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 conduct experiments 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 $\mu$ 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 $O\left(\frac{\Delta K \log T}{d(\mu - \Delta, \mu)}\right)^{\frac{1}{4}}$.

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-

\[ d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \]

is the KL divergence of a Bernoulli distribution with success probability $p$ from a Bernoulli distribution with success probability $q$.
algorithm 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 $\epsilon$ 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 $\epsilon$. 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 $\epsilon$. 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 $G := (\mathcal{N}, \mathcal{E})$ 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 $i$, i.e., $\mathcal{E}$ includes all self-loops. We assume that graph $G$ does not vary over time. For each $i \in \mathcal{N}$, let set $\mathcal{N}_i \subseteq \mathcal{N}$ collect arm $i$ and all its neighbors in $G$. In each round $t$, the learner pulls an arm $I_t \in \mathcal{N}$ and then observes the reward of each arm in $\mathcal{N}_i$. 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.
Let $\mu_i$ 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 $R(T)$, which is defined as

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

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 $\bigcup_{C \in \mathcal{C}} \mathcal{N}$ 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 $R(T)$ can be further expressed as

$$ R(T) = \sum_{i \in \mathcal{N}} \sum_{t=1}^{T} \mathbb{E} [\mathbb{1}\{I_t = i\}] \cdot \Delta_i $$

$$ \leq \sum_{C \in \mathcal{C}} \mathbb{E} \left[ \sum_{i \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}\{I_t = i\} \cdot \Delta_i \right], $$

where $R_C(T) := \sum_{i \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{1}\{I_t = i\} \cdot \Delta_i$ 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 $\mu^\text{max}_C := \max_{i \in C \setminus \{1\}} \mu_i$, $\Delta^\text{max}_C := \max_{i \in C \setminus \{1\}} \Delta_i$, and $\Delta^\text{min}_C := \min_{i \in C \setminus \{1\}} \Delta_i$.

### 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 $O \left( \sum_{C \in \mathcal{C}} \frac{\Delta^\text{max}_C \ln(T)}{(\Delta^\text{min}_C)^2} \right)$ 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 $\delta$ 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 $\delta$ 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\footnote{Their algorithm admits regret bounds even if $\mathcal{G}$ varies over time.} 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 $O \left( \frac{\ln(T)}{\alpha(\mathcal{G})} \right)$, 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
compared to the standard value of the term in the upper confidence bound is enlarged as the empirical mean of arm $i$ until the end of round $t$. Let $\hat{\mu}_i(t) := \bar{\mu}_i O_i(t)$ 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 $\sqrt{2 \ln(1/\Delta_i)}$. 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., $I_t := \arg \max_{i \in \mathcal{N}_t} \hat{\mu}_i(t)$. 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

1: Set $O_i \leftarrow 0, \hat{\mu}_i O_i \leftarrow 0, \forall i \in \mathcal{N}$
2: for $t = 1 : T$ do
3: Set $\tilde{\mu}_i(t) = \bar{\mu}_i O_i(t-1) + \sqrt{\frac{2 \ln(1/\Delta_i)}{O_i(t-1)}}, \forall i \in \mathcal{N}$
4: Pull arm $I_t \leftarrow \arg \max_{i \in \mathcal{N}} \tilde{\mu}_i(t)$
5: for $i \in \mathcal{N}_{I_t}$ do
6: Set $O_i \leftarrow O_i + 1$; Observe $X_i(t)$
7: Set $\hat{\mu}_i O_i \leftarrow \frac{\hat{\mu}_i O_i(-1) + X_i(t)}{O_i}$
9: end for

4.2 REGRET ANALYSIS

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

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

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

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 $\mathcal{C} = \{1\}, \forall i \in \mathcal{N}$ to recover the regret bound of this classic setting. From $\mathcal{C} = \{1\}, \forall i \in \mathcal{N}$ we have $\Delta_{\text{min}} = \Delta_{\text{max}} = \Delta_i$ 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-NE 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 $O(|\mathcal{C}|)$ while UCB-NE improves to $O \left( \frac{\ln(N_C)}{\Delta_i} \right)$ when the clique size is large. However, when taking the trivial clique covering $\mathcal{C} = \{1\}, \forall i \in \mathcal{N}$, UCB-N boils down to the same regret bound as UCB1 while UCB-NE needs to pay an additional price of $\frac{2 \ln(1/\Delta_i)}{\Delta_i}$ 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 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 $L_C := \frac{8 \ln(N_C \cdot T)}{\left(\Delta_{\text{min}}^2\right)^2}$, where we recall that $N_C := \max_{i \in \mathcal{C}} \left\{ |\mathcal{N}_i|^{1/4} \right\}$. For the rounds when clique $C$ is in the under-sampled regime, the total regret is at most $L_C \cdot \Delta_{\text{max}}^2$, 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 $|\mathcal{N}_i|^{1/4}$.
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 $\alpha, \beta$. 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)

1: Set $O_i \leftarrow 0, Q_i \leftarrow 0, \forall i \in \mathcal{N}$
2: for $t = 1 : T$ do
3: Sample $\theta_i(t)$ from $Beta(Q_i + 1, O_i - Q_i + 1), \forall i \in \mathcal{N}$
4: Pull arm $I_t \leftarrow \arg \max_{i \in \mathcal{N}} \theta_i(t)$
5: for $i \in N_{I_t}$ do
6: Set $O_i \leftarrow O_i + 1; \text{Observe } X_i(t)$
7: Set $Q_i \leftarrow Q_i + X_i(t)$
8: end for
9: end for

5.2 REGRET ANALYSIS

Let $N_C := \max_{i \in C} |\mathcal{N}_i|$ and, for $a, b \in [0, 1]$, define $d(a, b) := a \ln \left( \frac{a}{b} \right) + (1 - a) \ln \left( \frac{1 - a}{1 - b} \right)$ 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

\[
\inf_C \left\{ \sum_{C \in C, C \neq \{1\}} \mathbb{E} \left[ R_C(T) \right] \right\} \\
\leq \inf_C \left\{ \sum_{C \in C, C \neq \{1\}} \frac{\Delta_{C}^{\text{max}} (1 + \epsilon_C) \ln(T)}{d(p_C^{\text{max}}, \mu_1)} + O \left( \frac{\ln(N_C) + 1}{(\epsilon_C)^2} \right) \right\},
\]

where $\epsilon_C$ can be any value in $(0, \min\left\{ \frac{d(\mu_{C}^{\text{max}}, \mu_1)}{d(\mu_1, \mu_1)}, 1 \right\})$ and $\mu_C \in (\mu_{C}^{\text{max}}, \mu_1)$ 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 $C = \{\{i\}, \forall i \in \mathcal{N}\}$ to represent the case where there is no feedback graph. Then we have $\mu_{C}^{\text{max}} = \mu_1$ 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 $\epsilon_C$. 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 $\epsilon_C$ in our regret bound. We use $\epsilon_C$ to control the problem-dependent constant term to make it scale logarithmically with the clique size. It is important to note that $\epsilon_C$ does not depend on the number of arms within clique $C$. Instead, $\epsilon_C$ only depends on $\mu_1$ and $\mu_{C}^{\text{max}}$ (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 $\epsilon_C$ 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 $\epsilon$ properly.

