
Graph-Dependent Regret Bounds in Multi-Armed Bandits with Interference

Fateme Jamshidi^{*1}

Mohammad Shahverdikondori^{*1}

Negar Kiyavash¹

¹College of Management of Technology, EPFL, Lausanne, Switzerland

Abstract

We study multi-armed bandits under network interference, where each unit’s reward depends on its own treatment and those of its neighbors in a given graph. This induces an exponentially large action space, making standard approaches computationally impractical. We propose a novel algorithm that uses the local graph structure to minimize regret. We derive a graph-dependent upper bound on cumulative regret that improves over prior work. Additionally, we provide the first lower bounds for bandits with arbitrary network interference, where each bound involves a distinct structural property of the graph. These bounds show that for both dense and sparse graphs, our algorithm is nearly optimal, with matching upper and lower bounds up to logarithmic factors. When the interference graph is unknown, a variant of our algorithm is Pareto optimal: no algorithm can uniformly outperform it across all instances. We complement our theoretical results with numerical experiments, showing that our approach outperforms the baseline methods.

1 INTRODUCTION

Multi-armed bandits (MABs) have become a fundamental tool for online decision-making across a variety of applications [Bouneffouf et al., 2020, Tewari and Murphy, 2017, Lattimore and Szepesvári, 2020, Bubeck et al., 2009]. Although classic MABs perform well in certain settings, they are prone to systematic errors when *interference* is present. Interference arises when a unit’s outcome depends not only on its own treatment but also on the treatment of others. For instance, the vaccination status of an individual can affect whether others fall ill, or a personal advertisement may in-

fluence friends of the targeted user. Such interference poses challenges to traditional experimental design and sequential learning frameworks, which often assume independent unit responses; formally, the *Stable Unit Treatment Value Assumption* (SUTVA) [Rubin, 1980]. Violations of SUTVA with cross-unit dependencies complicates the applicability of traditional methods, leading to suboptimal decisions.

In recent years, certain methods have been developed that account for interference in settings where only the quality of the final output is of interest [e.g., Ugander et al., 2013, Athey et al., 2018]. The cumulative performance of the experimentation, which is particularly relevant in most online settings, has received relatively less attention, in part because it is more difficult to analyze. We aim to advance this line of research by studying the MAB problem in the presence of arbitrary network interference. Specifically, we consider N units representing entities such as users on an online platform or a medical trial. There are k treatment arms (or arms) and a time horizon of T rounds. At each round, the learner assigns one arm to each unit and observes the resulting reward, with the goal of maximizing the total accumulated reward over all rounds. Unlike the traditional MAB framework, where rewards are assumed to be independent, interference introduces a dependency structure, with the reward of each unit determined not only by its own arm but also by the arms assigned to others. This dependency can significantly increase the complexity of exploration, as the action space may include up to k^N different actions, growing exponentially in the number of units.

Classical MAB algorithms, such as the Upper Confidence Bound (UCB) method [Auer, 2002], yield a regret scaling of $\tilde{O}(\sqrt{\frac{T}{N}} k^N)^1$, which becomes prohibitive as N grows. Furthermore, without imposing any assumptions on the interference structure, the regret is guaranteed to scale as $\Omega(\sqrt{\frac{T}{N}} k^N)$, as shown by lower bounds in the MAB litera-

^{*}Authors contributed equally.

¹Assuming each unit’s reward is 1-sub-Gaussian, the average reward over N units follows a $(1/\sqrt{N})$ -sub-Gaussian distribution.

ture [Lattimore and Szepesvári, 2020].

To address this challenge, we consider a setting where the reward of any given unit is influenced by its own arm and the arms assigned to its neighboring units, as defined by a graph that represents the interferences among the units. This graph is referred to as the *interference graph*.

Contributions Our main contributions are as follows.

- We introduce the Partitioned Upper Confidence Bound with Interference (PUCB-I) algorithm for the MAB problem in the presence of network interference. Using the local structure of the interference graph, we establish a graph-dependent upper bound on cumulative regret. The algorithm outperforms state-of-the-art methods in all of the settings.
- We derive the first lower bounds on the regret for bandits with arbitrary network interference. We provide two distinct lower bounds that quantify the worst-case regret of any algorithm based on the topological properties of the underlying interference graph. Specifically, these properties pertain to i) the symmetries in neighborhoods of the nodes (representing the units) and ii) the structure of doubly-independent sets of the interference graph. We demonstrate that the upper bound of our algorithm matches these lower bounds up to logarithmic factors for both dense and sparse graphs, thereby establishing its near-optimality. Furthermore, in general, the gap between the upper and lower bounds is at most $\sqrt{N}$.
- We extend our analysis to the setting where the interference graph is unknown. We introduce a variant of our algorithm, PUCB-UI, which assumes the true interference graph is complete and runs PUCB-I accordingly. We show that PUCB-UI is in the Pareto optimal set of algorithms, meaning that no algorithm can uniformly outperform it across all instances. This result establishes a fundamental limit on learning under an unknown interference structure.
- Through simulations, we show that our algorithm outperforms baseline methods.

Paper Outline We begin with related work, followed by background and key definitions in Section 2. Section 3 presents our algorithm and its regret upper bound, and Section 4 provides matching lower bounds. Section 5 addresses the setting with an unknown interference graph. Simulations appear in Section 6 and all the proofs are included in Appendix A.

1.1 RELATED WORK

Cross-unit interference has been studied in statistics [Hudgens and Halloran, 2008, Eckles et al., 2017, Basse and Feller, 2018, Li and Wager, 2022, Leung, 2023], computer

science [Ugander et al., 2013, Yuan et al., 2021, Ugander and Yin, 2023] and medical research [Tchetgen and VanderWeele, 2012]. To address the challenge of interfering units, researchers have proposed tailor-made methodologies in structured interference models of their specific settings, such as intragroup interference [Rosenbaum, 2007, Hudgens and Halloran, 2008], network neighborhoods [Ugander et al., 2013, Bhattacharya et al., 2020, Yu et al., 2022, Gao and Ding, 2023], bipartite graphs [Pouget-Abadie et al., 2019, Bajari et al., 2021, 2023] and general graph models through exposure mappings [Aronow, 2012, Aronow and Samii, 2017, Adhikari et al., 2025].

Despite this extensive literature, all the aforementioned works have focused merely on strategies that maximize the final reward. The harder problem of understanding cumulative performance over time remains relatively unexplored. The MAB framework is naturally well-suited to address this gap by balancing exploration and exploitation. Studies in structured bandits have shown that incorporating problem structure, such as dependencies among actions or feedback, can lead to more efficient learning [Chen et al., 2014, Lattimore et al., 2016, Alon et al., 2017, Jamshidi et al., 2024, Shahverdikondori et al., 2025, Mahrooghi et al., 2025]. In our case, the structure arises from interference between units. A recent work by Zhang and Wang [2024] studied online experimental design under network interference, focusing on a trade-off between reward and estimation. In contrast, our focus is on minimization of regret in sequential learning under arbitrary interference. Faruk and Zheleva [2025] studied a contextual MAB setting with interference in a social network. In their framework, at each round, a single unit is assigned a treatment, which may activate its neighboring units randomly, with the objective of maximizing the total number of activated nodes over time. The linear contextual bandit setting with network interference is also studied recently by Xu et al. [2024]. Our setting can be formulated as a specific instance of the combinatorial bandit framework [Cesa-Bianchi and Lugosi, 2012, Chen et al., 2013b]. However, we show that no algorithm achieving optimal regret in the general combinatorial setting can be optimal in ours. We show this by deriving an upper bound that improves over the matching upper and lower bounds known for combinatorial bandits (see Remark 3.5). These results underscore the theoretical significance of our setting as a structured subclass that merits separate analysis. Another line of work is multiple-play bandits [Anantharam et al., 1987, Chen et al., 2013a, Lagr e et al., 2016, Jia et al., 2023] that considers settings where the learner selects multiple arms simultaneously and observes feedback for each arm. However, they assume that the rewards are independent across arms, which is not the case in the presence of interference.

Recently, Jia et al. [2024] studied a batched adversarial bandit framework where N units lie on the $\sqrt{N} \times \sqrt{N}$ unit grid. By limiting the action space to those with an identical

arm for all units, they achieved a regret bound that does not depend exponentially on N . However, such an approach is quite limiting in practice, as the optimal action may be one that assigns heterogeneous arms to units.

A common approach to address the curse of dimensionality is to impose sparsity constraints, where only a fraction of actions yield non-zero rewards. Agarwal et al. [2024] studied stochastic MABs under network interference, modeling rewards over the hypercube $[k]^N$ with each unit's reward influenced by its own and its neighbors' arms. Their regression-based algorithm achieves regret $\tilde{O}(T^{2/3})$, which is worse than the classical MAB settings, highlighting the impact of interference. They further proposed a sequential action elimination algorithm with regret scaling as $\sqrt{T}$, albeit growing with N . In this paper, we use the neighborhood structure to establish an upper bound on regret that improves over previous work by a factor between $\sqrt{N}$ and N , depending on the interference graph. In particular, we avoid regrets growing with N , in contrast to previous work. Furthermore, we derive matching lower bounds and prove that our algorithm is near-optimal for both sparse and dense graphs.

