# Stochastic Graphical Bandits with Heavy-Tailed Rewards

**Yutian Gou**[1] **Jinfeng Yi**[2] **Lijun Zhang**[1*]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[2]JD AI Research, Beijing 100176, China

## Abstract

We consider stochastic graphical bandits, where after pulling an arm, the decision maker observes rewards of not only the chosen arm but also its neighbors in a feedback graph. Most of existing work assumes that the rewards are drawn from bounded or at least sub-Gaussian distributions, which however may be violated in many practical scenarios such as social advertising and financial markets. To settle this issue, we investigate stochastic graphical bandits with heavy-tailed rewards, where the distributions have finite moments of order $1 + \epsilon$, for some $\epsilon \in (0, 1]$. Firstly, we develop one UCB-type algorithm, whose expected regret is upper bounded by a sum of gap-based quantities over the *clique covering* of the feedback graph. The key idea is to estimate the reward means of the selected arm's neighbors by more refined robust estimators, and to construct a graph-based upper confidence bound for selecting candidates. Secondly, we design another elimination-based strategy and improve the regret bound to a gap-based sum with size controlled by the *independence number* of the feedback graph. For benign graphs, the *independence number* could be smaller than the size of the *clique covering*, resulting in tighter regret bounds. Finally, we conduct experiments on synthetic data to demonstrate the effectiveness of our methods.

## 1 INTRODUCTION

As one of the most classical problem in online sequential decision-making, Multi-Armed Bandits (MAB) has been successfully applied to various real-world scenes such as medical trials [Villar et al., 2015, Gutiérrez et al., 2017], news recommendation [Li et al., 2010], online advertising [Chen et al., 2013, Xu et al., 2013, Schwartz et al., 2017], resource allocations [Lattimore et al., 2014], and online routing [Kveton et al., 2015]. In its original stochastic form [Robbins, 1952], at each round $t$, a player has to select an arm $i$ from $K$ available candidates and receives a reward generated independently from an unknown but fixed distribution. The player's goal is to minimize the *regret* over $T$ steps of the game, namely the difference between the cumulative rewards of the chosen arms and that of the optimal arm in hindsight. In order to achieve this goal, the player needs to overcome the dilemma of exploration (learning new information about all arms) and exploitation (selecting the optimal arm based on available information). In the seminal work, Lai and Robbins [1985] establish an $\Omega(K \log T)$ asymptotic regret lower bound and propose UCB policy that attains this lower bound asymptotically. In the past decades, plentiful algorithms and theoretical results for bandits have been well developed [Bubeck and Cesa-Bianchi, 2012, Lattimore and Szepesvári, 2020].

However, one limitation of the stochastic MAB is that the regret bound scales linearly with $K$, and thus may become vacuous when the arm set gets very large. To address this limitation, Mannor and Shamir [2011] introduce an important variant of MAB termed Graphical Bandits (GB). In this scene, there exists an undirected feedback graph with node set consisting of $K$ arms and edge set revealing the relationship between arms. After pulling an arm, the decision maker will observe the rewards from not only the chosen arm but also its neighbors in the graph. Later, Caron et al. [2012] consider the stochastic version of GB. They present UCB-based algorithms for stochastic GB with bounded rewards and provide regret bounds depending on the *clique covering* of the feedback graph, whose size can be much smaller than $K$ for benign graphs. For stochastic GB with bounded rewards, Caron et al. [2012] proposed a lower bound of $\Omega(\log T)$. However, no lower bounds have been proposed for stochastic GB under the sub-Gaussian setting [Marinov et al., 2022b].

While the stochastic GB has been extensively studied in

*Corresponding author.