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

$$\inf C \left\{ \sum_{C \in C, C \neq \{1\}} \mathbb{E} [R_C(T)] \right\} \leq \inf C \left\{ \sum_{C \in C, C \neq \{1\}} \frac{(3 + \lambda_C^2) \Delta_{\text{max}} \ln(T)}{2(1 - e)^2 \Delta_{\text{max}}^2} + O \left( \frac{\Delta_{\text{max}}}{\epsilon^2 \Delta_{\text{max}}^2} \right) \right\},$$

where $\lambda_C := \log \left( \frac{\mu_1 - \epsilon \Delta_{\text{min}}}{\mu_{\text{max}}} \right)$.

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

**Notation and definitions:** Before presenting the analysis, we first introduce some important notation and definitions. Let $T_i(t) := \sum_{s=1}^{t-1} 1 \{ I_s = i \}$ 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., $T_C(t) := \sum_{s=1}^{t-1} 1 \{ \exists j \in C \ s.t. \ I_s = j \}$. Different from UCB-NE, in TS-N, $\hat{\mu}_i(t) = O_i(t) = \frac{Q_i(t)}{O_i(t) + 1}$ is defined as the empirical mean of arm $i$ at round $t$. $F_t$ collects all the history information until the end of round $t$ sequentially, which is $F_t = \{ I_s, X_i(s), \forall i \in \mathcal{N}_t, s = 1, 2, \ldots, t \}$. Define $F_0 = \{ \}$. Also note that $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_{T-1}$ always holds. For each arm $i$, $O_i(t), Q_i(t)$, and $\hat{\mu}_i(t)$ are determined by $F_{t-1}$. Similarly, the distribution that generates $\theta_i(t)$ is determined by $F_{t-1}$.

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.

$$R_C(T) = \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i \} \cdot \Delta_i$$

$$= \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, T_C(t) < L_C \} \cdot \Delta_i$$

$$\leq L_C \Delta_{\text{max}}$$

$$+ \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, T_C(t) \geq L_C \} \cdot \Delta_i$$

The first term in (3) is upper bounded by $L_C \cdot \Delta_{\text{max}}$ 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 $x_C \in (\mu_{\text{max}}^+, \mu_1), y_C \in (\mu_{\text{max}}^+, \mu_1)$, and a sufficiently small $0 < \epsilon_C < 1$ such that the following hold simultaneously: (i) $\mu_{\text{max}}^+ < x_C < y_C < \mu_1$; (ii) $d(x_C, \mu_1) = \frac{1}{1 + \epsilon_C} \cdot d(\mu_{\text{max}}^+, \mu_1)$; (iii) $d(x_C, y_C) = \frac{1}{1 + \epsilon_C} \cdot d(x_C, \mu_1)$; (iv) $d(x_C, y_C) \geq d(x_C, \mu_{\text{max}}^+)$. After fixing $x_C, y_C$, and $\epsilon_C$ that satisfy all the conditions in Lemma 1, set $L_C := \frac{\ln \left( \sum_{i \in C} T_i(t) \right)}{d(x_C, y_C)} + 2$, where $N_C = \max |\mathcal{N}_t|$ and $\eta_C := \frac{d(x_C, y_C)}{d(x_C, \mu_{\text{max}}^+)}$. Several remarks are in order for Lemma 1 and the choice of $L_C$. Regarding the choice of $x_C, y_C$, and $\epsilon_C$ in the standard Thompson Sampling analysis in (Agrawal and Goyal, 2017), $\epsilon_C$ can be any value in $(0, 1)$.

The second term $\Psi$ in (3) can be further decomposed into $\Psi_1$, $\Psi_2$, and $\Psi_3$ by introducing event

$$E^\theta_C(t) := \left\{ \max_{i \in C \setminus 1} \theta_i(t) \leq y_C \right\} \quad \text{and} \quad E^\mu_C(t) := \left\{ \max_{i \in C \setminus 1} \mu_i(t) \leq x_C \right\},$$

which is shown in (4).

$$\Psi = \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, E^\mu_C(t), T_C(t) \geq L_C \} \cdot \Delta_i$$

$$\Psi_1$$

$$+ \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, E^\theta_C(t), E^\mu_C(t), T_C(t) \geq L_C \} \cdot \Delta_i$$

$$\Psi_2$$

$$+ \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, E^\mu_C(t), E^\theta_C(t), T_C(t) \geq L_C \} \cdot \Delta_i$$

$$\Psi_3$$

After the aforementioned further regret decomposition,
Lemma 6 is proved roughly by taking a union bound over
Proof of Theorem 2.
which is roughly analogous to Lemmas 2.9 and 2.10 of
it does not grow with $T$
next, we upper bound
Lemma 8.

Proof of Theorem 2. As we are analyzing the regret for clique $C$, for ease of presentation we drop the subscript $C$ in $e_C$. Let $\Psi_C := \ln(\rho_{C}(1-p_{C}^{\max})/(1-p_1)) > 0$, $\Delta_C := \mu_1 - y_C$, and $d_{C} := d(y_C, \mu_1)$. Recall conditions (i) to (iv) in Lemma 1 when choosing $x_C, y_C$, and $\epsilon$. From condition (ii), $d(x_C, \mu_1) = \frac{d(\mu_1^{\max}, \mu_1)}{(1+\epsilon)\epsilon}$, we have $x_C - \mu_1^{\max} \geq \epsilon \frac{d(\mu_1^{\max}, \mu_1)}{(1+\epsilon)\epsilon}$ due to the convexity of function $x \mapsto d(x, \mu_1)$ when $x \in [\mu_1^{\max}, \mu_1]$. Then from Pinsker’s inequality we have $\frac{1}{2(x_C - \mu_1^{\max})^2} \leq \frac{(1+\epsilon)^2\Phi}{2\epsilon^2(d(\mu_1^{\max}, \mu_1))^2}$. Putting together condition (ii) and condition (iii), i.e., $d(x_C, y_C) = \frac{d(x_C, \mu_1)}{1+\epsilon}$ and $d(x_C, \mu_1) = \frac{d(\mu_1^{\max}, \mu_1)}{(1+\epsilon)\epsilon}$, we have $d(x_C, y_C) = d(\mu_1^{\max}, \mu_1)$.

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

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 $x_C \in (\mu_1^{\max}, \mu_1)$ and $y_C \in (\mu_1^{\max}, \mu_1)$ such that $\mu_1^{\max} < x_C < y_C < \mu_1$ and $d(x_C, y_C) = d(\mu_1^{\max}, \mu_1)$ hold simultaneously.