2 PROBLEM SETUP

We consider a stochastic multi-armed bandit setting with N units and k treatment arms (or simply, arms). At each round $t \in [T]$, the learner selects an assignment $A_t \in [k]^N$, where A_{ti} denotes the treatment assigned to unit $i \in [N]$. The reward for unit i is given by $Y_{ti} : [k]^N \rightarrow \mathbb{R}$, which depends on both the unit's own arm and those of its neighbors due to interference.

Interference is modeled by a graph $\mathcal{G} := ([N], \mathcal{E})$, where nodes represent units, and an edge $(i, j) \in \mathcal{E}$ indicates that treatments assigned to i and j affect each other's rewards. The neighborhood of unit i is denoted by $N(i)$, which includes i and its neighbors, with size $|N(i)| = d_i + 1$.

At each round t , the learner selects a treatment assignment A_t based on past observations and observes the rewards $Y_{ti}(A_t)$ ² for all $i \in [N]$.

Formally, a policy $\pi := (\pi_1, \dots, \pi_T)$ is a sequence of mappings:

$$\pi_t : ([k]^N \times \mathbb{R}^N)^{t-1} \rightarrow \mathcal{P}([k]^N),$$

where $\mathcal{P}([k]^N)$ denotes the space of probability distributions over treatment assignments and the domain represents the history of past assignments and rewards. The treatment at

round t is drawn as:

$$A_t \sim \pi_t(A_1, Y_1, \dots, A_{t-1}, Y_{t-1}).$$

To evaluate a policy over T rounds, we define cumulative regret as the gap between the optimal reward in hindsight and the expected reward achieved by the policy. The expected reward for unit i under treatment A is given by

$$\mu_i(A) := \mathbb{E}[Y_i | A_{N(i)}],$$

$$\text{with } \mu_i(A) \in [0, 1] \text{ for all } i \in [N], A \in [k]^N.$$

Definition 2.1 (Regret). The regret of policy π on instance $\mathcal{V}$ is defined as:

$$\text{Reg}_T(\pi, \mathcal{V}) := \frac{1}{N} \mathbb{E}_{A_t \sim \pi} \left[\max_{A \in [k]^N} T \sum_{i \in [N]} \mu_i(A) - \sum_{\substack{i \in [N] \\ t \in [T]}} \mu_i(A_t) \right],$$

where the first term is the cumulative reward under an optimal treatment assignment, and the second corresponds to the cumulative reward achieved by the policy π . The instance $\mathcal{V}$ includes the interference graph and the reward distributions. For simplicity, we may omit π or $\mathcal{V}$ when clear from context.

Remark 2.2. If the interference graph $\mathcal{G}$ is not connected, each connected component can be treated as an independent instance. The problem can then be solved separately for each component, and the total regret is the weighted average of regrets over them. The same applies to proving lower bounds. Hence, we assume $\mathcal{G}$ is connected throughout the paper.

3 UPPER BOUND

In this section, we propose the Partitioned UCB with Interference (PUCB-I) algorithm to address the MAB problem under network interference. The algorithm relies on a partitioning of units, based on a structural property of the interference graph, to choose its next treatment assignment.

Specifically, we define a binary relation over pairs of units $i, i' \in [N]$, where $i \sim i'$ if they are connected and have identical neighborhoods in the interference graph $\mathcal{G}$; that is,

$$i \sim i' \iff N(i) = N(i'). \quad (1)$$

This relation is an equivalence, partitioning $[N]$ into M classes $\{P_1, P_2, \dots, P_M\}$. Let $m_j = |P_j|$ denote the size of partition P_j and D_j the common degree of its nodes, since they share the same neighborhood. Figure 1 depicts such a partitioning with $P_1 = \{A, B\}$, $P_2 = \{C\}$, $P_3 = \{D, E\}$, and $P_4 = \{F, G, H\}$. For a graph $\mathcal{G}$, we denote by $M(\mathcal{G})$ the number of partitions induced by the relation in Equation (1) and use M when $\mathcal{G}$ is clear from context.

To explain the key components of the algorithm, we first introduce two essential definitions on graphs.

²We assume that for each treatment A and unit i , the reward distribution $Y_i(A)$ is 1-sub-Gaussian, a standard assumption in the bandit literature [Lattimore and Szepesvári, 2020, Bubeck et al., 2012]. Moreover, conditional on A , the rewards $\{Y_i(A)\}_{i \in [N]}$ are mutually independent.

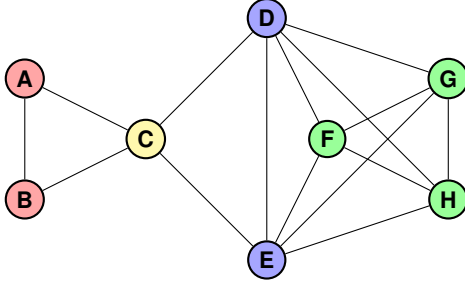


Figure 1: Graph partitions from Eq. (1); same-color nodes share a partition.

Definition 3.1 (Doubly-Independent Set). Let $\mathcal{G}$ be a connected graph with N nodes. A set of nodes S is called a *doubly-independent set* if no two nodes in S are adjacent, nor do they share a common neighbor outside S . The family of all doubly-independent sets of $\mathcal{G}$ is denoted by $DI(\mathcal{G})$.

Definition 3.2 (Square Chromatic Number). The *square chromatic number* of a graph $\mathcal{G}$ is the minimum number of colors required to color its nodes such that all nodes of the same color form a doubly-independent set. The square chromatic number of $\mathcal{G}$ is denoted by $\chi(\mathcal{G}^2)$.

Initialization. In this phase, the algorithm aims to collect at least one sample from each possible treatment $A_{N(i)}$ for every unit i and its neighbors. Exhaustively exploring all k^N joint treatments is highly inefficient, as it results in collecting excessive samples when the degrees are small, leading to high regret during initialization.

To address this, we propose a graph-based initialization strategy that efficiently explores the relevant treatment space in a small number of rounds. First, we partition the graph nodes into sets $S_1, S_2, \dots, S_{\chi(\mathcal{G}^2)}$ according to Definition 3.2, where each S_l is a doubly-independent set (note that these partitions differ from those defined by Relation 1). Let n_l denote the maximum degree of nodes in S_l .

For each S_l , we allocate k^{n_l+1} rounds to ensure that each unit $i \in S_l$ receives at least one sample for every configuration of $A_{N(i)}$. Due to the doubly-independent property, the treatment assignments for one unit and its neighbors do not interfere with the assignments of other units in the same set or their neighbors. This enables us to simultaneously explore all treatment combinations for all units in S_l and their neighborhoods. By devoting k^{n_l+1} rounds to each S_l , we assign treatments in a symmetric manner to ensure that every unit $i \in S_l$ receives exactly $k^{n_l-d_i}$ samples from each possible configuration of $A_{N(i)}$. Let $\Delta := \max_i d_i$ denote the maximum neighborhood size. Since $n_l \leq \Delta$ for all l , the total number of rounds required for initialization is at most:

$$\sum_{l \in [\chi(\mathcal{G}^2)]} k^{n_l+1} \leq \chi(\mathcal{G}^2) \cdot k^{\Delta+1}.$$

This method is significantly more efficient than exhaustive

Algorithm 1 Partitioned UCB with Interference (PUCB-I)

- 1: **Input:** Number of rounds T , interference graph $\mathcal{G}$, number of arms k , confidence parameter δ .
 - 2: **Initialization:** Perform the initialization phase described above and set $T_0 = \sum_{l \in [\chi(\mathcal{G}^2)]} k^{n_l+1}$.
 - 3: **for** each round $t = T_0, \dots, T-1$ **do**
 - 4: **for** each partition $P_j \in \{P_1, \dots, P_M\}$ **do**
 - 5: Compute $UCB_{tP_j}(A)$ for all $A \in [k]^N$ using Eq. (2).
 - 6: **end for**
 - 7: Select A_{t+1} using Eq. (3).
 - 8: Observe rewards $Y_{t+1,i}(A_{t+1})$ for all $i \in [N]$.
 - 9: Update $\hat{\mu}_{t+1,i}(A)$ and $n_{t+1P_j}(A)$ for the explored treatments.
 - 10: **end for**
-

enumeration over all k^N treatments or even taking one round for each treatment configuration per unit and its neighbors, which would require $\sum_{i \in [N]} k^{d_i+1}$ rounds. We will later show that $\chi(\mathcal{G}^2)$ can be upper bounded in terms of Δ ; see Lemma 4.6.

Empirical Mean Reward. For each unit $i \in [N]$ and treatment $A_{N(i)} \in [k]^{d_i+1}$, the empirical mean reward is estimated as:

$$\hat{\mu}_{ti}(A) = \frac{\sum_{t'=1}^t Y_{t'i}(A_{N(i)})}{n_{ti}(A)}$$

$$\text{with } n_{ti}(A) = \sum_{t'=1}^t \mathbb{1}\{A_{t'N(i)} = A_{N(i)}\},$$

where $n_{ti}(A)$ is the number of times $A_{N(i)}$ has been assigned to unit i up to time t . Since $n_{ti}(A) = n_{ti'}(A)$ for all $i, i' \in P_j$, we define $n_{tP_j}(A) := n_{ti}(A)$ for any unit $i \in P_j$.