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

the literature [Buccapatnam et al., 2014, 2017, Cohen et al., 2016, Tossou et al., 2017, Liu et al., 2018a,b, Hu et al., 2019, Lykouris et al., 2020, Marinov et al., 2022a], most previous studies assume that the rewards are drawn from either bounded or at least sub-Gaussian distributions. Since the sub-Gaussian random variables possess the characteristic of exponentially decaying tails, we can use the empirical mean to estimate the reward means of each arm, and guarantee exponential deviations by the standard concentration of measure techniques [Hoeffding, 1963]. However, there do exist practical scenarios which do not behave sub-Gaussian but can be modeled by the heavy-tailed distributions [Foss et al., 2011], such as frequent price fluctuations for financial markets [Rachev, 2003], preferential attachment in social networks [Mahanti et al., 2013] and unevenly distributed clicks of slogans in social advertising [Park et al., 2013]. Unfortunately, as heavy-tailed rewards no longer enjoy exponentially decaying tails, the empirical mean estimator can only provide polynomial concentration properties [Catoni, 2012], making it much harder to estimate the reward means of each arm.

In this study, we investigate stochastic GB with heavy-tailed rewards, where the reward distributions are assumed to have bounded $(1+\epsilon)$-th moments for some $\epsilon \in (0, 1]$. We present two novel algorithms for this setting based on more refined robust estimators. Firstly, we design one UCB-type algorithm named RUNE, whose expected regret is upper bounded by a sum of gap-based quantities over the *clique covering* of the feedback graph. The key idea is to estimate the reward means of the selected arm's neighbors by truncated empirical mean or median of means, and to construct a graph-based upper confidence bound for selecting candidates. Secondly, we propose another elimination-based algorithm termed RAAE and provide a regret bound by a gap-based sum whose size is controlled by the *independence number* of the feedback graph. For benign graphs, the *independence number* could be smaller than the size of the *clique covering*, resulting in tighter regret bounds. To the best of our knowledge, we provide the first regret bounds for stochastic GB with heavy-tailed rewards. Please refer to Table 1 for a comparison between our results and the previous results in stochastic graphical bandits. The contributions of this work are summarized as follows:

- We propose one novel UCB-type algorithm for stochastic GB with heavy-tailed rewards, named RUNE. Our algorithm obtains a gap-based logarithmic regret bound of $O(\sum_{C\in\mathcal{C}} \frac{v^{1/\epsilon}\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} + \sum_{C\in\mathcal{C}} \Delta_C^{\max})$, where $\mathcal{C}$ is a clique cover of $G$, $N_C$ is a quantity related to clique $C \in \mathcal{C}$, $\Delta_C^{\max}$ is the maximum reward gap of $C$, and $\Delta_C^{\min}$ is the minimum nonzero reward gap of $C$.
- To further improve the regret bound, we design another elimination-based algorithm termed RAAE and provide a gap-based logarithmic regret bound of $O(\sum_{i\in S} \frac{v^{1/\epsilon}\log T}{\Delta_i^{1/\epsilon}} + \Delta_{\max}\log T)$, where $\Delta_{\max}$ is the maximum suboptimal reward gap, $S$ is a subset of the first $\alpha$ suboptimal arms with the ties broken arbitrarily and $\alpha$ is the *independence number* of $G$. This regret bound is a substantial improvement over RUNE since the *independence number* is smaller than the size of *clique covering* for benign graphs.
- To demonstrate the effectiveness of our methods, we present synthetic experiments for comparing RUNE and RAAE with previous algorithms. The empirical results support our theoretical results.

## 2 PRELIMINARIES AND RELATED WORK

In this section, we first provide a formal description of our problem setup, and then review related work about stochastic bandits, including stochastic graphical bandits and stochastic bandits with heavy-tailed rewards.

### 2.1 PROBLEM SETUP AND DEFINITIONS

We consider stochastic GB with a fixed undirected feedback graph $G = (V, E)$, where $V = \{1, 2, \cdots, K\}$ denotes the arm set, and $E \subseteq V \times V$ reveals the relationship between arms. An edge $(i, j) \in E$ means that when the arm $i$ (or $j$) is pulled at round $t$, the player will receive a reward from $i$ and also observe the reward of $j$. For each arm $i \in V$, we assume that the reward $X_{i,t}$ at round $t$ is sampled independently from an unknown but fixed distribution $\mathcal{P}_i$ with mean $\mu_i$ and bounded $(1+\epsilon)$-th moments, i.e., $\mathbb{E}_{X\sim\mathcal{P}_i}[|X|^{1+\epsilon}] \le v$ or $\mathbb{E}_{X\sim\mathcal{P}_i}[|X-\mu_i|^{1+\epsilon}] \le v$.