Proof of Lemma 2. Fix $\epsilon \in (0, 1)$ and then set $y_C = \mu_1 - \epsilon\Delta_C^{\min}$. Clearly, $y_C \in (\mu_1^{\max}, \mu_1)$ as $\epsilon \in (0, 1)$. Then we construct a monotonic function $h(b) = d(b, y_C) - d(b, \mu_1^{\max})$ where $b \in [\mu_1^{\max}, y_C]$. Note that $h(b)$ is strictly decreasing when $b \in [\mu_1^{\max}, y_C]$ since $h'(b) = \ln(\mu_1^{\max} - b)/(\mu_1 - \mu_1^{\max}) < 0$. Also, we know that $h(\mu_1^{\max}) = d(\mu_1^{\max}, y_C) < 0$ and $h(y_C) = -d(y_C, \mu_1^{\max}) < 0$. Therefore, there exists a unique $m' \in (\mu_1^{\max}, y_C)$ such that $h(m') = d(m', y_C) - d(m', \mu_1^{\max}) = 0$ and $m' = \mu_1^{\max} + d(\mu_1^{\max}, y_C)/(1-\epsilon)\Delta_C^{\min}$ by using the linearity of the function $h$. Now, set $x_C = m'$. Note that setting $x_C = m'$ guarantees $\mathbb{E}[R_C(T)] = 1$, concluding the proof.
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 \( L_C = \frac{\ln(N_C y_C T)}{d(x_C y_C)} + 2 \), where

\[
\eta_C = \frac{d(x_C y_C)}{d(x_C y_C \max)} = 1.
\]

Then we have

\[
\frac{1}{d(x_C y_C \max)} \leq \frac{(1 + d(y_C \max))}{d(x_C y_C \max)} = \frac{1 + d(y_C \max)}{d(x_C y_C \max)} = (1 + \frac{\Delta_C \min}{\Delta_C \max})^2,
\]

by using Pinsker’s inequality.

Let \( \zeta_C := \frac{(1 + \frac{\Delta_C \min}{\Delta_C \max})^2}{2(1 - \epsilon)^2} \). Now we upper bound \( \zeta_C \). Let \( V_C := \frac{\mu_1 - \epsilon \Delta_C \max}{\mu_\max} > 1 \). From Lemma 4 in (Yang and Barron, 1998) and the symmetric property of the squared Helliger distance, we have

\[
d(\mu_1 - \epsilon \Delta_C \max, \mu_\max) \leq (2 + \log(V_C)) \cdot d_H^2(\mu_1 - \epsilon \Delta_C \max, \mu_\max) = (2 + \log(V_C)) \cdot d_H^2(\mu_\max, \mu_1 - \epsilon \Delta_C \max) \leq (2 + \log(V_C)) \cdot d(\mu_\max, \mu_1 - \epsilon \Delta_C \max).
\]

Recalling that \( \Delta_C = \mu_1 - \mu_C \) 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

\[
\Delta_C = \epsilon \Delta_C \min + D_C = d(\mu_1 - \epsilon \Delta_C \min, \mu_1) \leq \epsilon^2(\Delta_C \min)^2.
\]

Now applying \( L_C \), Lemma 4, Lemma 6, and Lemma 8 to (5), we have that \( \mathbb{E}[R_C(T)] \) is at most

\[
L_C \cdot \Delta_C \max + \frac{\Delta_C \max}{d(x_C, \mu_\max)} + \frac{\Delta_C \max}{2(1 - \epsilon)^2} + O\left(\frac{\Delta_C \max}{e^4(\Delta_C \min)^4}\right)
\]

\[
\leq \frac{(3 + \log(V_C))^2 \Delta_C \max}{2(1 - \epsilon)^2(\Delta_C \min)^2} + O\left(\frac{\Delta_C \max}{e^4(\Delta_C \min)^4}\right),
\]

where \( V_C = \frac{\mu_1 - \epsilon \Delta_C \min}{\mu_\max} \). As we explain at the end of the proof of Lemma 8, instead of paying \( O\left(\frac{\Delta_C \max}{e^4(\Delta_C \min)^4}\right) \), an alternative is to pay \( O\left(\frac{\Delta_C \max}{e^4(\Delta_C \min)^4}\right) \), \( \square \)

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 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). We set the gaps for the suboptimal arms to be the same 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, starting at 2 arms per clique (hence 4 arms total) until we hit 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-N, UCB-NE, 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 why 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 in detail. Compared to TS-N, instead of pulling the arm with the highest posterior sampling value, i.e., \( I_t \leftarrow \arg\max_{i \in N} \theta_i(t) \), in TS-MaxN the learner pulls the arm with the highest empirical mean among all the neighboring arms of \( I_t \), i.e., \( I_t \leftarrow \arg\max_{i \in N} \mu_i(t) \). 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 \( O\left(\frac{\ln(N)}{\Delta}\right) \) while UCB-NE’s regret bound is \( O\left(\frac{\ln(T) + \ln(|N|)}{\Delta}\right) \). 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)

1: Set $O_i \leftarrow 0, Q_i \leftarrow 0, \forall i \in \mathcal{N}$
2: for $t = 1 : T$ do
3: Sample $\theta_i(t)$ from $Beta(Q_i + 1, O_i - Q_i + 1), \forall i \in \mathcal{N}$
4: Locate arm $J_i \leftarrow \arg \max_{i \in \mathcal{N}} \theta_i(t)$
5: Pull arm $I_i \leftarrow \arg \max_{i \in \mathcal{N}} \tilde{\mu}_i(t)$
6: for $i \in \mathcal{N}_J$ do
7: Set $O_i \leftarrow Q_i + 1; \text{Observe } X_i(t)$
8: Set $Q_i \leftarrow O_i + X_i(t)$
9: end for
10: 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 $O\left(\sum_{C \in \mathcal{C}} \frac{\Delta_C^{\max} \ln \left(\frac{\max N_i}{\Delta_i^{\min}}\right)}{(\Delta_i^{\min})^2}\right)$, we still believe that the same constant term can be achieved for UCB-N, i.e., 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 $\frac{8\Delta^{\max} \ln(T)}{\Delta^{\min} T^2}$ for each clique. Our conjecture is that it is possible to only pay a price of $O\left(\frac{\ln(T)}{\Delta^{\min}}\right)$ for the leading term of each 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 $O(\alpha(G)) \cdot \ln(|\mathcal{N}|)$ 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 $\alpha(G) = |\mathcal{C}|$, 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.
References


A Proofs not appearing in the paper

A.1 Proof of Theorem 1

Proof of Theorem 1. To derive a regret bound of UCB-NE, let $T_i(t) := \sum_{s=1}^{t} 1\{S_s = i\}$ be the total number of times that arm $i$ has been pulled until the end of round $t$ and $T_C(t)$ be the total number of times that the learner pulls any arm in clique $C$ until the end of round $t$, i.e., $T_C(t) := \sum_{s=1}^{t} 1\{\exists j \in C \text{ s.t. } S_s = j\}$. Recall that $N_C = \max_{i \in C} \{ |N_i|^{\frac{1}{2}} \}$ and $L_C = \left[ \frac{8 \ln(N_C T)}{(\Delta_C^{\text{min}})^2} \right]$. Then we have

