\phi_C := \ln\left(\frac{\mu_1(1-\mu_C^{\max})}{\mu_C^{\max}(1-\mu_1)}\right) > 0$, $\Delta_C' := \mu_1 - y_C$, and $D_C := d(y_C,\mu_1)$. Recall conditions (i) to (iv) in Lemma 1 when choosing x_C, y_C, and ϵ. From condition (ii), $d(x_C,\mu_1) = \frac{d(\mu_C^{\max},\mu_1)}{(1+\epsilon)}$, we have $x_C - \mu_C^{\max} \geq \frac{\epsilon}{\epsilon+1}\frac{d(\mu_C^{\max},\mu_1)}{\phi_C}$ due to the convexity of function $x \mapsto d(x,\mu_1)$ when $x \in [\mu_C^{\max},\mu_1]$. Then from Pinsker's inequality we have $\frac{1}{d(x_C,\mu_C^{\max})} \leq \frac{1}{2(x_C-\mu_C^{\max})^2} \leq \frac{(1+\epsilon)^2\phi_C^2}{2\epsilon^2(d(\mu_C^{\max},\mu_1))^2}$. Putting together condition (ii) and condition (iii), i.e., $d(x_C,y_C) = \frac{d(x_C,\mu_1)}{1+\epsilon}$ and $d(x_C,\mu_1) = \frac{d(\mu_C^{\max},\mu_1)}{1+\epsilon}$, we have $d(x_C,y_C) = \frac{d(\mu_C^{\max},\mu_1)}{(1+\epsilon)^2}$.

Now, rewriting $L_C = \frac{\ln(T)}{d(x_C,y_C)} + \frac{\ln(N_C)}{d(x_C,\mu_C^{\max})} + 2$ and by applying $\frac{1}{d(x_C,y_C)} = \frac{(1+\epsilon)^2}{d(\mu_C^{\max},\mu_1)}$ and $\frac{1}{d(x_C,\mu_C^{\max})} \leq \frac{(1+\epsilon)^2\phi_C^2}{2\epsilon^2(d(\mu_C^{\max},\mu_1))^2}$ to L_C, we have $L_C \leq \frac{(1+\epsilon)^2\ln(T)}{d(\mu_C^{\max},\mu_1)} + \frac{(1+\epsilon)^2\phi_C^2\ln(N_C)}{2\epsilon^2(d(\mu_C^{\max},\mu_1))^2} + 2$.

From (3) we have $\mathbb{E}[R_C(T)] \leq L_C \cdot \Delta_C^{\max} + \mathbb{E}[\Psi]$ and by applying Lemmas 4, 6, and 8, and using the above rewrite of L_C, we further have that $\mathbb{E}[R_C(T)]$ is at most

$$
\begin{aligned}
&L_C \cdot \Delta_C^{\max} + \underbrace{\frac{\Delta_C^{\max}}{d(x_C,\mu_C^{\max})}}_{\text{Lemma 4}} + \underbrace{\Delta_C^{\max}}_{\text{Lemma 6}} \\
&+ \underbrace{\frac{24\Delta_C^{\max}}{\Delta_C'^2} + O\left(\frac{\Delta_C^{\max}}{\Delta_C'^2} + \frac{\Delta_C^{\max}}{\Delta_C' D_C} + \frac{\Delta_C^{\max}}{\Delta_C'^4}\right)}_{\text{Lemma 8}} \\
&\leq \underbrace{\frac{\Delta_C^{\max}(1+\epsilon)^2\ln(T)}{d(\mu_C^{\max},\mu_1)} + \frac{\Delta_C^{\max}(1+\epsilon)^2\phi_C^2(\ln(N_C)+1)}{2\epsilon^2(d(\mu_C^{\max},\mu_1))^2}}_{(L_C-2)\cdot\Delta_C^{\max}} \\
&\quad + \underbrace{3\Delta_C^{\max} + \frac{24\Delta_C^{\max}}{\Delta_C'^2} + O\left(\frac{\Delta_C^{\max}}{\Delta_C'^2} + \frac{\Delta_C^{\max}}{\Delta_C' D_C} + \frac{\Delta_C^{\max}}{\Delta_C'^4}\right)}_{O(1)} \\
&\leq \frac{\Delta_C^{\max}(1+\epsilon')\ln(T)}{d(\mu_C^{\max},\mu_1)} + O\left(\frac{\ln(N_C)+1}{\epsilon'^2}\right) + O(1) \quad , \qquad (5)
\end{aligned}
$$

where $\epsilon' = 3\epsilon$ and the Big-O notations in the last inequality hide problem-dependent constants. □