The player's goal is to minimize the (pseudo) *expected regret* over $T$ steps of the game, which is defined as

$$\mathbb{E}[R_T] = T\mu^\star - \sum_{t=1}^{T}\mu_{I_t} = \sum_{i\in V}\Delta_i\mathbb{E}[T_i(T)]\,, \qquad (1)$$

where $\mu^\star = \max_{i\in V}\mu_i$, $I_t$ is the arm chosen by the player at round $t$, $\Delta_i := \mu^\star - \mu_i$ denotes the reward gap of arm $i$ relative to the optimal arm, and $T_i(T) = \sum_{t=1}^{T}\mathbb{I}_{\{I_t=i\}}$ refers to the number of pulls for arm $i$ up to time $T$.

Note that the player in MAB can only observe the rewards from the selected arm $I_t$ at round $t$, whereas in GB, the rewards from its neighbors can be also observed. In other words, the main difference between GB and MAB lies in the fact that the number of observations made for arm $i$ until round $T$ is no longer $T_i(T)$ in (1) but

$$O_i(T) = \sum_{t=1}^{T}\mathbb{I}_{\{I_t\in N(i)\}}\,, \qquad (2)$$

where $N(i)$ denotes the set consisting of arm $i$ and its adjacent nodes in $G$. By the definitions, it can be verified that

Table 1: Comparison between different algorithms for stochastic graphical bandits. $T$ is the number of rounds, $K$ is the number of arms, $\Delta_i$ is the reward gap of arm $i$, $v$ is an upper bound of the $(1+\epsilon)$-th moments, $\delta$ is the maximum degree in the feedback graph $G$, $\mathcal{C}$ is a clique cover of $G$, $N_C$ is a quantity related to clique $C \in \mathcal{C}$, $\Delta_C^{\max}$ is the maximum reward gap of $C$, $\Delta_C^{\min}$ is the minimum nonzero reward gap of $C$, $\Delta_{\max}$ is the maximum suboptimal reward gap, $S$ is a subset of the first $\alpha$ suboptimal arms with the ties broken arbitrarily and $\alpha$ is the *independence number* of $G$.

| Algorithm | Regret (bounded in $[0,1]$) | Regret (bounded $(1+\epsilon)$-th raw moments) | Regret (bounded $(1+\epsilon)$-th central moments) |
|---|---|---|---|
| UCB-N [Caron et al., 2012] | $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log T}{(\Delta_C^{\min})^2} + K)$ | \ | \ |
| UCB-NE [Hu et al., 2019] | $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^2} + \lvert\mathcal{C}\rvert)$ | \ | \ |
| UCB-LP [Buccapatnam et al., 2014] | $O(\sum_{i\in D} \frac{\log T}{\Delta_i} + K\delta)$ | \ | \ |
| AAE-AlphaSample [Cohen et al., 2016] | $O(\sum_{i\in S} \frac{\log T}{\Delta_i})$ | \ | \ |
| RUNE-TEM (Theorem 2) | $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^2} + \lvert\mathcal{C}\rvert)$ | $O(\sum_{C\in\mathcal{C}} \frac{v^{1/\epsilon}\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} + \sum_{C\in\mathcal{C}}\Delta_C^{\max})$ | \ |
| RUNE-MoM (Theorem 3) | $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^2} + \lvert\mathcal{C}\rvert)$ | \ | $O(\sum_{C\in\mathcal{C}} \frac{v^{1/\epsilon}\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} + \sum_{C\in\mathcal{C}}\Delta_C^{\max})$ |
| RAAE-TEM (Theorem 4) | $O(\sum_{i\in S} \frac{\log T}{\Delta_i})$ | $O(\sum_{i\in S} \frac{v^{1/\epsilon}\log T}{\Delta_i^{1/\epsilon}} + \Delta_{\max}\log T)$ | \ |
| RAAE-MoM (Theorem 5) | $O(\sum_{i\in S} \frac{\log T}{\Delta_i})$ | \ | $O(\sum_{i\in S} \frac{v^{1/\epsilon}\log T}{\Delta_i^{1/\epsilon}} + \Delta_{\max}\log T)$ |

$O_i(T) \geq T_i(T)$ holds for any feedback graph. Thus, the player can provide a more accurate estimate for the mean of each arm's reward distribution, by utilizing the side information of the feedback graph.