$$
\mathbb{E} \left[ R_C(T) \right] = \sum_{i \in C} \sum_{t=1}^{T} \mathbb{E} \left[ 1\{I_t = i\} \right] \cdot \Delta_i \\
= \sum_{i \in C \setminus 1} \sum_{t=1}^{T} \mathbb{E} \left[ 1\{I_t = i\} \right] \cdot \Delta_i \\
\leq \sum_{i \in C \setminus 1} \sum_{t=1}^{T} \mathbb{E} \left[ 1\{\exists i \in C \setminus 1 \text{ s.t. } \bar{\mu}_i(t) \geq \bar{\mu}_1(t), T_i(t) > T_i(t-1), T_C(t-1) \leq L_C \} \right] \cdot \Delta_C^{\text{max}} \\
+ \sum_{i \in C \setminus 1} \sum_{t=1}^{T} \mathbb{E} \left[ 1\{\exists i \in C \setminus 1 \text{ s.t. } \bar{\mu}_i(t) > \bar{\mu}_1(t), T_i(t) > T_i(t-1), T_C(t-1) > L_C \} \right] \cdot \Delta_C^{\text{max}} \\
\leq L_C \cdot \Delta_C^{\text{max}} + \sum_{i \in C \setminus 1} \max_{\bar{\mu}_i(t)} \mathbb{E} \left[ 1\{\exists i \in C \setminus 1 \text{ s.t. } \bar{\mu}_i(t) > \bar{\mu}_1(t), T_i(t) > T_i(t-1), T_C(t-1) > L_C \} \right] \cdot \Delta_C^{\text{max}} .
$$

(6)

Now, we analyze term (a). We have

$$
(a) = \mathbb{E} \left[ 1\{\exists i \in C \setminus 1 \text{ s.t. } \bar{\mu}_i(t) \geq \bar{\mu}_1(t), T_i(t) > T_i(t-1), T_C(t-1) > L_C \} \right] \\
\leq \mathbb{E} \left[ 1\left\{ \max_{i \in C \setminus 1} \bar{\mu}_i(t) > \bar{\mu}_1(t), T_i(t) > T_i(t-1), T_C(t-1) > L_C \right\} \right] \\
\leq \mathbb{E} \left[ 1\left\{ \max_{i \in C \setminus 1} \left( \tilde{\mu}_i, \tilde{O}_1(t-1) \right) + \sqrt{\frac{2 \ln(|N_i| \cdot t)}{O_1(t-1)}} \right) \geq \tilde{\mu}_1, \tilde{O}_1(t-1) + \sqrt{\frac{2 \ln(|N_1| \cdot t)}{O_1(t-1)}}, T_i(t) > T_i(t-1), T_C(t-1) > L_C \right\} \right] \\
\leq \mathbb{E} \left[ 1\left\{ \max_{L_C < O_C < t \in i \in C \setminus 1} \left( \tilde{\mu}_i, O_C \right) + \sqrt{\frac{2 \ln(|N_i| \cdot t)}{O_C}} \right) \geq \min_{0 < O_1 < t} \tilde{\mu}_1, O_1 + \sqrt{\frac{2 \ln(|N_1| \cdot t)}{O_1}} , T_i(t) > T_i(t-1) \right\} \right] \\
\leq \sum_{O_C = L_C}^{t-1} \sum_{O_1 = 1}^{t-1} \mathbb{E} \left[ 1\left\{ \max_{i \in C \setminus 1} \left( \tilde{\mu}_i, O_C \right) + \sqrt{\frac{2 \ln(|N_i| \cdot t)}{O_C}} \right) \geq \tilde{\mu}_1, O_1 + \sqrt{\frac{2 \ln(|N_1| \cdot t)}{O_1}} , T_i(t) > T_i(t-1) \right\} \right] \\
\leq \sum_{O_C = L_C}^{t-1} \sum_{O_1 = 1}^{t-1} \mathbb{E} \left[ \max_{i \in C \setminus 1} \left( \tilde{\mu}_i, O_C \right) + \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}} \right) \geq \tilde{\mu}_1, O_1 + \sqrt{\frac{2 \ln(|N_1| \cdot t)}{O_1}} \right] .
$$

(7)
If \((\beta)\) holds, it means at least one of the following three inequalities must hold:

\[
\begin{cases}
\hat{\mu}_{1,O_1} \leq \mu_1 - \sqrt{\frac{2 \ln(|N_1|^\frac{1}{2} \cdot t)}{O_1}}, \\
\max_{i \in C \setminus 1} \hat{\mu}_{i,O_C} \geq \mu_c^{\text{max}} + \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}}, \\
\mu_1 < \mu_c^{\text{max}} + 2 \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}}.
\end{cases}
\]

(8)\(\quad\) (9)\(\quad\) (10)

Note that when \(O_C \geq \frac{8 \ln(N_C \cdot T)}{(\Delta_c^{\text{max}})^2}\), inequality \(\{\mu_1 < \mu_c^{\text{max}} + 2 \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}}\}\) cannot be true.

Then we have

\[
(\alpha) \leq \frac{t-\ln}{O_C = L_C} \sum_{O_1=1}^{t-1} \mathbb{E} \left[1 \left\{ \max_{i \in C \setminus 1} \hat{\mu}_{i,O_C} \geq \mu_c^{\text{max}} + \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}} \right\} + 1 \left\{ \hat{\mu}_{1,O_1} \leq \mu_1 - \sqrt{\frac{2 \ln(|N_1|^\frac{1}{2} \cdot t)}{O_1}} \right\} \right]
\]

\[
= \sum_{O_C = L_C}^{t-1} \sum_{O_1=1}^{t-1} \left[ \mathbb{P} \left\{ \max_{i \in C \setminus 1} \hat{\mu}_{i,O_C} \geq \mu_c^{\text{max}} + \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}} \right\} + \mathbb{P} \left\{ \hat{\mu}_{1,O_1} \leq \mu_1 - \sqrt{\frac{2 \ln(|N_1|^\frac{1}{2} \cdot t)}{O_1}} \right\} \right].
\]

(11)

By applying Hoeffding’s inequality, we have

\[
\mathbb{P} \left\{ \hat{\mu}_{1,O_1} \leq \mu_1 - \sqrt{\frac{2 \ln(|N_1|^\frac{1}{2} \cdot t)}{O_1}} \right\} \leq \frac{1}{|N_1|} \cdot \frac{1}{|N_1|} \leq \frac{1}{n},
\]

(12)

and

\[
\sum_{i \in C \setminus 1} \mathbb{P} \left\{ \hat{\mu}_{i,O_C} \geq \mu_c^{\text{max}} + \sqrt{\frac{2 \ln(N_C \cdot t)}{O_C}} \right\} \leq \frac{1}{|N_1|} \cdot \frac{1}{|N_1|} \leq \frac{1}{n}.
\]

(13)

By plugging (12) and (13) into (11) we have that \((\alpha) \leq \frac{2}{n}. Then by plugging the upper bound of term \((\alpha)\) and \(L_C\) into (6) we have

\[
\mathbb{E}[R_C(T)] \leq \frac{8 \ln(N_C \cdot T)}{(\Delta_c^{\text{max}})^2} \Delta_c^{\text{max}} + \left(1 + \frac{\pi^2}{3}\right) \Delta_c^{\text{max}},
\]

(14)

where \(N_C = \max_{i \in C} \left\{ |N_i|^\frac{1}{2} \right\}\).