Before presenting the proof of Theorem 3, we present a new lemma that gives a novel way to choose x_C and y_C. After fixing x_C and y_C, we prove Theorem 3 by exploiting the properties of the squared Hellinger distance (Tsybakov, 2009) and its link to the KL divergence $d(a,b)$. The squared Hellinger distance between two Bernoulli distributions with success probabilities a and b is defined as $d_H^2(a,b) := (\sqrt{a}-\sqrt{b})^2 + (\sqrt{1-a}-\sqrt{1-b})^2$.

Lemma 2. *For clique C and any $\epsilon \in (0,1)$, we can always find $x_C \in (\mu_C^{\max},\mu_1)$ and $y_C \in (\mu_C^{\max},\mu_1)$ such that $\mu_C^{\max} < x_C < y_C < \mu_1$ and $d(x_C,y_C) = d(x_C,\mu_C^{\max})$ hold simultaneously.*

Proof of Lemma 2. Fix $\epsilon \in (0,1)$ and then set $y_C = \mu_1 - \epsilon\Delta_C^{\min}$. Clearly, $y_C \in (\mu_C^{\max},\mu_1)$ as $\epsilon \in (0,1)$. Then we construct a monotonic function $h(b) = d(b,y_C) - d(b,\mu_C^{\max})$ where $b \in [\mu_C^{\max},y_C]$. Note that $h(b)$ is strictly decreasing when $b \in [\mu_C^{\max},y_C]$ since $h'(b) = \ln\left(\frac{\mu_C^{\max}}{y_C}\frac{1-y_C}{1-\mu_C^{\max}}\right) < 0$. Also, we know that $h(\mu_C^{\max}) = d(\mu_C^{\max},y_C) > 0$ and $h(y_C) = -d(y_C,\mu_C^{\max}) < 0$. Therefore, there exists a unique $m' \in (\mu_C^{\max},y_C)$ such that $h(m') = d(m',y_C) - d(m',\mu_C^{\max}) = 0$ and $m' = \mu_C^{\max} + \frac{d(\mu_C^{\max},y_C)(1-\epsilon)\Delta_C^{\min}}{d(\mu_C^{\max},y_C)+d(y_C,\mu_C^{\max})}$ by using the linearity of the function h. Now, set $x_C = m'$. Note that setting $x_C = m'$ guarantees $\eta_C = \frac{d(x_C,y_C)}{d(x_C,\mu_C^{\max})} = 1$, concluding the proof. □

[5] Lemmas 3 through 8 are in the supplementary material.

Proof of Theorem 3. After fixing x_C and y_C that satisfy the conditions in Lemma 2, all the proofs of Lemmas 3 through 8 still hold as only Lemma 5 needs to use the condition $\eta_C \geq 1$. Just as when proving Theorem 2, let $L_C = \frac{\ln((N_C)^{\eta_C} \cdot T)}{d(x_C, y_C)} + 2$, where $\eta_C = \frac{d(x_C, y_C)}{d(x_C, \mu_C^{\max})} = 1$. Then we have $\frac{1}{d(x_C, y_C)} = \frac{1}{d(x_C, \mu_C^{\max})} \leq \frac{\left(1 + \frac{d(y_C, \mu_C^{\max})}{d(\mu_C^{\max}, y_C)}\right)^2}{2(1-\epsilon)^2(\Delta_C^{\min})^2} = \frac{\left(1 + \frac{d(\mu_1 - \epsilon\Delta_C^{\min}, \mu_C^{\max})}{d(\mu_C^{\max}, \mu_1 - \epsilon\Delta_C^{\min})}\right)^2}{2(1-\epsilon)^2(\Delta_C^{\min})^2}$ by using Pinsker's inequality.