Before stating existing results, we introduce two standard graph-theoretic definitions [West, 2001], which will be used to describe regret bounds.

**Definition 1** *A clique in graph $G = (V, E)$ is a subset of vertices $C \subseteq V$ such that all arms in $C$ are neighbors with each other. A clique covering $\mathcal{C}$ of $G$ is a set of cliques such that $V = \cup_{C\in\mathcal{C}} C$. The clique covering number $\bar{\chi}(G)$ is the size of the smallest clique covering in $G$.*

**Definition 2** *An independent set in graph $G = (V, E)$ is a subset of vertices $S \subseteq V$ that are not connected by any edges with each other. Namely, $S$ is independent if for any $u, v \in S, u \neq v$, then $(u, v) \notin E$. The independence number $\alpha(G)$ is the size of the maximum independent set in $G$.*

Note that each node in a maximum independent set must consume one clique to cover, thus $\alpha(G) \leq \bar{\chi}(G)$ for any graph $G$, and the gap between them can be very large [Mannor and Shamir, 2011].

## 2.2 STOCHASTIC GRAPHICAL BANDITS

To fully exploit the side information of the feedback graph $G$, previous work [Caron et al., 2012, Buccapatnam et al., 2014, 2017, Cohen et al., 2016, Tossou et al., 2017, Liu et al., 2018a,b, Hu et al., 2019, Lykouris et al., 2020] has used the structural information of the feedback graph to characterize their regret bounds.

One category of classical methods for stochastic GB is based on UCB [Lai and Robbins, 1985, Agrawal, 1995, Auer et al., 2002]. For stochastic MAB with bounded rewards, Auer et al. [2002] propose UCB1 according to the principle of Optimism in the Face of Uncertainty (OFU), which attains an optimal regret bound of $O(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i} + \sum_{i=1}^K \Delta_i)$. Afterward, Caron et al. [2012] extend UCB1 [Auer et al., 2002] to UCB-N for stochastic GB with bounded rewards, where the main improvement is to update the estimated values of not only the chosen arm but also its neighbors at each round. They show that UCB-N attains an $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log T}{(\Delta_C^{\min})^2} + \sum_{i=1}^K \Delta_i)$ gap-based regret bound. In addition, they present an $\Omega(\log T)$ regret lower bound for this setting. Later, Hu et al. [2019] modify the index of UCB-N to enlarge the exploration phase and improve the bound to $O(\sum_{C\in\mathcal{C}} \frac{\Delta_C^{\max}\log(N_C T)}{(\Delta_C^{\min})^2} + \sum_{C\in\mathcal{C}}\Delta_C^{\max})$, where $N_C = \max_{i\in C} |N(i)|^{\frac{1}{4}}$ is determined by the maximum degree of clique $C \in \mathcal{C}$. Note that the sum of gap-based quantities in this bound is taken over the *clique covering* of the feedback graph, instead of the whole arm set.

Besides, there is another class of algorithms for stochastic GB based on the elimination technique [Even-Dar et al., 2006, Auer and Ortner, 2010]. Buccapatnam et al. [2014, 2017] propose a strategy termed UCB-LP, which leverages a linear programming (LP) induced by the feedback graph to explicitly guide the exploration stage. UCB-LP obtains an $O(\sum_{i\in D} \frac{\log T}{\Delta_i} + K\delta)$ gap-based regret bound, where $D$ is a particularly selected dominated set of $G$ (i.e., every node in the graph is either in $D$ or has at least one neighbor in $D$) and $\delta$ is the maximum degree in the feedback graph. Furthermore, they established an LP-based lower bound, which is also logarithmic with respect to $T$. Later, Cohen et al. [2016] consider a harder setting where the feedback graph may be