UCB. For each partition P_j at round t , compute the UCB for every treatment A as:

$$UCB_{tP_j}(A) := \sum_{i \in P_j} \hat{\mu}_{ti}(A) + \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A)}}. \quad (2)$$

Treatment Assignment. At each round, the learner selects the treatment that maximizes the sum of UCBs across all partitions:

$$A_{t+1} = \arg \max_{A \in [k]^N} \sum_{j \in [M]} UCB_{tP_j}(A). \quad (3)$$

Following standard practice in the combinatorial bandit literature [Chen et al., 2013b, Cesa-Bianchi and Lugosi, 2012, Chen et al., 2016, Wang and Chen, 2017, Perrault et al., 2020, Tzeng et al., 2023], we assume access to a computationally efficient offline oracle that, given the current estimates (e.g., mean rewards or UCBs), returns the

treatment assignment maximizing the objective in Equation (3). This abstraction, commonly referred to as a linear maximization (LM) oracle, is standard in the combinatorial bandits. Without such an oracle, even solving the offline optimization problem with known rewards is generally intractable, rendering regret minimization in the online setting computationally infeasible.

Regarding the existence of such an oracle, the recent work Shahverdikondori et al. [2026] studies the computational complexity of this optimization problem. They show that it is NP-hard for general graphs and propose a dynamic programming algorithm that computes the exact optimizer in time $O(k^{tw(\mathcal{G}^2)} N^2)$, where $tw(\mathcal{G}^2)$ denotes the *treewidth* of the squared graph. They further prove that this running time is essentially optimal under standard complexity assumptions. Treewidth is a graph parameter that measures how “tree-like” a graph is; for example, every tree has treewidth 1 (see Cygan et al. [2015]). This result highlights that the optimization is computationally tractable when $tw(\mathcal{G}^2)$ is small. In addition, they show that when rewards are positive, efficient constant-factor approximation algorithms exist for classes of practical graphs, such as those with constant maximum degree Δ .

Reward Observation and Update. After selecting A_{t+1} , the algorithm observes $Y_{t+1i}(A_{t+1})$ for all $i \in [N]$, and updates $\hat{\mu}_{t+1i}(A)$ and $n_{t+1P_j}(A)$ accordingly.

The following theorem establishes a graph-dependent upper bound on the expected cumulative regret of Algorithm 1:

Theorem 3.3. *[Graph-Partitioned Regret Upper Bound] The expected cumulative regret of Algorithm 1 with $\delta = (T^2 N \sum_{j \in [M]} k^{D_j+1})^{-1}$ interacting with any instance with 1-sub-Gaussian rewards and interference graph $\mathcal{G}$, partitioned into $P_1, P_2, \dots, P_M$, satisfies:*

$$Reg_T \in \mathcal{O} \left(\sqrt{\frac{T}{N^2} \log(TN)} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

The dependency on k^{D_j+1} reflects the complexity introduced by interference through each partition’s neighborhood size. A perhaps more interpretable form of the bound can be obtained in terms of the maximum degree $\Delta := \max_{i \in [N]} d_i$:

Corollary 3.4. *Let Δ denote the maximum degree of $\mathcal{G}$. Then,*

$$Reg_T \in \mathcal{O} \left(\sqrt{\frac{TM}{N} k^{\Delta+1} \log(TN)} \right).$$

This follows from Cauchy-Schwarz inequality:

$$\begin{aligned} \sum_{j \in [M]} \sqrt{m_j k^{\Delta+1}} &\leq \sqrt{M \sum_{j \in [M]} m_j k^{\Delta+1}} \\ &= \sqrt{MN k^{\Delta+1}}. \end{aligned}$$

Corollary 3.4 allows comparison to the state-of-the-art. Agarwal et al. [2024] derived a regret upper bound of

$$\tilde{\mathcal{O}} \left(\sqrt{TN k^{\Delta+1}} \right),$$

which is a factor $N/\sqrt{M}$ worse than PUCB-I. At one end, when $M = N$ (i.e., each unit forms its own partition), the gap is $\sqrt{N}$, while at the other end, when $M = 1$, it can be as large as N .

Remark 3.5. Compared to combinatorial bandits, our setting can be regarded as a special case with $\sum_{i \in [N]} k^{d_i+1}$ base arms and k^N super arms where each super arm consists of N arms and represents a joint treatment assignment to all N units. At each round, the learner selects one super arm and observes outcomes for all individual units. In terms of feedback structure, this aligns with the *semi-bandit* model studied in the combinatorial bandit literature, where the learner observes individual rewards for each selected base arm. For this setting, the tight lower bound in the combinatorial bandit literature [Cesa-Bianchi and Lugosi, 2012, Audibert et al., 2014, Merlis and Mannor, 2020] is

$$\Omega \left(\sqrt{\frac{T}{N} \sum_{i \in [N]} k^{d_i+1}} \right),$$

which can be larger than our upper bound in Corollary 3.4. For example, on regular graphs where all nodes have the same degree, the gap is $\sqrt{\frac{N}{M}}$. This gap indicates that algorithms optimal for the general combinatorial bandit problem are not optimal in our setting. As a result, specialized algorithms are required to exploit the structure induced by network interference.

If partitioning is ignored and UCBs are computed per unit, the regret bound becomes

$$\mathcal{O} \left(\sqrt{\frac{T}{N^2} \log(TN)} \sum_{i \in [N]} \sqrt{k^{d_i+1}} \right).$$

This bound is always at least as large as the bound in Theorem 3.3 since

$$\sum_{i \in [N]} \sqrt{k^{d_i+1}} = \sum_{j \in [M]} m_j \sqrt{k^{D_j+1}} \geq \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}}.$$

This indicates that partitioning reduces the total regret from partition P_j by a factor of $\sqrt{m_j}$. For example, in the case of a complete graph, this yields a $\sqrt{N}$ improvement.

4 LOWER BOUNDS

In this section, we establish the first lower bounds on the expected regret for MABs with arbitrary network interference. We derive two distinct bounds that quantify the worst-case regret of any algorithm based on the topological properties of the underlying interference graph. Specifically, these properties pertain to i) the symmetries in neighborhoods of the nodes and ii) the structure of doubly-independent sets of the interference graph. Proofs are provided in Appendix A.2.

Both lower bounds have a gap with the proposed upper bound for PUCB-I in Theorem 3.3. Subsequently, we identify classes of graphs where these gaps are constant, showing that our algorithm is nearly optimal (up to logarithmic factors). The first lower bound demonstrates that PUCB-I is near-optimal for classes of dense graphs, while the second proves that it is near-optimal for sparse graphs. Sparse graphs are particularly significant in multi-armed bandit problems with interference, as many practical applications involve interference graphs that are sparse due to limited local interactions [Agarwal et al., 2024, Yang et al., 2016]. The following theorem establishes our first lower bound on the expected regret.

Theorem 4.1. *[Graph-Partitioned Regret Lower Bound]*

Let $\mathcal{G}$ be a connected graph with N nodes and partitions $\{P_1, P_2, \dots, P_M\}$ with $|P_j| = m_j$. If $k > \frac{2^{\frac{\Delta+1}{2}}}{2^{\frac{\Delta+1}{2}} - 1}$ and $T \geq \frac{4(k-1)^{\Delta+1}}{M}$, then for any policy π , there exists a bandit instance with interference graph $\mathcal{G}$ whose reward is distributed as 1-Gaussian with means in $[0, 1]$ such that

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T}{N^2 M}} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

This lower bound has a gap of $\sqrt{M}$ compared to the upper bound of PUCB-I. This indicates that the algorithm achieves better performance for graphs with a smaller number of partitions.

We now derive a second lower bound on the expected regret that nearly matches the upper bound of our proposed algorithm for graphs with a bounded maximum degree.

Theorem 4.2. *[Doubly-Independent Set Regret Lower Bound]*

Let $\mathcal{G}$ be a connected graph with N nodes. If $T \geq k^{\Delta+1} - 1$, for any policy π , there exists a bandit instance with interference graph $\mathcal{G}$, 1-Gaussian reward distributions and means in $[0, 1]$ such that

$$\text{Reg}_T(\pi) \in \Omega \left(\max_{S \in DI(\mathcal{G})} \sqrt{\frac{T}{N^2}} \sum_{i \in S} \sqrt{k^{d_i+1}} \right), \quad (4)$$

where $DI(\mathcal{G})$ is defined in Definition 3.1.

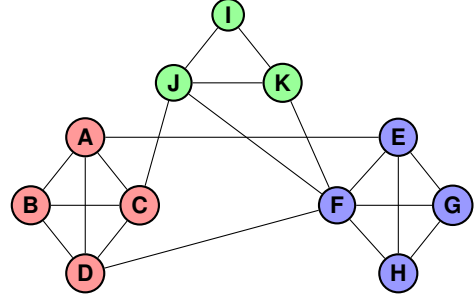


Figure 2: A $(3, 2)$ -clique-sparse graph with colored clusters.

The following corollary relates doubly-independent sets to the square chromatic number $\chi(\mathcal{G}^2)$, allowing comparison between the two lower bounds.