Let $\zeta_C := \frac{\left(1 + \frac{d(\mu_1 - \epsilon\Delta_C^{\min}, \mu_C^{\max})}{d(\mu_C^{\max}, \mu_1 - \epsilon\Delta_C^{\min})}\right)^2}{2(1-\epsilon)^2}$. Now we upper bound ζ_C. Let $V_C := \frac{\mu_1 - \epsilon\Delta_C^{\min}}{\mu_C^{\max}} > 1$. From Lemma 4 in (Yang and Barron, 1998) and the symmetric property of the squared Hellinger distance, we have $d(\mu_1 - \epsilon\Delta_C^{\min}, \mu_C^{\max}) \leq (2 + \log(V_C)) \cdot d_H^2(\mu_1 - \epsilon\Delta_C^{\min}, \mu_C^{\max}) = (2 + \log(V_C)) \cdot d_H^2(\mu_C^{\max}, \mu_1 - \epsilon\Delta_C^{\min}) \leq (2 + \log(V_C)) \cdot d(\mu_C^{\max}, \mu_1 - \epsilon\Delta_C^{\min})$. Then we have $\zeta_C \leq \frac{(3+\log(V_C))^2}{2(1-\epsilon)^2}$.

Recalling that $\Delta_C^{'} = \mu_1 - y_C$ and $D_C = d(y_C, \mu_1)$ and by applying $y_C = \mu_1 - \epsilon\Delta_C^{\min}$ to $\Delta_C^{'}$ and D_C, we have $\Delta_C^{'} = \epsilon\Delta_C^{\min}$ and $D_C = d(\mu_1 - \epsilon\Delta_C^{\min}, \mu_1) \leq \epsilon^2(\Delta_C^{\min})^2$. Now applying L_C, Lemma 4, Lemma 6, and Lemma 8 to (5), we have that $\mathbb{E}[R_C(T)]$ is at most

$$
\begin{aligned}
& L_C \cdot \Delta_C^{\max} + \underbrace{\frac{\Delta_C^{\max}}{d(x_C, \mu_C^{\max})}}_{\text{Lemma 4}} + \underbrace{\Delta_C^{\max}}_{\text{Lemma 6}} \\
& + \underbrace{\frac{24\Delta_C^{\max}}{\epsilon^2(\Delta_C^{\min})^2} + O\left(\frac{\Delta_C^{\max}}{\epsilon^4(\Delta_C^{\min})^4}\right)}_{\text{Lemma 8}} \\
& \leq \frac{(3 + \log(V_C))^2 \Delta_C^{\max} \ln(N_C \cdot T)}{2(1-\epsilon)^2(\Delta_C^{\min})^2} \\
& \quad + \frac{(3 + \log(V_C))^2 \Delta_C^{\max}}{2(1-\epsilon)^2(\Delta_C^{\min})^2} + O\left(\frac{\Delta_C^{\max}}{\epsilon^4(\Delta_C^{\min})^4}\right) \quad ,
\end{aligned}
$$

where $V_C = \frac{\mu_1 - \epsilon\Delta_C^{\min}}{\mu_C^{\max}}$. As we explain at the end of the proof of Lemma 8, instead of paying $O\left(\frac{\Delta_C^{\max}}{\epsilon^4(\Delta_C^{\min})^4}\right)$, an alternative is to pay $O\left(\frac{\Delta_C^{\max} \ln(T \cdot \epsilon\Delta_C^{\min})}{(\epsilon)^2(\Delta_C^{\min})^2}\right)$. □

6 EXPERIMENTAL RESULTS

We conducted experiments with fixed (i.e., not time-varying) undirected feedback graphs with two equally-sized cliques. The reward for each arm is generated i.i.d. according to a Bernoulli distribution and the rewards of all the arms in a given round are independently generated. In the experiments, there is only one optimal arm, which means one clique can include the unique optimal arm while the other clique only contains sub-optimal arms. Also, we set all the sub-optimal arms with the same mean reward (and hence the same gap) so as to let $\Delta_C^{\max} = \Delta_C^{\min} =: \Delta$, thereby removing other factors that may impact the regret. Consequently, only the clique size (which we vary over our experiments) impacts the regret. In our experiments, we double the number of arms in each clique to study the effect of clique size on the regret. We start at 2 arms per clique (hence 4 arms total) until hitting 1024 arms per clique (2056 arms total). Each experiment is run for $T = 35,000$ rounds for each run, and we take the average of 100 independent runs.