directed, time-variant, and not entirely revealed to the player. They propose an elimination-based algorithm and obtain an $O(\sum_{i\in S}\frac{\log T}{\Delta_i})$ gap-based regret bound, where $S$ is the set of the $\alpha_{\max}\log K$ arms with the smallest gap and $\alpha_{\max}$ is an upper bound of the *independence number* of the feedback graph over $T$ rounds. Recently, Lykouris et al. [2020] propose a novel layering technique by using the *independent set* for sampling and derive a similar $O(\sum_{i\in I}\frac{\log^2 T}{\Delta_i})$ regret bound for UCB-N, where $I$ is any independent set of $G$. In addition, other work [Tossou et al., 2017, Liu et al., 2018a,b, Hu et al., 2019, Lykouris et al., 2020] applies Thompson Sampling [Thompson, 1933] to stochastic GB and provides the corresponding theoretical guarantees.

### 2.3 STOCHASTIC HEAVY-TAILED BANDITS

Liu and Zhao [2011] are the first to investgate stochastic MAB with heavy-tailed rewards. In particular, they consider reward distributions with finite moments of order $1+\epsilon$ for some $\epsilon \in (0,1]$. They propose an algorithm based on a deterministic sequencing of exploration and exploitation, which attains a polynomial regret of $O(T^{\frac{1}{1+\epsilon}})$. In a subsequent work, Bubeck et al. [2013] design a framework termed Robust UCB by replacing the empirical mean in UCB1 [Auer et al., 2002] with more refined robust estimators, such as truncated empirical mean or median of means. They obtain the first gap-based logarithmic regret of $O(\sum_{i:\Delta_i>0}(\frac{v}{\Delta_i})^{\frac{1}{\epsilon}}\log T+\sum_{i=1}^{K}\Delta_i)$, where $v$ is an upper bound of the $(1+\epsilon)$-th moments. For stochastic MAB with finite variances ($\epsilon=1$), this regret bound recovers the optimal regret under the bounded or sub-Gaussian assumption [Lai and Robbins, 1985, Auer et al., 2002]. Besides, Bubeck et al. [2013] also provide a matching lower bound of $O(\Delta_i^{-\frac{1}{\epsilon}}\log T)$. Later, Medina and Yang [2016] extend the results to stochastic linear bandits with infinite action sets. They design two algorithms both with sublinear regret bounds, which are subsequently improved to be nearly optimal by Shao et al. [2018]. Recently, robust estimators are applied by Xue et al. [2020] to design algorithms for stochastic linear bandits with finite action sets, and nearly optimal sublinear regret bounds are established. For other settings, robust estimators are also employed by Lu et al. [2019], Tao et al. [2022] to design algorithms for stochastic lipschitz bandits with heavy-tailed rewards and stochastic MAB with heavy-tailed rewards in the (local) differential privacy model, respectively. In addition, heavy-tailed distributions have been extensively studied in the offline setting [Brownlees et al., 2015, Hsu and Sabato, 2016, Zhang and Zhou, 2018].

## 3 MAIN RESULTS

We first propose two UCB-type algorithms termed RUN and RUNE for stochastic graphical bandits with heavy-tailed rewards. Next, we present another elimination-based algorithm named RAAE with an improved regret bound. All the technical lemmas and proofs are deferred to the supplementary due to the space limitation.

**Algorithm 1** RUN-TEM

**Input:** Graph $G=(V;E)$, $\epsilon\in(0,1]$, $(1+\epsilon)$-th raw moment bound $v$, confidence level $\delta\in(0,1)$
**Initialize:** Set $O_i(0)=0$ for each arm $i\in V$. Let $\widehat{\mu}_i(t)$ be the estimate mean value based on the first $s$ observed values $X_{i,1},\cdots,X_{i,s}$ of arm $i$ up to time $t$
**for** $t>1$ **do**
  Pull arm
$$I_t=\operatorname*{argmax}_{i\in V}\mathrm{UCB}_i(t),$$
  where $\mathrm{UCB}_i(t)$ is computed by (8)
  Receive reward $X_{I_t,t}$ and observe rewards $X_{k,t}$ ($k\in N(I_t)$)
  **for** arm $k\in N(I_t)$ **do**
    $O_k(t)=O_k(t-1)+1$
    Compute the truncation level $B_{k,t,\delta}$ by (3)
    Update the estimate value:
$$\widehat{\mu}_k(t)=\frac{O_k(t-1)\widehat{\mu}_k(t-1)+X_{k,t}\mathbb{I}_{\{|X_{k,t}|\le B_{k,t,\delta}\}}}{O_k(t-1)+1}$$
  **end for**
**end for**