Corollary 4.3. *A connected graph $\mathcal{G}$ can be colored with $\chi(\mathcal{G}^2)$ distinct colors, where each color class is a doubly-independent set. Therefore, the maximum in (4) is at least equal to $\frac{1}{\chi(\mathcal{G}^2)}$ times the sum on all the nodes, implying*

$$\text{Reg}_T(\pi) \in \Omega \left(\frac{1}{\chi(\mathcal{G}^2)} \sqrt{\frac{T}{N^2}} \sum_{i \in [N]} \sqrt{k^{d_i+1}} \right). \quad (5)$$

The first lower bound has a $\sqrt{M(\mathcal{G})}$ gap with the upper bound; the second has a gap of $\chi(\mathcal{G}^2)$. Thus, the tighter bound depends on whether $\chi(\mathcal{G}^2) > \sqrt{M(\mathcal{G})}$ or not. Overall, the gap is at most

$$\min \left(\sqrt{M(\mathcal{G})}, \chi(\mathcal{G}^2) \right)$$

which is always bounded by $\sqrt{N}$. Therefore, for graphs with constant $\sqrt{M(\mathcal{G})}$ or $\chi(\mathcal{G}^2)$, PUCB-I achieves near-optimal regret.

4.1 GRAPHS WITH TIGHT BOUNDS

This section explores the classes of graphs where $\min \left(\sqrt{M(\mathcal{G})}, \chi(\mathcal{G}^2) \right)$ is constant, yielding tight regret bounds.

To characterize the graphs for which the first lower bound shows the near-optimality of our algorithm, i.e., where $\sqrt{M(\mathcal{G})}$ is constant, we define a class of dense graphs, called *Clique-Sparse Graphs*.

Definition 4.4 (Clique-Sparse Graph). A graph $\mathcal{G}$ with N nodes is (R, r) -Clique-Sparse if its nodes can be partitioned into R clusters $C_1, C_2, \dots, C_R$ such that:

- Each cluster C_i forms a complete graph, and
- There are at most r edges between any pair of clusters.

For example, Figure 2 shows a $(3, 2)$ -clique-sparse graph, while a complete graph is $(1, 0)$ -clique-sparse.

The next lemma shows that if $\mathcal{G}$ is (R, r) -clique-sparse with constant R and r , then $M(\mathcal{G})$ is also constant, establishing that PUCB-I is near-optimal for this class of graphs.

Lemma 4.5. *For a (R, r) -clique-sparse graph $\mathcal{G}$, the number of partitions $M(\mathcal{G})$ induced by the equivalence relation in Equation (1) satisfies:*

$$M(\mathcal{G}) \leq R + rR(R - 1).$$

The following lemma shows that for graphs with constant maximum degree Δ , the square chromatic number is also constant, implying that PUCB-I is near optimal for these graphs as well.

Lemma 4.6. *For any graph $\mathcal{G}$ with maximum degree Δ , the square chromatic number satisfies*

$$\chi(\mathcal{G}^2) \leq \Delta^2 + 1.$$

5 UNKNOWN INTERFERENCE GRAPH

The assumption of knowing the interference graph may be unrealistic in certain scenarios. In this section, we study regret minimization when the graph is unknown. We show that, without any prior information on the graph, no algorithm can uniformly outperform the one that assumes a complete graph. For a policy π , graph $\mathcal{G}$, we denote by $R_T(\pi, \mathcal{G})$ the worst-case regret of policy π over all instances with 1-sub-Gaussian rewards and true graph $\mathcal{G}$ when the graph is unknown.

A natural extension of PUCB-I to this setting is Partitioned UCB with Unknown Interference (PUCB-UI) which assumes that for any instance, the interference graph is complete and runs PUCB-I accordingly. Clearly, this algorithm satisfies:

$$\forall \mathcal{G} : R_T(\pi_{\text{PUCB-UI}}, \mathcal{G}) \in \mathcal{O} \left(\sqrt{\frac{T}{N}} k^N \right),$$

matching the worst-case regret of PUCB-I with a known, complete graph.

The following theorem shows that no algorithm can uniformly outperform PUCB-UI across all problem instances.

Theorem 5.1. *[Unknown Graph Lower Bound] For any policy π and graph $\mathcal{G}$, if $k > \frac{2^{\frac{1}{N}}}{2^{\frac{1}{N}} - 1}$,*

$$R_T(\pi, \mathcal{G}) R_T(\pi, K_N) \in \Omega \left(\frac{T}{N} k^N \right).$$

The proof appears in Appendix A.3.

Corollary 5.2. *Theorem 5.1 implies that for a policy π , if there exists a constant c_1 such that*

$$\forall \mathcal{G} : R_T(\pi, \mathcal{G}) \leq c_1 \sqrt{\frac{T}{N}} k^N,$$

then there exists a constant c_2 such that

$$\forall \mathcal{G} : R_T(\pi, \mathcal{G}) \geq c_2 \sqrt{\frac{T}{N}} k^N.$$

Thus, no algorithm can uniformly outperform PUCB-UI across all instances.

6 EXPERIMENTS

We empirically evaluate PUCB-I (Algorithm 1) against the following baselines in simulations to validate our theoretical guarantees.

Classical UCB. This algorithm ignores the interference graph, treats each treatment in $k^{[N]}$ as an independent arm, and performs the UCB algorithm.

Combinatorial UCB (CUCB). Proposed by Chen et al. [2013b], CUCB is one of the most widely used algorithms in the combinatorial bandit framework. By formulating our problem within this framework, CUCB can be applied to the MAB setting with interference.

Network Explore-Then-Commit (ETC). Proposed by Agarwal et al. [2024], this algorithm runs in two phases: it first assigns treatments uniformly at random, then estimates the reward parameters via least squares regression, using the known interference graph. In the second phase, it plays the arm with the highest estimated reward for the remaining rounds.

Sequential Action Elimination (SAE). Also from Agarwal et al. [2024], this elimination-based algorithm begins with all $k^{[N]}$ arms active. In each epoch, it pulls each active arm equally and eliminates those with poor observed performance.

The complete implementation details are provided in the supplementary material. In all experiments, we set $k = 2$ and $\Delta = 3$. The total number of rounds is set to $T_{\max} = 10 \cdot 2^N$, ensuring sufficient exploration of all 2^N arms, as required by classical UCB and SAE algorithms. The rewards for each unit are independently drawn from a 1-Gaussian distribution. Additionally, the theoretical value of δ suggested in Theorem 3.3 is overly conservative in practice. To obtain more reasonable confidence bounds, we used a larger δ satisfying $\log(2/\delta) = \frac{N}{5} \log(0.02)$. We conducted two sets of experiments, analyzing the cumulative regret as a function of T and N . All results are averaged over 50 independent runs, with the standard deviation of the regret displayed in the plots.

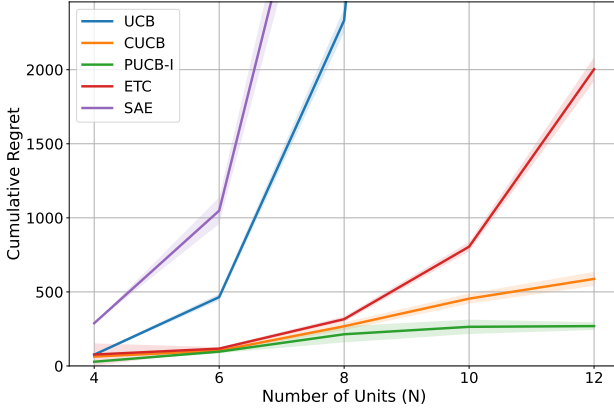


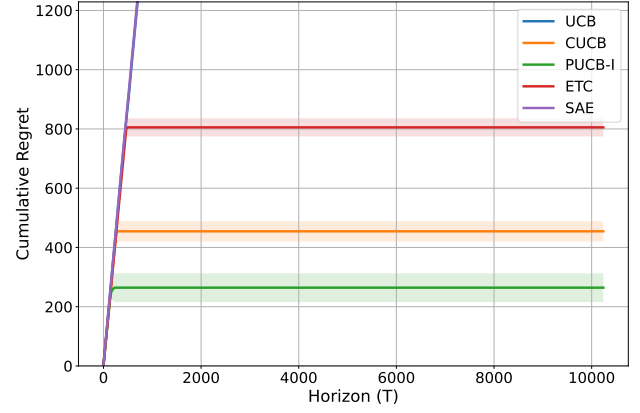
Figure 3: Average regret vs. number of units (N).

Scaling of Regret with T . Figures 4a and 4b present the cumulative regret of different algorithms over time for $N = 10$ and $N = 12$ units. The results highlight the superior performance of PUCB-I in both settings. In comparison, UCB and SAE exhibit a linear increase in regret during their long exploration phases, resulting in significantly higher overall regret and lower practical effectiveness.

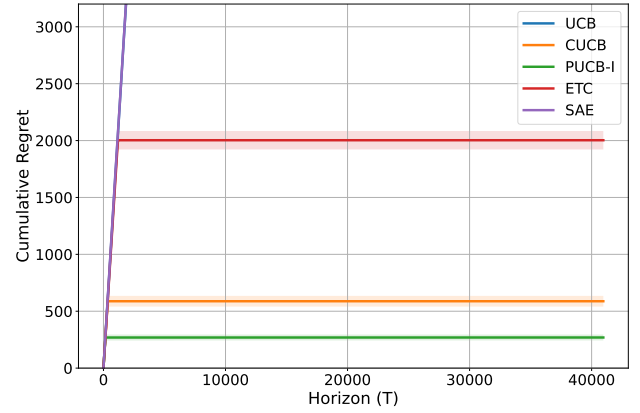
Scaling of Regret with N . To evaluate the impact of the number of units N , we analyzed the cumulative regret of the algorithms across instances by varying N in $\{4, 6, 8, 10, 12\}$. As shown in Figure 3, when N increases, the regret of UCB grows exponentially, confirming that it was not able to incorporate the graph structure. In contrast, the regret of PUCB-I grows more slowly as N increases compared to all other baselines because it benefits from partitioning the units with similar neighborhoods and graph-based initialization.