Then the regret of UCB-NE is at most

\[
R(T) \leq \inf_C \left\{ \sum_{i \in C, i \neq \{1\}} \frac{8 \Delta_i^{\text{max}} \ln(N_C \cdot T)}{(\Delta_i^{\text{max}})^2} + \left(1 + \frac{\pi^2}{3}\right) \Delta_c^{\text{max}} \right\},
\]

where \(N_C = \max_{i \in C} \left\{ |N_i|^\frac{1}{2} \right\}\). \(\square\)
A.2 Proof of Theorem 2

Proof of Lemma 1. First we construct a monotonic function $g(a) = d(a, \mu_1) - 2d(a, \mu_c^{\text{max}})$ where $a \in [\mu_c^{\text{max}}, \mu_1]$. Now we claim that $g(a)$ is a strictly decreasing function when $a \in [\mu_c^{\text{max}}, \mu_1]$. It is trivial to prove this claim as $g'(a) = \ln \left( \frac{(\mu_c^{\text{max}})^2}{d(a, \mu_1)} \cdot \frac{(1-a)(1-\mu_1)}{(1-d(a, \mu_c^{\text{max}}))^2} \right) < 0$ when $a \in [\mu_c^{\text{max}}, \mu_1]$. Also, we know that $g(\mu_c^{\text{max}}) = d(\mu_c^{\text{max}}, \mu_1) > 0$ and $g(\mu_1) = -2d(\mu_1, \mu_c^{\text{max}}) < 0$. There thus exists a unique $m_c \in (\mu_c^{\text{max}}, \mu_1)$ such that $g(m_c) = 0$. Therefore, we have $g(a) \geq 0$ when $a \in (\mu_c^{\text{max}}, m_c]$ while $g(a) < 0$ when $a \in (m_c, \mu_1]$.

Now, we choose $x_C$ such that $x_C \in (\mu_c^{\text{max}}, m_c]$ and $d(x_C, \mu_1) > \frac{1}{2}d(\mu_c^{\text{max}}, \mu_1)$ hold simultaneously. The guarantee that $x_C \in (\mu_c^{\text{max}}, m_c]$ and $d(x_C, \mu_1) > \frac{1}{2}d(\mu_c^{\text{max}}, \mu_1)$ hold simultaneously implies $\epsilon$ can be any value in $(0, \min \{ \frac{d(\mu_c^{\text{max}}, \mu_1)}{d(m_c, \mu_1)} - 1, 1 \})$. As $x_C \in (\mu_c^{\text{max}}, m_c]$, it means $g(x_C) = d(x_C, \mu_1) - 2d(x_C, \mu_c^{\text{max}}) \geq g(m_c) = 0$, which guarantees $d(x_C, \mu_1) > 2d(x_C, \mu_c^{\text{max}})$. Then we can always find $y_C \in (x_C, \mu_1)$ such that $d(x_C, y_C) = \frac{d(x_C, \mu_1)}{1+\epsilon}$. After fixing $x_C$ and $y_C$, it is trivial to prove condition (iv) since $d(x_C, y_C) - d(x_C, \mu_c^{\text{max}}) = \frac{d(x_C, \mu_1)}{1+\epsilon} - d(x_C, \mu_c^{\text{max}}) \geq \frac{d(x_C, \mu_c^{\text{max}})}{1+\epsilon} - d(x_C, \mu_c^{\text{max}}) \geq 0$.

As there is no closed-form expression of $m_c$, we give a lower bound of $m_c$. From $g'(a) = \ln \left( \frac{(\mu_c^{\text{max}})^2}{d(a, \mu_1)} \cdot \frac{(1-a)(1-\mu_1)}{(1-d(a, \mu_c^{\text{max}}))^2} \right)$ we know $g''(a) < 0$. Then from the concavity of $g(a)$ we have $m_c \geq \mu_c^{\text{max}} + \frac{d(\mu_c^{\text{max}}, \mu_1)}{2d(\mu_1, \mu_c^{\text{max}}) + d(\mu_c^{\text{max}}, \mu_1)}$.

To analyze the first term in (4), i.e., term $\Psi_1$, we prepare Lemma 3 and Lemma 4. Lemma 3 claims that after an 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 $x_C$. Lemma 4 uses a union bound over all the arms in clique $C$ based on Lemma 3.

Lemma 3. For $i \in C \setminus 1$ we have

\[
\sum_{t=1}^{T} \mathbb{P} \{ \hat{\mu}_i(t) > x_C, O_i(t+1) > O_i(t), O_i(t) \geq L_C \} \leq \frac{1}{N_C} \cdot \frac{1}{d(x_C, \mu_c^{\text{max}})}.
\]

Proof of Lemma 3. The proof uses the fact that in each round, we can get at most one observation for any arm. Let $\tau_k$ denote the time stamp when the $k$-th observation of arm $i$ happens. Note that $\{ O_i(t+1) > O_i(t) \}$ cannot happen during the rounds when $t \in \{ \tau_k + 1, \cdots, \tau_{k+1} - 1 \}$ since no new update can be conducted at the end of these rounds. Then we have

\[
\sum_{t=1}^{T} \mathbb{P} \{ \hat{\mu}_i(t) > x_C, O_i(t+1) > O_i(t), O_i(t) \geq L_C \} \leq \mathbb{E} \left[ \sum_{k=L_C}^{\tau_{k+1}-1} \sum_{t=\tau_k}^{\tau_{k+1}-1} \mathbb{1} \{ \hat{\mu}_i(t) > x_C, O_i(t+1) > O_i(t) \} \right]
\]

\[
= \mathbb{E} \left[ \sum_{k=L_C}^{\tau_{k+1}-1} \sum_{t=\tau_k}^{\tau_{k+1}-1} \mathbb{1} \{ \hat{\mu}_i(t) > x_C \} \cdot \mathbb{1} \{ O_i(t+1) > O_i(t) \} \right]
\]

\[
= \mathbb{E} \left[ \sum_{k=L_C}^{\tau_{k+1}-1} \mathbb{1} \{ \hat{\mu}_i(\tau_k) > x_C \} \right].
\]

The first inequality in (15) uses the fact the number of observations starts from at least $L_C$ and increments to at most $T$. All the $T$ rounds are segmented into multiple intervals and in each interval, only one observation is obtained
except for the last time interval during which zero observation may be obtained. The last equality uses the fact that $1\{O_i(t+1) > O_i(t)\} = 0$ when $t \in \{\tau_k + 1, \cdots, \tau_{k+1} - 1\}$ and $1\{O_i(t+1) > O_i(t)\} = 1$ only when $t = \tau_k$.