### 3.1 ROBUST UCB STRATEGY WITH FEEDBACK GRAPH

The basic idea behind existing algorithms for stochastic GB is to exploit the side information by sampling through the feedback graph. In this section, we begin with a simple strategy termed RUN and then present an improved algorithm named RUNE.

Following the seminal work of Caron et al. [2012], we propose Robust UCB-N (RUN) policy for stochastic GB with heavy-tailed rewards. Since the rewards of each arm no longer follow the sub-Gaussian distribution, their used empirical mean estimator can only provide polynomial deviations [Catoni, 2012]. To settle this issue, we employ RUN with Truncated Empirical Mean (TEM) estimator, which can guarantee exponential deviations for even heavy-tailed rewards [Bubeck et al., 2013]. The key idea of TEM is to truncate large rewards while computing the average value. Since truncation will bias the distribution, we cannot use a fixed truncation level uniformly over all time. Instead, we use an increasing truncation levels sequence for each arm $i\in V$:

$$B_{i,t,\delta}=\left(\frac{vO_i(t)}{\log(1/\delta)}\right)^{\frac{1}{1+\epsilon}}, \tag{3}$$

where $\delta \in (0,1)$ is a confidence level predetermined by the player. At each round $t$, we will compute the average truncated reward of each arm $i \in V$:

$$\widehat{\mu}_i(t) = \frac{\sum_{s=1}^{t} X_{i,s}\mathbb{I}_{\{|X_{i,s}|\le B_{i,s,\delta}\cap I_s\in N(i)\}}}{O_i(t)} , \quad (4)$$

which is updated incrementally in our algorithm to reduce the time complexity. Under the truncation level (3), we can obtain the concentration properties of TEM in the following proposition.

**Proposition 1** *Let $\delta \in (0,1), \epsilon \in (0,1]$ be positive parameters. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables sampling from fixed distribution $\mathcal{P}$ with finite mean $\mu$ and bounded $(1+\epsilon)$-th raw moments, i.e., $\mathbb{E}_{X\sim\mathcal{P}}[|X|^{1+\epsilon}] \le v$. Consider the TEM estimator*

$$\widehat{\mu}_T = \frac{1}{n}\sum_{t=1}^{n} X_t\mathbb{I}_{\{|X_t|\le B_{t,\delta}\}} , \quad (5)$$

*where $B_{t,\delta} = (\frac{vt}{\log(1/\delta)})^{\frac{1}{1+\epsilon}}$, then with probability at least $1-\delta$,*

$$\widehat{\mu}_T \ge \mu - 5v^{\frac{1}{1+\epsilon}}\left(\frac{\log(1/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} , \quad (6)$$

*and also, with probability at least $1-\delta$,*

$$\widehat{\mu}_T \le \mu + 5v^{\frac{1}{1+\epsilon}}\left(\frac{\log(1/\delta)}{n}\right)^{\frac{\epsilon}{1+\epsilon}} . \quad (7)$$

By using the concentration properties in Proposition 1, we construct an upper confidence bound based on the sum of average truncated reward and a confidence term:

$$\mathrm{UCB}_i(t) = \widehat{\mu}_i(t-1) + 5v^{\frac{1}{1+\epsilon}}\left(\frac{\log(1/\delta)}{O_i(t-1)}\right)^{\frac{\epsilon}{1+\epsilon}} , \quad (8)$$

where we take the convention $\sqrt{1/0} = +\infty$ so that all arms get observed at least once.

At each round $t$, following the principle of OFU, we first pull the arm $I_t$ with the maximum UCB index defined in (8) with ties broken arbitrarily. After that, we will receive reward $X_{I_t,t}$ of the selected arm and also observe rewards $X_{k,t}$ of all arm $k$ in its neighbor set $N(I_t)$. Finally, we update the observation number $O_k(t)$ and the estimate value $\widehat{\mu}_k(t)$ for all the arms $k \in N(I_t)$ by the truncation level defined in (3). The above procedure is summarized in Algorithm 1, and is referred to as RUN-TEM.