We compare the performance of UCB-NE, UCB-N, TS-N, the elimination-based algorithm of Cohen et al. (2016), and an algorithm called TS-MaxN devised by Tossou et al. (2017). The reason that we do not compare to UCB-MaxN is that it becomes equivalent to UCB-N for our choice of feedback graphs. Algorithm 3 presents the TS-MaxN algorithm. Compared to TS-N, instead of pulling the arm with the highest posterior sampling value, i.e., $J_t \leftarrow \arg\max_{i \in \mathcal{N}} \theta_i(t)$, in TS-MaxN the learner pulls the arm with the highest empirical mean among all the neighboring arms of J_t, i.e., $I_t \leftarrow \arg\max_{i \in \mathcal{N}_{J_t}} \hat{\mu}_i(t)$. Note that the learner needs the knowledge of the feedback graph for TS-MaxN.

As can be seen from our experimental results (see Figure 2, the elimination-based algorithm does not perform well practically. Also, the regret bound of the elimination-based algorithm is $O\left(|\mathcal{C}|\frac{\ln(|\mathcal{N}|) \cdot \ln(T)}{\Delta}\right)$ while UCB-NE's regret bound is $O\left(|\mathcal{C}|\frac{\ln(T) + \ln(|\mathcal{N}|)}{\Delta}\right)$. Hence, our selected problem instances are ones for which UCB-NE's theoretical guarantee is better than that of the elimination-based algorithm. Figure 1 shows the regret of all the remaining algorithms except for the elimination algorithm. We can see that although the number of arms per clique increases exponentially, the regret grows almost linearly with respect to $\ln(|C|)$ for UCB-NE and TS-N. Also, UCB-N always performs better than UCB-NE, TS-N always performs better than UCB-N and UCB-NE, and TS-MaxN performs better than TS-N.

Algorithm 3 TS-MaxN (Tossou et al., 2017)

Set $O_i \leftarrow 0, Q_i \leftarrow 0, \forall i \in \mathcal{N}$
for $t = 1 : T$ **do**
  Sample $\theta_i(t)$ from $Beta(Q_i + 1, O_i - Q_i + 1), \forall i \in \mathcal{N}$
  Locate arm $J_t \leftarrow \arg\max_{i \in \mathcal{N}} \theta_i(t)$
  Pull arm $I_t \leftarrow \arg\max_{i \in \mathcal{N}_{J_t}} \hat{\mu}_i(t)$
  for $i \in \mathcal{N}_{I_t}$ **do**
    Set $O_i \leftarrow O_i + 1$; Observe $X_i(t)$
    Set $Q_i \leftarrow Q_i + X_i(t)$
  end for
end for

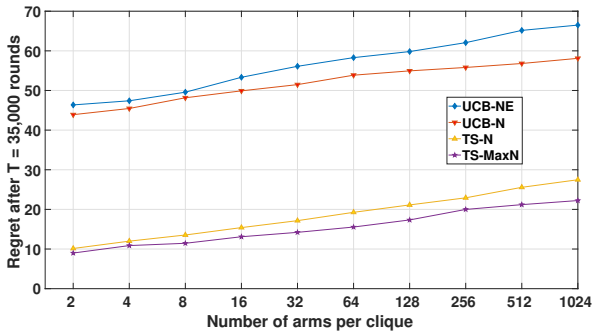

Figure 1: Regret for UCB-NE, UCB-N, TS-N, and TS-MaxN with different number of arms per clique.