7 CONCLUSION

We studied the MAB problem with network interference, where each unit’s reward depends on its own treatment and those of its neighbors. To address the resulting challenges posed by interference, we proposed the PUCB-I algorithm, which partitions units based on their neighborhood structure to minimize cumulative regret. We derived a graph-dependent regret upper bound and established the first regret lower bounds for bandits with arbitrary network interference, showing that our proposed algorithm is near-optimal (up to logarithmic factors) for both sparse and dense graphs. We also considered the setting where the interference graph is unknown and showed that a variant of our algorithm is Pareto optimal. Our work highlights the importance of accounting for the graph structure in sequential decision-making under interference.



(a) Average regret over time for an instance with $N = 10$, $k = 2$ and $\Delta = 3$.



(b) Average regret over time for an instance with $N = 12$, $k = 2$ and $\Delta = 3$.

Figure 4: Comparison of average regret for various algorithms.

References

- Shishir Adhikari, Sourav Medya, and Elena Zheleva. Learning exposure mapping functions for inferring heterogeneous peer effects. *arXiv preprint arXiv:2503.01722*, 2025.
- Abhineet Agarwal, Anish Agarwal, Lorenzo Masoero, and Justin Whitehouse. Multi-armed bandits with network interference. *arXiv preprint arXiv:2405.18621*, 2024.
- Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.
- Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

- Peter M Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.
- Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. 2017.
- Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.
- Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.
- P Auer. Finite-time analysis of the multiarmed bandit problem, 2002.
- Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas Richardson, and Ido M Rosen. Multiple randomization designs. *arXiv preprint arXiv:2112.13495*, 2021.
- Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas S Richardson, and Ido M Rosen. Experimental design in marketplaces. *Statistical Science*, 38(3):458–476, 2023.
- Guillaume Basse and Avi Feller. Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113(521):41–55, 2018.
- Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*, pages 1028–1038. PMLR, 2020.
- Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.
- Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, pages 23–37. Springer, 2009.
- Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.
- Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.
- Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013a.
- Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013b.
- Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.
- Marek Cygan, Fedor V. Fomin, Łukasz Kowalik, Daniel Lokshantov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. doi: 10.1007/978-3-319-21275-3.
- Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017.
- Ahmed Sayeed Faruk and Elena Zheleva. Leveraging heterogeneous spillover in maximizing contextual bandit rewards. In *Proceedings of the ACM on Web Conference 2025*, pages 3049–3060, 2025.
- Mengsi Gao and Peng Ding. Causal inference in network experiments: regression-based analysis and design-based properties. *arXiv preprint arXiv:2309.07476*, 2023.
- Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.
- Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Confounded budgeted causal bandits. In *Causal Learning and Reasoning*, pages 423–461. PMLR, 2024.
- Su Jia, Nishant Oli, Ian Anderson, Paul Duff, Andrew A Li, and Ramamoorthi Ravi. Short-lived high-volume bandits. In *International Conference on Machine Learning*, pages 14902–14929. PMLR, 2023.
- Su Jia, Peter Frazier, and Nathan Kallus. Multi-armed bandits with interference. *arXiv preprint arXiv:2402.01845*, 2024.
- Paul Lagrée, Claire Vernade, and Olivier Cappe. Multiple-play bandits in the position-based model. *Advances in Neural Information Processing Systems*, 29, 2016.

- Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Michael P Leung. Network cluster-robust inference. *Econometrica*, 91(2):641–667, 2023.
- Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. *The Annals of Statistics*, 50(4):2334–2358, 2022.
- Ilia Mahrooghi, Mahshad Moradi, Sina Akbari, and Negar Kiyavash. Multi-armed bandits with missing outcomes. In *Conference on Uncertainty in Artificial Intelligence*, pages 2844–2875. PMLR, 2025.
- Nadav Merlis and Shie Mannor. Tight lower bounds for combinatorial multi-armed bandits. In *Conference on Learning Theory*, pages 2830–2857. PMLR, 2020.
- Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 33:5429–5440, 2020.
- Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.
- Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200, 2007.
- Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.
- Mohammad Shahverdikondori, Amir Mohammad Abouei, Alireza Rezaeimoghadam, and Negar Kiyavash. Optimal best arm identification with post-action context. *arXiv preprint arXiv:2502.03061*, 2025.
- Mohammad Shahverdikondori, Sepehr Elahi, Patrick Thiran, and Negar Kiyavash. Neighborhood-aware graph labeling problem. *arXiv preprint arXiv:2602.08098*, 2026.
- Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.
- Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517, 2017.
- Ruo-Chun Tzeng, Po-An Wang, Alexandre Proutiere, and Chi-Jen Lu. Closing the computational-statistical gap in best arm identification for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 36: 18391–18403, 2023.
- Johan Ugander and Hao Yin. Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1): 20220014, 2023.
- Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.
- Aad W. Van der Vaart and Jon A. Wellner. Weak convergence and empirical processes: Introduction to nonparametric estimation. *Springer Series in Statistics.*, 1996.
- Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. *Advances in Neural Information Processing Systems*, 30, 2017.
- Yang Xu, Wenbin Lu, and Rui Song. Linear contextual bandits with interference. *arXiv preprint arXiv:2409.15682*, 2024.
- Jing Yang, Stark C Draper, and Robert Nowak. Learning the interference graph of a wireless network. *IEEE Transactions on Signal and Information Processing over Networks*, 3(3):631–646, 2016.
- Christina Lee Yu, Edoardo M Airolidi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.
- Yuan Yuan, Kristen Altenburger, and Farshad Kooti. Causal network motifs: Identifying heterogeneous spillover effects in a/b tests. In *Proceedings of the Web Conference 2021*, pages 3359–3370, 2021.
- Zhiheng Zhang and Zichen Wang. Online experimental design with estimation-regret trade-off under network interference. *arXiv preprint arXiv:2412.03727*, 2024.

A OMITTED PROOFS

A.1 PROOFS OF SECTION 3

Theorem 3.3. [Graph-Partitioned Regret Upper Bound] *The expected cumulative regret of Algorithm 1 with $\delta = (T^2 N \sum_{j \in [M]} k^{D_j+1})^{-1}$ interacting with any instance with 1-sub-Gaussian rewards and interference graph $\mathcal{G}$, partitioned into $P_1, P_2, \dots, P_M$, satisfies:*

$$\text{Reg}_T \in \mathcal{O} \left(\sqrt{\frac{T}{N^2} \log(TN)} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

Proof. By the definition of A_{t+1} , we have:

$$\sum_{i \in [N]} \hat{\mu}_{ti}(A_t) + \sum_{j \in [M]} \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A_t)}} \geq \sum_{i \in [N]} \hat{\mu}_{ti}(A^*) + \sum_{j \in [M]} \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A^*)}}, \quad (6)$$

where $A^* \in \arg \max_{A \in [k]^N} \sum_{i \in [N]} \mu_i(A)$ denotes the optimal treatment. We define the good event G , ensuring that the empirical mean reward is close to the true mean for all times t , partitions j , and treatments A :

$$G := \mathbb{1} \left\{ \forall t \in [T], j \in [M], A \in [k]^N : \left| \sum_{i \in P_j} \hat{\mu}_{ti}(A) - \mu_{ti}(A) \right| \leq \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A)}} \right\}.$$

Applying the Hoeffding's inequality, we obtain:

$$P \left(\left| \sum_{i \in P_j} \hat{\mu}_{ti}(A) - \mu_{ti}(A) \right| \geq \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A)}} \right) \leq \delta.$$

By a union bound over all t, j and A :

$$\begin{aligned} P(G^c) &\leq \sum_{t \in [T]} \sum_{j \in [M]} \sum_{A \in [k]^N} P \left(\left| \sum_{i \in P_j} \hat{\mu}_{ti}(A) - \mu_{ti}(A) \right| \geq \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A)}} \right) \\ &\leq T \sum_{j \in [M]} k^{D_j+1} \delta, \end{aligned}$$

where D_j denotes the degree of each unit in the j -th partition.

By the law of total expectations, we can write the regret under event G as:

$$\begin{aligned} N \cdot \text{Reg}_T &= \sum_{t \in [T]} \sum_{i \in [N]} \mu_{ti}(A^*) - \mu_{ti}(A_t) \\ &\stackrel{(a)}{\leq} \sum_{t \in [T]} \sum_{i \in [N]} \hat{\mu}_{ti}(A^*) + \sum_{j \in [M]} \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A^*)}} - \left[\hat{\mu}_{ti}(A_t) - \sum_{j \in [M]} \sqrt{2 \log(2/\delta) \frac{m_j}{n_{tP_j}(A_t)}} \right] \\ &\stackrel{(b)}{\leq} 2\sqrt{2 \log(2/\delta)} \sum_{t \in [T]} \sum_{j \in [M]} \sqrt{\frac{m_j}{n_{tP_j}(A_t)}} \end{aligned}$$

where (a) and (b) hold by definition of event G and (6), respectively.