Finally, we establish the following expected regret bound for RUN-TEM.

**Theorem 1** *Let $G = (V; E)$, $\epsilon \in (0,1]$ and $v > 0$. Assume that the reward distributions $\mathcal{P}_i$ satisfy that,*

$$\mathbb{E}_{X\sim\mathcal{P}_i}\left[|X|^{1+\epsilon}\right] \le v \ (\forall i \in V) , \quad (9)$$

*then the expected regret of Algorithm 1 (RUN-TEM) with $\delta = \frac{1}{t^4}$ after $T$ steps is upper bounded by*

$$\mathbb{E}[R_T] \le \inf_{\mathcal{C}}\left\{40\left(\sum_{C\in\mathcal{C}}\frac{(10v)^{1/\epsilon}\Delta_C^{\max}}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}}\right)\log T\right\} + \left(1+\frac{\pi^2}{3}\right)\sum_{i=1}^{K}\Delta_i , \quad (10)$$

*where $\Delta_C^{\min} := \min_{i\in C\setminus\{i^*\}}\Delta_i$ is the minimum nonzero reward gap in clique $C$ and $\Delta_C^{\max} := \max_{i\in C}\Delta_i$ is the maximum reward gap in clique $C$.*

**Remark.** If we choose $\mathcal{C}$ as the trivial covering $\{\{i\} : i \in V\}$, the above regret bound reduces exactly to

$$40\sum_{i\in V:\Delta_i>0}\left(\frac{10v}{\Delta_i}\right)^{\frac{1}{\epsilon}}\log T + \left(1+\frac{\pi^2}{3}\right)\sum_{i=1}^{K}\Delta_i , \quad (11)$$

which matches the optimal regret bound for heavy-tailed MAB proved by Bubeck et al. [2013]. Moreover, if $G$ is a complete graph, then the whole graph constitute a clique covering and we further obtain the regret bound

$$40\left(\frac{(10v)^{1/\epsilon}\Delta_{\max}}{\Delta_{\min}^{(1+\epsilon)/\epsilon}}\right)\log T + \left(1+\frac{\pi^2}{3}\right)\sum_{i=1}^{K}\Delta_i , \quad (12)$$

which is a substantial improvement over the regret bounds of Robust UCB [Bubeck et al., 2013] since the leading term is independent of $K$. Except for these two extremes, if we choose more proper $\mathcal{C}$, the regret bound of RUN-TEM can be also improved significantly compared to the standard regret bounds of heavy-tailed MAB, since we make effective utilization on the side information of $G$.

In addition, when the rewards are generated from distributions with finite variances ($\epsilon = 1$), RUN-TEM yields regret bound

$$\mathbb{E}[R_T] \le \inf_{\mathcal{C}}\left\{40\left(\sum_{C\in\mathcal{C}}\frac{\sqrt{10v}\Delta_C^{\max}}{(\Delta_C^{\min})^2}\right)\log T\right\} + \left(1+\frac{\pi^2}{3}\right)\sum_{i=1}^{K}\Delta_i , \quad (13)$$

which enjoys the same order as the regret bound of UCB-N [Caron et al., 2012]. However, when the rewards are generated from distributions with infinite variances ($0 < \epsilon < 1$) [Shao and Nikias, 1993], the theoretical results of UCB-N are no longer applicable, while our method still enjoys a gap-based regret bound (10) with the leading term scales logarithmically with $T$.