Then we can use the definition of $\hat{\mu}_i(\tau_k)$ and further have

$$
\mathbb{E}
\left[
\sum_{k=L_C}^{T} 1\{\hat{\mu}_i(\tau_k) > x_C\}
\right]
= \sum_{k=L_C}^{T} \mathbb{P}\{\hat{\mu}_i(\tau_k) > x_C\}
= \sum_{k=L_C}^{T} \mathbb{P}\{\frac{Q_i(\tau_k)}{x_C} > x_C\}
< \sum_{k=L_C}^{T} \mathbb{P}\{\frac{Q_i(\tau_k)}{x_C} > x_C\}
\leq \sum_{k=L_C}^{T} e^{-(k-1)\cdot d(x_C, \mu)}
< e^{-\frac{(t_C-2)\cdot d(x_C, \mu)}{d(x_C, \hat{\mu}_i)}}
\leq e^{-\ln((N_C)^{\eta_C} \cdot T) \cdot \frac{d(x_C, \mu_{\text{max}})}{d(x_C, \hat{\mu})}} \cdot \frac{1}{d(x_C, \mu_{\text{max}})}
= \left(\frac{1}{N_C}\right)^{\eta_C} \cdot \frac{1}{d(x_C, \mu_{\text{max}})} \cdot \frac{1}{d(x_C, \mu_{\text{max}})}
= \frac{1}{N_C} \cdot \frac{1}{d(x_C, \mu_{\text{max}})} \cdot \frac{1}{d(x_C, \mu_{\text{max}})}
\leq \frac{1}{N_C} \cdot \frac{1}{d(x_C, \mu_{\text{max}})} \cdot \frac{1}{d(x_C, \mu_{\text{max}})}
\leq \frac{1}{N_C} \cdot \frac{1}{d(x_C, \mu_{\text{max}})} \cdot .
$$

(16)

The first inequality in (16) uses the definition of $\hat{\mu}_i(\tau_k)$, the empirical mean of $k-1$ observations. Note that although $\tau_k$ is the time stamp when the $k$-th observation happens, $O_i(t)$ and $Q_i(t)$ will only be updated at the end of round $\tau_k$. This is why we only have $k-1$ observations at round $\tau_k$. The third inequality uses the Chernoff-Hoeffding bound (Fact 1 in (Agrawal and Goyal, 2017)). Note that this lemma does not need to use $\eta_C \geq 1$ during the proof. □
Lemma 4. For any $C$ we have

$$E\left[\sum_{t=1}^{T} \sum_{i \in C} 1\{I_t = i, E_C^t(t), T_C(t) \geq L_C\}\right] \cdot \Delta_i \leq \frac{\Delta_{\max}^{C}}{d(x_C, \mu_{\max}^{C})} \cdot \Delta_{\max}^{C}$$

Proof of Lemma 4. For clique $C$ we have

$$E\left[\sum_{t=1}^{T} \sum_{i \in C} 1\{I_t = i, E_C^t(t), T_C(t) \geq L_C\}\right] \cdot \Delta_i$$

$$= E\left[\sum_{t=1}^{T} \sum_{i \in C \setminus 1} 1\{I_t = i, E_C^t(t), T_C(t) \geq L_C\}\right] \cdot \Delta_{\max}^{C}$$

(Use the definition of $E_C^t(t) = \left\{\max_{j \in C \setminus 1} \hat{\mu}_j(t) \leq x_C\right\}$ and then the union bound)

$$\leq \sum_{j \in C \setminus 1} E\left[\sum_{t=1}^{T} 1\{\exists i \in C \setminus 1 s. t. I_t = i, \hat{\mu}_j(t) > x_C, T_C(t) \geq L_C\}\right] \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C}$$

(Pulling any arm in clique makes arm $j$ observed)

$$\leq \sum_{j \in C \setminus 1} \sum_{t=1}^{T} P\{\exists i \in C \setminus 1, I_t = i, \hat{\mu}_j(t) > x_C, T_C(t) \geq L_C\} \cdot \Delta_{\max}^{C}$$

\[\text{Lemma 3}\]

$$\leq \sum_{j \in C \setminus 1} \frac{1}{N_C} \cdot \frac{1}{d(x_C, \mu_{\max}^{C})} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C}$$

$$\leq \frac{\Delta_{\max}^{C}}{d(x_C, \mu_{\max}^{C})} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C} \cdot \Delta_{\max}^{C}$$
To analyze the second term in (4), i.e., term $\Psi_2$, we prepare Lemma 5 and Lemma 6. Lemma 5 claims that after an 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 6 uses a union bound over all the arms within clique $C$ based on Lemma 5.

**Lemma 5.** For $i \in C \setminus 1$ we have

$$
\sum_{t=1}^{T} P \{ \hat{\mu}_i(t) \leq x_C, \theta_i(t) > y_C, O_i(t) \geq L_C \} \leq \frac{1}{N_C}.
$$

**Proof of Lemma 5.** For any sub-optimal arm $i \in C$ we have

$$
\sum_{t=1}^{T} P \{ \hat{\mu}_i(t) \leq x_C, \theta_i(t) > y_C, O_i(t) \geq L_C \}
\leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ \hat{\mu}_i(t) \leq x_C, \theta_i(t) > y_C, O_i(t) \geq L_C \} \right]
\leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{E} \left[ \mathbb{1} \{ \hat{\mu}_i(t) \leq x_C, \theta_i(t) > y_C, O_i(t) \geq L_C \} | \mathcal{F}_{t-1} \right] \right]
\leq \mathbb{E} \left[ \sum_{t=1}^{T} \mathbb{1} \{ \hat{\mu}_i(t) \leq x_C, O_i(t) \geq L_C | \mathcal{F}_{t-1} \} \cdot P \{ \theta_i(t) > y_C | \mathcal{F}_{t-1} \} \right].
$$

Let $F_{\alpha, \beta}^B (\cdot)$ denote the CDF of Beta($\alpha, \beta$) and $F_{n,p}^B (\cdot)$ denote the CDF of binomial distribution with parameter $n, p$. We categorize all the instantiations of $\mathcal{F}_{t-1}$ into two types based on whether a specific instantiation $F_{t-1}$ can make the indicator function $\omega(t)$ return 1 or not. Let $\gamma(t) := \omega(t) \cdot v(t)$. In each round $t$, for the instantiation $F_{t-1}$ of $\mathcal{F}_{t-1}$ that makes $\omega(t) = 0$, we have $\gamma(t) = 0$, while for the instantiation $F_{t-1}$ of $\mathcal{F}_{t-1}$ that makes $\omega(t) = 1$, i.e., both events $\hat{\mu}_i(t) \leq x_C$ and $O_i(t) \geq L_C$ are true, we only need to analyze $v(t) = P \{ \theta_i(t) > y_C | \mathcal{F}_{t-1} = F_{t-1} \}$. Note that $\theta_i(t)$ is sampled from Beta($Q_i(t) + 1, O_i(t) - Q_i(t) + 1$). Then we have

$$
P \{ \theta_i(t) > y_C | \mathcal{F}_{t-1} = F_{t-1} \}
= 1 - F_{Q_i(t) + 1, O_i(t) - Q_i(t) + 1}^B (y_C)
= 1 - F_{\hat{\mu}_i(t)|O_i(t) + 1, 1 - \hat{\mu}_i(t)|O_i(t) + 1}^B (y_C)
\leq 1 - F_{x_C(O_i(t) + 1), 1 - x_C(O_i(t) + 1)}^B (y_C)
= F_{\frac{B}{O_i(t) + 1}, y_C}^B (x_C(O_i(t) + 1))
\leq e^{-O_i(t) + 1/d(x_C, y_C)} \leq e^{-L_C \cdot d(x_C, y_C)}
= e^{-\frac{\ln(N_C)}{d(x_C, y_C)} \cdot d(x_C, y_C)}
= \frac{1}{N_C^{\frac{1}{d(x_C, y_C)}}}
\leq \frac{1}{N_C^T}.
$$

The third equality uses $F_{\alpha, \beta}^B (y) = 1 - F_{\alpha + 1, \beta - 1}^B (y - 1)$ (Fact 3 in (Agrawal and Goyal, 2017)) and the inequality followed by this equality uses the Chernoff-Hoeffding bound again. The last inequality uses $\eta_C \geq 1$. Note that without
the condition \( \eta_C \geq 1 \), it is not easy to make the clique size scale logarithmically with the clique size. Applying \( v(t) = \mathbb{P}\{\theta_i(t) > y_C|\mathcal{F}_{t-1} = F_{t-1}\} \leq \frac{1}{N_C t} \) to (18) concludes the proof.