7 CONCLUSION AND OPEN PROBLEMS

In this work, we have shown new problem-dependent regret bounds for the stochastic multi-armed bandit problem with feedback graphs. Our UCB-style algorithm, UCB-NE, is the first algorithm of this type that provably obtains regret that is linear in the size of a clique covering rather than linear in the total number of arms. Our regret bounds for the Thompson Sampling-style algorithm TS-N are the first problem-dependent regret bounds for Thompson Sampling that improve with side observations. To ensure that the regret bound is linear in the size of a clique covering rather than linear in the total number of arms, we required important innovations to the previous analysis of Agrawal and Goyal (2017).

While UCB-NE achieves this by improving the constant term (relative to UCB-N) to $O\left(\sum_{C \in \mathcal{C}} \frac{\Delta_C^{\max} \ln\left(\max_{i \in C} N_i\right)}{(\Delta_C^{\min})^2}\right)$, we still believe that a similar constant term can be achieved for UCB-N without modifying the way of constructing upper confidence bounds. We make this conjecture due to the following facts: *(i)* Even without modifying TS-N to make it explore more, we were still able to obtain regret bounds with a constant term that scales logarithmically with the clique size. *(ii)* Our experimental results for UCB-N in Figure 1 show that the regret increases roughly linearly as the clique size increases exponentially. Another open problem is whether we need to pay the leading term $\frac{8\Delta_C^{\max} \ln(T)}{(\Delta_C^{\min})^2}$ for each clique. Our conjecture is that it is possible to only pay a price of $O\left(\frac{\ln(T)}{\Delta_C^{\min}}\right)$ for the leading term of each clique.

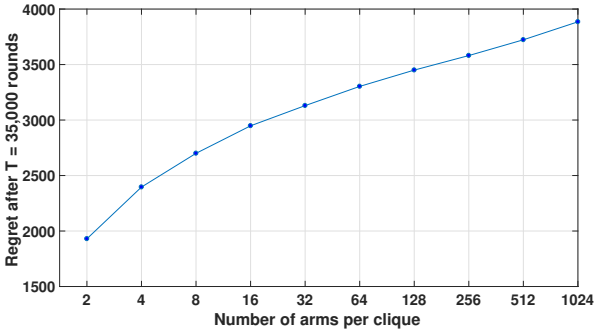

Figure 2: Regret for elimination algorithm with different number of arms per clique.

Regarding the elimination-based algorithm in (Cohen et al., 2016), although the independence number is always no greater than clique covering number, their regret bound's leading term scales with the worst $O(\alpha(\mathcal{G}) \cdot \ln(|\mathcal{N}|))$ arms. Instead, for regret bounds that depend on clique coverings, for each clique we pay for its worst arm once. If an undirected feedback graph satisfies $\alpha(\mathcal{G}) = |\mathcal{C}|$, the leading term of the elimination-based algorithm is $O\left(|\mathcal{C}| \frac{\ln(|\mathcal{N}|) \cdot \ln(T)}{\Delta}\right)$ while UCB-NE can achieve $O\left(|\mathcal{C}| \frac{\ln(T)}{\Delta}\right)$.

Just like the experimental results in Tossou et al. (2017), our experimental results in Figure 1 also confirm that TS-MaxN outperforms TS-N practically. Therefore, it is desirable to have a problem-dependent regret bound for TS-MaxN and we also believe that the constant term also scales logarithmically with the clique size.

Acknowledgements

This work was partially supported by the NSERC Discovery Grant RGPIN-2018-03942, CFI, and BCKDF.

References

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson Sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2): 55–65, 2010.

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

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 142–151. AUAI Press, 2012.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pages 811–819, 2016.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pages 359–376, 2011.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

Fang Liu, Swapna Buccapatnam, and Ness Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3643–3650, 2018a.

Fang Liu, Zizhan Zheng, and Ness Shroff. Analysis of Thompson Sampling for graphical bandits without the graphs. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 13–22. AUAI Press, 2018b.

Aristide CY Tossou, Christos Dimitrakakis, and Devdatt Dubhashi. Thompson Sampling for stochastic bandits with graph feedback. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 2660–2666, 2017.

Alexandre B Tsybakov. Introduction to nonparametric estimation. Revised and extended from the 2004 french original. Translated by vladimir zaiats, 2009.

Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.