Next, we bound $\sum_{t \in [T]} \frac{1}{\sqrt{n_{tP_j}(A_t)}}$ as

$$\begin{aligned}
\sum_{t \in [T]} \frac{1}{\sqrt{n_{tP_j}(A_t)}} &= \sum_{A \in [k]^{D_j+1+1}} \sum_{t \in [T]} \mathbb{1}\{\forall i \in P_j : A_{tN(i)} = A_{N(i)}\} \frac{1}{\sqrt{n_{tP_j}(A)}} \\
&= \sum_{A \in [k]^{D_j+1}} \sum_{t \in [n_{TP_j}(A)]} \frac{1}{\sqrt{t}} \\
&\stackrel{(a)}{\leq} 2 \sum_{A \in [k]^{D_j+1}} \sqrt{n_{TP_j}(A)} \\
&\stackrel{(b)}{\leq} 2 \sqrt{k^{D_j+1} \sum_{A \in [k]^{D_j+1}} n_{TP_j}(A)} \\
&\leq 2\sqrt{T k^{D_j+1}}
\end{aligned} \tag{7}$$

where (a) holds since $\forall n \in \mathbb{N} : \sum_{i \in [n]} 1/\sqrt{i} \leq 2\sqrt{n}$, and (b) follows from Jensen inequality.

We prove the inequality $\sum_{i=1}^n \frac{1}{\sqrt{i}} \leq 2\sqrt{n}$ using integral approximation. Since the function $f(x) = \frac{1}{\sqrt{x}}$ is decreasing,

$$\sum_{i=1}^n \frac{1}{\sqrt{i}} \leq 1 + \int_1^n \frac{1}{\sqrt{x}} dx.$$

Evaluating the integral,

$$\int_1^n \frac{1}{\sqrt{x}} dx = 2\sqrt{n} - 2.$$

Thus,

$$\sum_{i=1}^n \frac{1}{\sqrt{i}} \leq 1 + (2\sqrt{n} - 2) \leq 2\sqrt{n}.$$

Therefore, the regret under event G is:

$$N.Reg_T \leq 4\sqrt{2\log(2/\delta)T} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \tag{8}$$

Thus, by choosing $\delta = \frac{1}{T^2 N \sum_{j \in [M]} k^{D_j+1}}$ we get the following by the law of total probability:

$$\begin{aligned}
Reg_T &\leq 4/N \sqrt{2\log(2/\delta)T} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} + T(T \sum_{j \in [M]} k^{D_j+1} \delta) \\
&\in \mathcal{O} \left(\sqrt{\frac{T}{N^2} \log(T^2 N \sum_{j \in [M]} k^{D_j+1})} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).
\end{aligned}$$

Therefore, since $T \geq k^{\Delta+1}$, it satisfies:

$$Reg_T \in \mathcal{O} \left(\sqrt{\frac{T}{N^2} \log(TN)} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

□

A.2 PROOFS OF SECTION 4

Theorem 4.1. [Graph-Partitioned Regret Lower Bound] Let $\mathcal{G}$ be a connected graph with N nodes and partitions $\{P_1, P_2, \dots, P_M\}$ with $|P_j| = m_j$. If $k > \frac{2^{\frac{1}{\Delta+1}}}{2^{\frac{1}{\Delta+1}} - 1}$ and $T \geq \frac{4(k-1)^{\Delta+1}}{M}$, then for any policy π , there exists a bandit instance with interference graph $\mathcal{G}$ whose reward is distributed as 1-Gaussian with means in $[0, 1]$ such that

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T}{N^2 M}} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

Proof. To prove this lower bound, we employ a change-of-measure argument, a well-known technique in the multi-armed bandit literature for deriving lower bounds. Consider an instance $\mathcal{V}$ with interference graph $\mathcal{G}$, Gaussian reward noises with unit variance, and the following reward means for each unit $i \in [N]$:

$$\mu(A_{N(i)}) = \begin{cases} \Delta_i & \text{if } A_{N(i)} = (1, 1, \dots, 1), \\ 0 & \text{otherwise,} \end{cases}$$

where Δ_i is a positive real number to be determined later. In this case, the optimal treatment is to assign treatment 1 to all units, yielding an expected reward of $\sum_{i \in [N]} \Delta_i$.

Now, fix a policy π operating on instance $\mathcal{V}$. For each treatment $A \in [k]^N$, let $T_\pi(A) = (t_{\pi,1}(A), t_{\pi,2}(A), \dots, t_{\pi,N}(A))$, where $t_{\pi,i}(A)$ is the expected number of times policy π applies a treatment assignment such that unit i and its neighbors receive treatment $A_{N(i)}$ over T rounds. For simplicity, we denote $T_\pi(A)$ and $t_{\pi,i}(A)$ as $T(A)$ and $t_i(A)$, respectively.

Let $S \subset [k]^N$ be the set of all treatments where no unit receives treatment 1, therefore, $|S| = (k-1)^N$. The following lemma implies that there is always an under-explored treatment.

Lemma A.1. For any policy π and any set of values Δ_i for $i \in [N]$, there exists a treatment $A' \in S$ such that:

$$\sum_{i \in [N]} t_i(A') \Delta_i^2 \leq T \left(\sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}} \right).$$

Proof. To prove this lemma, we employ a double-counting technique. Consider a matrix with $(k-1)^N$ rows and N columns where row i corresponds to a treatment $A_i \in S$ and column i corresponds to unit i . On the element in j -th row and i -th column of the matrix, we write $t_i(A_j) \Delta_i^2$ as follows:

$$\mathbf{M} := \begin{bmatrix} t_1(A_1) \Delta_1^2 & t_2(A_1) \Delta_2^2 & \cdots & t_N(A_1) \Delta_N^2 \\ t_1(A_2) \Delta_1^2 & t_2(A_2) \Delta_2^2 & \cdots & t_N(A_2) \Delta_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1(A_{(k-1)^N}) \Delta_1^2 & t_2(A_{(k-1)^N}) \Delta_2^2 & \cdots & t_N(A_{(k-1)^N}) \Delta_N^2 \end{bmatrix}.$$

To calculate the sum of the numbers in the first column, we have

$$\sum_{A_i \in S} t_1(A_i) \Delta_1^2 = \Delta_1^2 (k-1)^{N-(d_1+1)} \sum_{A_{N(1)} \in S_1} t(A_{N(1)}),$$

where $S_1 = \{2, 3, \dots, k\}^{d_1+1}$ is the set of all treatments that can be assigned to unit one and its neighbors without using treatment 1 and $t(A_{N(1)})$ is the expected number of times that unit one and its neighbors are assigned $A_{N(1)}$ during T rounds of interaction between π and $\mathcal{V}$. The equation is true because of the symmetry in the problem which implies that each combination of the treatment of unit 1 and its neighbors exists in $(k-1)^{N-(d_1+1)}$ number of members of S . On the other hand, we have

$$\sum_{A_{N(1)} \in S_1} t(A_{N(1)}) = T.$$

Writing the same equation for all the columns implies that the sum of the numbers in the whole matrix is equal to

$$T \sum_{i \in [N]} \Delta_i^2 (k-1)^{N-(d_i+1)}.$$

Dividing this number by the number of rows shows that there exists a row j such that

$$\sum_{i \in [N]} t_i(A_j) \Delta_i^2 \leq T \left(\sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}} \right).$$

□

Now, based on the under-explored treatment A' , we design a *confusing* instance $\mathcal{V}'$ which for each treatment A has the mean expected rewards as follows:

$$\mu'(A_{N(i)}) = \begin{cases} \Delta_i & \text{if } A_{N(i)} = (1, 1, \dots, 1), \\ 2\Delta_i & \text{if } A_{N(i)} = A'_{N(i)}, \\ 0 & \text{otherwise.} \end{cases}$$

We denote $\mathbb{P}_{\mathcal{V}}$ and $\mathbb{P}_{\mathcal{V}'}$ the probability measures over the bandit model induced by T rounds of interaction between the policy π and the instances $\mathcal{V}$ and $\mathcal{V}'$, respectively. In this case, using the divergence decomposition lemma [see Lemma 15.1 in Lattimore and Szepesvári, 2020] and the fact that the KL-divergence (denoted by $D_{KL}(\cdot \| \cdot)$) between two 1-Gaussian distributions with means μ_1 and μ_2 is equal to $(\mu_1 - \mu_2)^2/2$, we derive the following equality

$$D_{KL}(\mathbb{P}_{\mathcal{V}} \| \mathbb{P}_{\mathcal{V}'}) = \sum_{i \in [N]} t_i(A') \frac{(2\Delta_i)^2}{2} = 2 \sum_{i \in [N]} t_i(A') \Delta_i^2. \quad (9)$$

Using Lemma A.1, we have

$$D_{KL}(\mathbb{P}_{\mathcal{V}} \| \mathbb{P}_{\mathcal{V}'}) \leq 2T \left(\sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}} \right). \quad (10)$$

For each $t \in [T]$, define the event $\mathcal{E}_t$ as:

$$\mathcal{E}_t := \left\{ \mathbb{E} \left[\sum_{i \in \mathcal{O}_t} \Delta_i \right] \geq \frac{1}{2} \sum_{i \in [N]} \Delta_i \right\},$$

where $\mathcal{O}_t := \{i \mid A_{tN(i)} = (1, \dots, 1)\}$ denotes the set of units i that receive treatment 1 at time t , along with all their neighbors. This event signifies that in round t , the expected sum of Δ_i for such units i is at least half of the total $\sum_{i \in [N]} \Delta_i$. We further define the event $\mathcal{E}$ as:

$$\mathcal{E} := \left\{ \sum_{t \in [T]} \mathbb{1}_{\mathcal{E}_t} \geq \frac{T}{2} \right\},$$

indicating that $\mathcal{E}_t$ occurs in at least half of the T rounds. Using the event $\mathcal{E}$, we can bound the regret of policy π for both instances $\mathcal{V}$ and $\mathcal{V}'$. For $\mathcal{V}$ we have:

$$\begin{aligned} \text{Reg}_T(\pi, \mathcal{V}) &\geq \mathbb{P}_{\mathcal{V}}(\mathcal{E}^c) (\text{Reg}_T(\pi, \mathcal{V}) \mid \mathcal{E}^c) \\ &\geq \mathbb{P}_{\mathcal{V}}(\mathcal{E}^c) \frac{T}{4N} \sum_{i \in [N]} \Delta_i, \end{aligned}$$

where, with a slight abuse of notation, we use $\text{Reg}_T(\pi, \mathcal{V}) \mid \mathcal{E}^c$ to denote regret (which is an expectation) conditioned on the event $\mathcal{E}^c$. The second line holds because, on $\mathcal{E}^c$, the algorithm incurs a regret of at least $\sum_{i \in [N]} \Delta_i$ in at least half of the rounds. Consequently, the regret of π interacting with $\mathcal{V}$ is at least $\frac{T}{4N} \sum_{i \in [N]} \Delta_i$.