**Lemma 6.** For any \( C \) we have

\[
\mathbb{E}\left[ \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, E^\theta_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta_i \leq \Delta^\max_C.
\]

**Proof of Lemma 6.** For clique \( C \) we have

\[
\mathbb{E}\left[ \sum_{t=1}^{T} \sum_{i \in C} 1 \{ I_t = i, E^\theta_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta_i
\]

= \[
\mathbb{E}\left[ \sum_{t=1}^{T} \sum_{i \in C \setminus \{i\}} 1 \{ I_t = i, E^\theta_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta_i
\]

\leq \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ \exists i \in C \setminus \{1\} \text{ s.t. } I_t = i, E^\theta_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

(Remove \{ \exists i \in C \setminus \{1\} \text{ such that } I_t = i \} from indicator function)

\leq \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ E^\theta_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

= \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ E^\theta_C(t), \max_{j \in C \setminus \{1\}} \theta_j(t) > y_C, T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

(Use the definition \( E^\theta_C(t) = \left\{ \max_{j \in C \setminus \{1\}} \theta_j(t) \leq y_C \right\} \) and then the union bound)

\leq \sum_{j \in C \setminus \{1\}} \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ E^\theta_C(t), \theta_j(t) > y_C, T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

(Use the definition \( E_C(t) := \left\{ \max_{k \in C \setminus \{1\}} \mu_k(t) \leq x_C \right\} \))

= \sum_{j \in C \setminus \{1\}} \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ \max_{k \in C \setminus \{1\}} \mu_k(t) \leq x_C, \theta_j(t) > y_C, T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

\leq \sum_{j \in C \setminus \{1\}} \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ \hat{\mu}_j(t) \leq x_C, \theta_j(t) > y_C, T_C(t) \geq L_C \} \right] \cdot \Delta^\max_C

\leq \sum_{j \in C \setminus \{1\}} \mathbb{E}\left[ \sum_{t=1}^{T} 1 \{ \hat{\mu}_j(t) \leq x_C, \theta_j(t) > y_C, \hat{O}_j(t) \geq L_C \} \right] \cdot \Delta^\max_C

= \sum_{j \in C \setminus \{1\}} \sum_{t=1}^{T} \mathbb{P}\{ \hat{\mu}_j(t) \leq x_C, \theta_j(t) > y_C, \hat{O}_j(t) \geq L_C \} \cdot \Delta^\max_C

\leq \frac{|C|}{N_C} \cdot \Delta^\max_C

\leq \Delta^\max_C.

\]

To analyze the third term in (4), i.e., term \( \Psi_3 \), we prepare Lemma 7 and Lemma 8. The key techniques in these two lemmas use the ideas in (Agrawal and Goyal, 2017) with slight modifications.
For $C \neq \{1\}$, define $p_{c,t} := P\{\theta_1(t) > y_C | F_{t-1}\}$ and recall that $\Delta'_C = \mu_1 - y_C$ and $D_C = d(y_C, \mu_1)$.

**Lemma 7.** For $i \in C \backslash 1$ we have

$$P\{\exists i \in C \backslash 1 \text{ s.t. } I_t = i, E_C^\mu(t), E_C^\theta(t) | F_{t-1}\} \leq \frac{1 - p_{c,t}}{p_{c,t}} \cdot P\{I_t = 1, E_C^\mu(t), E_C^\theta(t) | F_{t-1}\}.$$  

**Proof of Lemma 7.** Recall that $p_{c,t} = P\{i, E_C^\mu(t), E_C^\theta(t) | F_{t-1}\}$. The proof uses a similar idea as when proving Lemma 2.8 in (Agrawal and Goyal, 2017). The key idea behind the proof is to exploit the feature that $\theta(t)$ for all $i \in N$ are generated independently in each round $t$ and, given $F_{t-1}$, the distribution that generates $\theta(t)$ is determined. Recall that whether $E_C^\mu(t)$ is true or not is determined by an instantiation $F_{t-1}$ of $F_{t-1}$. If the instantiation $F_{t-1}$ is the one that makes event $E_C^\mu(t)$ false, it is trivial to prove since both sides in Lemma 7 are 0. If the instantiation $F_{t-1}$ is the one that makes event $E_C^\mu(t)$ true, then it suffices to prove that for all such instantiations $F_{t-1}$ we have

$$P\{\exists i \in C \backslash 1 \text{ s.t. } I_t = i, E_C^\mu(t), F_{t-1} = F_{t-1}\} \leq \frac{1 - p_{c,t}}{p_{c,t}} \cdot P\{I_t = 1, E_C^\mu(t), F_{t-1} = F_{t-1}\}.$$  

For clique $C$, recall that $E_C^\theta(t) = \left\{ \max_{i \in C \backslash 1} \theta_i(t) \leq y_C \right\}$.

Now, we analyze term $\omega$ in (21) and have

$$\omega = P\{\exists i \in C \backslash 1 \text{ s.t. } I_t = i, E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$\leq P\{\theta_1(t) \leq y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$= P\{\theta_1(t) \leq y_C | E_C^\mu(t), F_{t-1} = F_{t-1}\} \cdot P\{\theta_1(t) \leq y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$= (1 - p_{c,t}) \cdot \underbrace{P\{\theta_1(t) \leq y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}}_{\beta}.$$  

Now, we analyze term $\gamma$ in (21) and have

$$\gamma = P\{I_t = 1 | E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$\geq P\{\theta_1(t) > y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$= P\{\theta_1(t) > y_C, E_C^\mu(t), F_{t-1} = F_{t-1}\} \cdot P\{\theta_1(t) \leq y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}$$

$$= p_{c,t} \cdot \underbrace{P\{\theta_1(t) \leq y_C, \forall j \in N \setminus E_C^\mu(t), F_{t-1} = F_{t-1}\}}_{\beta}.$$  

Then we have

$$P\{\exists i \in C \backslash 1 \text{ s.t. } I_t = i, E_C^\mu(t), F_{t-1} = F_{t-1}\} \cdot \frac{1}{1 - p_{c,t}} \leq \beta \leq P\{I_t = 1 | E_C^\mu(t), F_{t-1} = F_{t-1}\} \cdot \frac{1}{1 - p_{c,t}},$$

which concludes the proof.