Similarly, for $\mathcal{V}'$, the regret of π can be bounded using $\mathcal{E}$ as:

$$\begin{aligned} \text{Reg}_T(\pi, \mathcal{V}') &\geq \mathbb{P}_{\mathcal{V}'}(\mathcal{E}) (\text{Reg}_T(\pi, \mathcal{V}') \mid \mathcal{E}) \\ &\geq \mathbb{P}_{\mathcal{V}'}(\mathcal{E}) \frac{T}{4N} \sum_{i \in [N]} \Delta_i. \end{aligned} \quad (11)$$

On the other hand, Bretagnolle-Huber inequality [Van der Vaart and A. Wellner, 1996] implies:

$$\begin{aligned} \mathbb{P}_{\mathcal{V}}(\mathcal{E}^c) + \mathbb{P}_{\mathcal{V}'}(\mathcal{E}) &\geq \frac{1}{2} \exp(-D_{KL}(\mathbb{P}_{\mathcal{V}} \parallel \mathbb{P}_{\mathcal{V}'})) \\ &\geq \frac{1}{2} \exp\left(-2T \sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}}\right), \end{aligned} \quad (12)$$

where the second line holds by Equation (10). From Equations (11) and (12), we have:

$$\text{Reg}_T(\pi, \mathcal{V}) + \text{Reg}_T(\pi, \mathcal{V}') \geq \frac{T}{8N} \left(\sum_{i \in [N]} \Delta_i \right) \exp\left(-2T \sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}}\right).$$

Next, for each $j \in [M]$ and each $l \in P_j$, set the value of Δ_l to:

$$\Delta_l = \sqrt{\frac{(k-1)^{D_j+1}}{TMm_j}},$$

which satisfies

$$T \sum_{i \in [N]} \frac{\Delta_i^2}{(k-1)^{d_i+1}} = T \sum_{j \in [M]} \sum_{l \in P_j} \frac{1}{Mm_j} = T \sum_{j \in [M]} \frac{1}{M} = 1,$$

and

$$\begin{aligned} \sum_{i \in [N]} \Delta_i &= \sqrt{\frac{1}{TM}} \sum_{j \in [M]} \sum_{l \in P_j} \sqrt{\frac{(k-1)^{D_j+1}}{m_j}} \\ &= \sqrt{\frac{1}{TM}} \sum_{j \in [M]} \sqrt{m_j (k-1)^{D_j+1}}. \end{aligned}$$

Note that by the assumption $T \geq \frac{4(k-1)^{\Delta+1}}{M}$, we have $\Delta_i \leq \frac{1}{2}$ which implies that the mean rewards are in $[0, 1]$.

Now, note that if $k > \frac{2^{\frac{1}{\Delta+1}}}{2^{\frac{1}{\Delta+1}} - 1}$, then $(k-1)^{D_j+1} > \frac{1}{2}k^{D_j+1}$. Using this, there exists a universal constant C such that at least one of $\text{Reg}_T(\pi, \mathcal{V})$ and $\text{Reg}_T(\pi, \mathcal{V}')$ is greater than

$$C \left(\sqrt{\frac{T}{N^2 M}} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

Therefore,

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T}{N^2 M}} \sum_{j \in [M]} \sqrt{m_j k^{D_j+1}} \right).$$

□

Theorem 4.2. [Doubly-Independent Set Regret Lower Bound] Let $\mathcal{G}$ be a connected graph with N nodes. If $T \geq k^{\Delta+1} - 1$, for any policy π , there exists a bandit instance with interference graph $\mathcal{G}$, 1-Gaussian reward distributions and means in $[0, 1]$ such that

$$\text{Reg}_T(\pi) \in \Omega \left(\max_{S \in DI(\mathcal{G})} \sqrt{\frac{T}{N^2}} \sum_{i \in S} \sqrt{k^{d_i+1}} \right), \quad (4)$$

where $DI(\mathcal{G})$ is defined in Definition 3.1.

Proof. To prove this theorem, we first provide two important lemmas.

Lemma A.2. For a graph $\mathcal{G}$, let $\mathcal{G}'$ be the graph obtained after removing an arbitrary edge from $\mathcal{G}$. Then, for any policy π , the worst-case regret of interacting with bandit instances on $\mathcal{G}$ is at least that of interacting with bandit instances on $\mathcal{G}'$. That is,

$$\sup_{\mathcal{V} \sim \mathcal{G}} \text{Reg}_T(\pi, \mathcal{V}) \geq \sup_{\mathcal{V}' \sim \mathcal{G}'} \text{Reg}_T(\pi, \mathcal{V}'),$$

where $\mathcal{V} \sim \mathcal{G}$ denotes a bandit instance whose interference graph is $\mathcal{G}$.

Proof. Assume $\mathcal{G}'$ is obtained by removing an edge between nodes i and j . To prove this lemma, consider any instance $\mathcal{V}'$ with the interference graph $\mathcal{G}'$. Construct an instance $\mathcal{V}$ with the interference graph $\mathcal{G}$ such that the reward distribution for each treatment $A_{N(i)}$ for unit i is identical to its distribution in $\mathcal{V}'$.

In this construction, the reward distribution for the unit i becomes independent of the node j , and the edge $i - j$ has no impact on the rewards. Consequently, no algorithm can achieve a lower regret on $\mathcal{V}$ compared to $\mathcal{V}'$. This implies that for every instance with the interference graph $\mathcal{G}'$, a harder instance exists with the interference graph $\mathcal{G}$, which completes the proof. $\square$

Lemma A.3. Let $\mathcal{G}$ be a graph with N nodes where d_i is the degree of node i . If $T \geq k^{d_i+1} - 1$, for any policy π and each $i \in [N]$, there exists a bandit instance with interference graph $\mathcal{G}$, 1-Gaussian reward distributions and means in $[0, 1]$ such that:

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T k^{d_i+1}}{N^2}} \right).$$

Proof. To prove this lemma, observe that the reward distribution of unit i for each treatment $A_{N(i)}$ assigned to unit i and its neighbors is independent of other treatments. This allows us to construct a corresponding classic multi-armed bandit instance with k^{d_i+1} arms, where each arm represents a treatment $A_{N(i)}$ and follows the same reward distribution.

In this scenario, based on the classic lower bound in the multi-armed bandit literature [Lattimore and Szepesvári, 2020], for any policy interacting with a bandit with K arms, there exists an instance where the regret $\text{Reg}_T(\pi)$ is at least $\sqrt{TK}$. Applying this to our problem, where total regret is defined as the average regret of units, implies that for any policy π , there exists an instance with the interference graph $\mathcal{G}$ such that:

$$\text{Reg}_T(\pi) = \Omega \left(\sqrt{\frac{T k^{d_i+1}}{N^2}} \right).$$

$\square$

We prove the theorem using the aforementioned lemmas as follows. For any set $S = \{s_1, s_2, \dots, s_m\} \in DI(\mathcal{G})$, let $\mathcal{G}_S$ be the graph obtained from $\mathcal{G}$ by removing all edges between nodes that are outside of S (i.e., edges where neither endpoint belongs to S). By Lemma A.2, the worst-case regret for instances with interference graph $\mathcal{G}$ is at least as large as the worst-case regret for instances with $\mathcal{G}_S$. The graph $\mathcal{G}_S$ consists of edges where at least one endpoint lies in S . Since $S \in DI(\mathcal{G})$, the resulting graph $\mathcal{G}_S$ consists of m disjoint connected components $\mathcal{G}_{s_1}, \mathcal{G}_{s_2}, \dots, \mathcal{G}_{s_m}$. Each component $\mathcal{G}_{s_i}$ is a star graph with $d_{s_i} + 1$ nodes.