**Lemma 8.** For any $C$ we have

$$E\{\sum_{t=1}^{T} \sum_{i \in C} I_i(t), E_C^\mu(t), T_C(t) \geq L_C\} \cdot \Delta_t \leq \frac{24\Delta_{\max}}{\Delta_C^2} + O \left( \frac{\Delta_{\max}}{\Delta_C} + \frac{\Delta_{\max}}{\Delta_C D_C} + \frac{\Delta_{\max}}{\Delta_C^2} \right).$$
Proof of Lemma 8. The proof uses the ideas that pulling arm 1 means it must be observed. Also, in each round \( t \), the learner can get at most one observation of arm 1. Let \( \tau_k \) be the time stamp where arm 1 gets the \( k \)-th observation and set \( \tau_0 = 0 \). Note that \( \theta_{c,t} \) cannot change during the rounds when \( t \in \{ \tau_k + 1, \cdots, \tau_{k+1} \} \) since the beta distribution that generates \( \theta_1(t) \) does not change.

For clique \( C \) we have

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \sum_{i \in C} \mathbb{1}\{I_t = i, E^u_C(t), E^g_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta_i
\]

\[
= \mathbb{E} \left[ \sum_{t=1}^{T} \sum_{i \in C} \mathbb{1}\{I_t = i, E^u_C(t), E^g_C(t), T_C(t) \geq L_C \} \right] \cdot \Delta_i
\]

(Remove \{ \( T_C(t) \geq L_C \) \} from the indicator function)

\[
\leq \mathbb{E} \left[ \sum_{i \in C} \mathbb{1}\{I_t = i, E^u_C(t), E^g_C(t) \} \right] \cdot \Delta_i
\]

\[
= \sum_{i \in C} \mathbb{P}\{I_t = i, E^u_C(t), E^g_C(t) \} \cdot \Delta_i
\]

\[
= \sum_{i \in C} \mathbb{E} \left[ \mathbb{P}\{I_t = i, E^u_C(t), E^g_C(t) | F_{t-1} \} \right] \cdot \Delta_i
\]

(By using Lemma 7 we can get the following)

\[
\leq \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1 - P_{c,t}}{P_{c,t}} \mathbb{1}\{I_t = 1, E^u_C(t), E^g_C(t) | F_{t-1} \} \right] \cdot \Delta_i
\]

\[
= \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1 - P_{c,t}}{P_{c,t}} \mathbb{1}\{I_t = 1, E^u_C(t), E^g_C(t) \} \right] \cdot \Delta_i
\]

\[
\leq \sum_{t=1}^{T} \mathbb{E} \left[ \frac{1 - P_{c,t}}{P_{c,t}} \mathbb{1}\{O_1(t) > O_1(t), E^u_C(t), E^g_C(t) \} \right] \cdot \Delta_i
\]

(Use the fact that \( \{O_1(t+1) > O_1(t)\} \) cannot be true when \( t \in \{ \tau_k + 1, \cdots, \tau_{k+1} + 1 \} \))

\[
\leq \sum_{t=0}^{T} \mathbb{E} \left[ \frac{1 - P_{c,\tau_{k+1}}}{P_{c,\tau_{k+1}}} \right] \cdot \Delta_i
\]

(Use the fact that \( p_{c,t} \) does not change when \( t \in \{ \tau_k + 1, \cdots, \tau_{k+1} \} \))

\[
= \sum_{t=0}^{T} \mathbb{E} \left[ \frac{1 - p_{c,\tau_{k+1}}}{p_{c,\tau_{k+1}}} \right] \cdot \Delta_i
\]

(Slightly modify Lemma 2.9 in (Agrawal and Goyal, 2017) we can get)

\[
\leq \frac{3 \Delta_{C}^{\max}}{\Delta_{C}^{2}} + \sum_{k = \tau_0}^{T} \frac{O}{k} \left( \frac{\Delta_{C}^{\max}}{e^{\Delta_{C}^{2}/2}} + \frac{\Delta_{C}^{\max}}{(k+1)\Delta_{C}^{2}e^{\Delta_{C}^{2}/4}} + \frac{\Delta_{C}^{\max}}{\Delta_{C}^{4}} \right)
\]

\[
\leq \frac{24 \Delta_{C}^{\max}}{\Delta_{C}^{2}} + O \left( \frac{\Delta_{C}^{\max}}{\Delta_{C}^{2}} + \frac{\Delta_{C}^{\max}}{\Delta_{C}^{2}D_{C}} + \frac{\Delta_{C}^{\max}}{\Delta_{C}^{4}} \right)
\]

The modification is only at the beginning of the proof of Lemma 2.9 in (Agrawal and Goyal, 2017). More specifically, we only need to modify the following: Let \( O_j(t) = j \) and \( Q_j(t) = s \). Let \( y = y_C \). Then we use \( p_{c,t} := \mathbb{P}\{\theta_1(t) >
\{y_{t} \mid \mathcal{F}_{t-1}\} instead of their $p_{t,j}$ during the proof. Another modification is to let $\tau_j + 1$ be the time stamp after the $j$-th observation of arm 1 instead of the time stamp after the $j$-th pull of arm 1.

For the second term in Big-O notation in the last inequality, in (Agrawal and Goyal, 2017), it is $\Theta \left( \frac{\Delta_{\max}}{\Delta_{C}^{2}D_{C}} \right)$ originally but it can be improved to $O \left( \frac{\Delta_{\max}}{\Delta_{C}^{2}D_{C}} \right)$. Also, an alternative way is to pay $O \left( \frac{\Delta_{C} \ln(T\Delta_{C}^{'})}{(\Delta_{C}^{'})^{2}} \right)$. For the last term in Big-O notation, instead of paying $O \left( \frac{\Delta_{\max}}{\Delta_{C}^{2}} \right)$, an alternative way is to pay $O \left( \frac{\Delta_{C} \ln(T\Delta_{C}^{'})}{(\Delta_{C}^{'})^{2}} \right)$. □
B UCB-MaxN

In this section, we provide more details about the issues with the proof of the regret bound for UCB-MaxN (Caron et al., 2012). In the analysis of Theorem 3 of Caron et al. (2012) (the regret bound for UCB-MaxN), one of the steps of the proof appears to be problematic. Specifically, there seems to be an issue with the second inequality below inequality (3) (the first inequality just after “The first summation can be bounded using the Chernoff-Hoeffding inequality as before:”). In our understanding, pulling arm $k_C$, the best sub-optimal arm within clique $C$, does not mean its upper confidence bound must be greater than that of the globally best arm. An example is that arm $k_C$ may have a neighboring arm $j$, not belonging to clique C, which has the highest upper confidence bound while, simultaneously, arm $k_C$ has the highest empirical mean among all the neighbors of arm $j$. In this example, arm $k_C$ is pulled but its upper confidence bound is not necessarily greater than or equal to that of the globally best arm. It is important to note that arm $k_C$ might be collecting observations from pulls of its neighbors that are not neighbors of $j$. Therefore, it is possible that the upper confidence bound of arm $k_C$ is no greater than that of arm $j$ while simultaneously, the empirical mean of arm $k_C$ is no smaller than that of arm $j$. 