By Lemma A.3, for any policy π interacting with $\mathcal{G}_{s_i}$, there exists a bandit instance where:

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T k^{d_{s_i}+1}}{N^2}} \right).$$

Since the components $\mathcal{G}_{s_i}$ are disjoint, the total regret over $\mathcal{G}_S$ equals the sum of the regrets for each component. Therefore, for any policy π and any $S \in DI(\mathcal{G})$, there exists an instance such that:

$$\text{Reg}_T(\pi) \in \Omega \left(\sqrt{\frac{T}{N^2}} \sum_{i \in S} \sqrt{k^{d_i+1}} \right).$$

This concludes the proof. $\square$

Lemma 4.5. *For a (R, r) -clique-sparse graph $\mathcal{G}$, the number of partitions $M(\mathcal{G})$ induced by the equivalence relation in Equation (1) satisfies:*

$$M(\mathcal{G}) \leq R + rR(R - 1).$$

Proof. For each cluster C_i , let $|C_i| = c_i$. It is given that the number of edges between the nodes in C_i and C_j is at most r for each pair i, j . This implies that, within cluster C_i , there are at most $r(R - 1)$ nodes with neighbors outside of C_i . Let D_i denote the set of nodes in C_i that have no neighbors outside C_i . Thus, we have:

$$|D_i| \geq c_i - r(R - 1).$$

Since C_i is a complete graph, it holds that:

$$\forall l \in D_i : N(l) = C_i.$$

This indicates that for all nodes $l \in D_i$, the set $N(l)$ is identical, which implies that all such nodes belong to the same partition. Furthermore, there are at most $r(R - 1)$ nodes in $C_i \setminus D_i$, meaning that in total these nodes can form:

$$\sum_{i \in [R]} |C_i \setminus D_i| \leq \sum_{i \in [R]} r(R - 1) = rR(R - 1)$$

distinct partitions. This implies that the total number of partitions $M(\mathcal{G})$ is bounded as:

$$M(\mathcal{G}) \leq R + rR(R - 1).$$

$\square$

Lemma 4.6. *For any graph $\mathcal{G}$ with maximum degree Δ , the square chromatic number satisfies*

$$\chi(\mathcal{G}^2) \leq \Delta^2 + 1.$$

Proof. To prove the lemma, we provide a coloring method using at most $\Delta^2 + 1$ colors, ensuring that two nodes with the same color are neither adjacent nor share a common neighbor. Consider an arbitrary order $\{v_1, v_2, \dots, v_N\}$ on the nodes, and let the available colors be $\{c_1, c_2, \dots, c_{\Delta^2+1}\}$. Start with v_1 , and for each v_i , assign the smallest color c_j such that no node already colored with c_j is adjacent to v_i or shares a common neighbor with v_i .

To prove that this coloring method is valid and does not require more than $\Delta^2 + 1$ colors, assume the process stops at v_i because no valid color is available. This would mean v_i has more than Δ^2 nodes that are either adjacent to it or share a common neighbor. However, since the maximum degree is Δ , v_i has at most Δ neighbors, and each neighbor can have at most $\Delta - 1$ other neighbors. This totals at most Δ^2 nodes, contradicting the assumption that more than Δ^2 nodes are involved. Thus, the coloring method works as intended. This shows that for constant values of Δ (independent of N and k), the value $\Delta^2 + 1$ is also constant.

$\square$

A.3 PROOFS OF SECTION 5

Theorem 5.1. [Unknown Graph Lower Bound] For any policy π and graph $\mathcal{G}$, if $k > \frac{2^{\frac{1}{N}}}{2^{\frac{1}{N}} - 1}$,

$$R_T(\pi, \mathcal{G}) R_T(\pi, K_N) \in \Omega\left(\frac{T}{N} k^N\right).$$

Proof. Fix an arbitrary graph $\mathcal{G}$ and consider the instance $\mathcal{V}$ with interference graph $\mathcal{G}$, 1-Gaussian rewards, and the following reward means for each unit i under treatment A ,

$$\mu(A_{N(i)}) = \begin{cases} \Delta & \text{if } A_{N(i)} = (1, 1, \dots, 1), \\ 0 & \text{otherwise,} \end{cases}$$

where Δ is a positive real number to be determined later. The optimal treatment is to assign treatment 1 to everyone, which yields an expected reward of $N\Delta$.

Let $S \subset [k]^N$ denote the set of all treatments where no unit receives treatment 1. For each treatment $A' \in S$, we construct the instance $\mathcal{V}_{A'}$ with the following properties: the interference graph is K_N (complete graph with N nodes), 1-Gaussian rewards with means

$$\mu'(A_{N(i)}) = \begin{cases} \Delta & \text{if } A_{N(i)} = (1, 1, \dots, 1), \\ 2\Delta & \text{if } A = A', \\ 0 & \text{otherwise,} \end{cases}$$

where $A_{N(i)}$ denotes the treatment assigned to the neighborhood of node i in graph $\mathcal{G}$. The instance $\mathcal{V}_{A'}$ differs from the base instance $\mathcal{V}$ only in the units' rewards assigned to treatment A' , which is now the optimal treatment in $\mathcal{V}_{A'}$.

Now consider a policy π . During the interaction of π with a problem instance under an unknown interference graph, we define the event $\mathcal{E}$ as:

$$\mathcal{E} = \left\{ \sum_{t \in [T]} n_t \geq \frac{TN}{2} \right\},$$

where n_t denotes the random number of nodes i in round t whose entire neighborhood receives treatment 1. Note that under $\mathcal{E}$, the regret of π when interacting with $\mathcal{V}_{A'}$ is at least $\frac{T\Delta}{2}$, and under $\mathcal{E}^c$, the regret of π interacting with $\mathcal{V}$ is at least $\frac{T\Delta}{2}$. This shows that

$$R_T(\pi, \mathcal{G}) + R_T(\pi, K_N) \geq R_T(\pi, \mathcal{V}) + R_T(\pi, \mathcal{V}_{A'}) \geq \frac{T\Delta}{2} (\mathbb{P}_{\mathcal{V}}(\mathcal{E}^c) + \mathbb{P}_{\mathcal{V}_{A'}}(\mathcal{E})),$$

where $\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{\mathcal{V}_{A'}}$ denote the probability measures over the bandit model induced by T rounds of interaction of the policy π and instances $\mathcal{V}$ and $\mathcal{V}'$, respectively, when the interference graph is unknown.

By Bretagnolle-Huber inequality [Van der Vaart and A. Wellner, 1996], we have

$$\mathbb{P}_{\mathcal{V}}(\mathcal{E}^c) + \mathbb{P}_{\mathcal{V}_{A'}}(\mathcal{E}) \geq \frac{1}{2} \exp(-D_{KL}(\mathbb{P}_{\mathcal{V}} \parallel \mathbb{P}_{\mathcal{V}_{A'}})).$$

Since $\mathcal{V}$ and $\mathcal{V}_{A'}$ only differ in the mean rewards of treatment A' , the KL divergence between their distributions can be computed as follows using the KL formula for two 1-Gaussian distributions:

$$D_{KL}(\mathbb{P}_{\mathcal{V}} \parallel \mathbb{P}_{\mathcal{V}_{A'}}) = N \frac{(2\Delta)^2}{2} \mathbb{E}_{\mathcal{V}}[T_{A'}],$$

where $\mathbb{E}_{\mathcal{V}}[T_{A'}]$ shows the expected number of times π plays A' on instance $\mathcal{V}$. Then we have:

$$2R_T(\pi, K_N) \geq R_T(\pi, \mathcal{G}) + R_T(\pi, K_N) \geq \frac{T\Delta}{4} \exp(-2N\Delta^2 \mathbb{E}_{\mathcal{V}}[T_{A'}]),$$

where the first inequality holds because, by Lemma A.2, the worst-case regret of any policy on instances with the fully connected graph is larger than or equal to the worst-case regret on instances with any other fixed graph. Then, by simple algebra, we have

$$\mathbb{E}_{\mathcal{V}}[T_{A'}] \geq \frac{\log\left(\frac{T\Delta}{4}\right) - \log(2R_T(\pi, K_N))}{2N\Delta^2}.$$

By setting $\Delta = \frac{12R_T(\pi, K_N)}{T}$,

$$\mathbb{E}_{\mathcal{V}}[T_{A'}] \geq \frac{\log(\frac{3}{2})T^2}{2 \times 144 \times NR_T(\pi, K_N)^2} = \frac{\log(\frac{3}{2})}{288} \frac{T^2}{NR_T(\pi, K_N)^2}.$$

As this inequality holds for every treatment $A' \in S$, then

$$\begin{aligned} R_T(\pi, \mathcal{G}) &\geq R_T(\pi, \mathcal{V}) \geq \frac{1}{N} \sum_{A' \in S} \mathbb{E}_{\mathcal{V}}[T_{A'}](N\Delta) \geq \frac{1}{N} \sum_{A' \in S} \frac{\log(\frac{3}{2})}{288} \frac{T^2}{NR_T(\pi, K_N)^2} \frac{12NR_T(\pi, K_N)}{T} \\ &= \frac{|S|}{N} \frac{\log(\frac{3}{2})}{288} \frac{12T}{R_T(\pi, K_N)} \\ &\geq \frac{k^N}{2N} \frac{\log(\frac{3}{2})}{288} \frac{12T}{R_T(\pi, K_N)} \in \Omega\left(\frac{Tk^N}{NR_T(\pi, K_N)}\right) \\ &\implies R_T(\pi, \mathcal{G})R_T(\pi, K_N) \in \Omega\left(\frac{T}{N}k^N\right), \end{aligned}$$

where the last inequality holds because $|S| = (k-1)^N > \frac{k^N}{2}$ for $k > \frac{2^{\frac{1}{N}}}{2^{\frac{1}{N}} - 1}$. □