Although RUN-TEM can obtain a graph-based logarithmic regret bound, the second term in (10) is still in the order of $O(K)$. To settle this issue, we further design Robust

UCB-NE (RUNE) strategy with an improved regret bound. Inspired by Hu et al. [2019], we embed the side information of the feedback graph into the principle of OFU and redefine a graph-based UCB index for each arm $i$ to enlarge the exploration stage properly:

$$\widehat{\mu}_i(t-1) + 5v^{\frac{1}{1+\epsilon}} \left( \frac{\log(|N(i)|/\delta)}{O_i(t-1)} \right)^{\frac{\epsilon}{1+\epsilon}} , \tag{14}$$

where $|N(i)|$ is the size of arm $i$'s neighbor set and $\widehat{\mu}_i(t-1)$ is computed by the TEM estimator with an altered truncation levels sequence

$$B_{i,t,\delta} = \left( \frac{vO_i(t)}{\log(|N(i)|/\delta)} \right)^{\frac{1}{1+\epsilon}} . \tag{15}$$

Except the above two parameters, other procedures follow the same as Algorithm 1, and this policy is called RUNE-TEM. Finally, we obtain a regret bound with constant terms taking sum over the *clique covering* of the feedback graph and logarithmic in the size of the cliques, which is summarized in the following theorem.

**Theorem 2** *Consider the same preconditions as Theorem 1. Let $\delta = \frac{1}{t^4}$, then the expected regret of RUNE-TEM after $T$ steps is upper bounded by*

$$\begin{aligned} \mathbb{E}[R(T)] \leq \inf_{\mathcal{C}} \Bigg\{ & 40 \left( \sum_{C \in \mathcal{C}} \frac{(10v)^{1/\epsilon} \Delta_C^{\max}}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} \right) \log T \\ & + \sum_{C \in \mathcal{C}} \left[ \left( \frac{40(10v)^{1/\epsilon} \Delta_C^{\max}}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} \right) \log N_C \right. \\ & \left. + \left( 1 + \frac{\pi^2}{3} \right) \Delta_C^{\max} \right] \Bigg\} , \end{aligned} \tag{16}$$

*where $\Delta_C^{\min} = \min_{i \in C \setminus \{i^\star\}} \Delta_i, \Delta_C^{\max} = \max_{i \in C} \Delta_i$, and $N_C = \max_{i \in C} |N(i)|^{\frac{1}{4}}$ is determined by the maximum degree of clique $C \in \mathcal{C}$.*

**Remark.** Here, we provide a discussion about the difference between RUN-TEM and RUNE-TEM. Given the same feedback graph $G$, the leading term of RUN-TEM and RUNE-TEM is the same. However, the constant term of RUN-TEM is in order $O(\sum_{i=1}^{K} \Delta_i)$ while RUNE-TEM improves it to $O(\sum_{C \in \mathcal{C}} [\frac{\Delta_C^{\max} \log(N_C)}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}} + \Delta_C^{\max}])$, where the sum is only taken over the *clique covering* of $G$, not all $K$ arms. As a result, RUNE-TEM can obtain a promotion in the constant term when the clique size is large.

Note that RUNE-TEM can only be applied to the rewards distributions with bounded $(1+\epsilon)$-th raw moments, which means that the selected arms may change along with the synchronized shift of all the reward distributions. Thus, it would be more desirable to obtain a regret bound in terms of the centered moments bound. To address this problem, we employ RUNE with Median of Means (MoM) estimator [Alon et al., 1999], and result in a translation-invariant algorithm termed RUNE-MoM. The main idea is to first divide the rewards $X_{i,1}, \cdots, X_{i,n}$ of each arm $i \in V$ into $k$ various disjoint blocks with size $N = \lceil n/k \rceil$:

$$\mathcal{X}_i = \{ X_{i,1:N}, \cdots, X_{i,((k-1)N+1):n} \} . \tag{17}$$

After that, we compute separately the standard empirical mean of each block $s \in [k]$ by

$$\widehat{\mu}_s = \frac{1}{N} \sum_{t=(s-1)N+1}^{sN} X_t . \tag{18}$$

Finally, we acquire the mean estimate value $\widehat{\mu}_i(t)$ by taking a median value of these empirical means within each block:

$$\widehat{\mu}_i(t) = \text{median}(\widehat{\mu}_1, \cdots, \widehat{\mu}_k) . \tag{19}$$

For a particular arm set the block size k as following

$$k = \lceil 8 \log(|N(i)| e^{-1/8}/\delta) \rceil , \tag{20}$$

where $\delta \in (0,1)$ is a confidence level predetermined by the player, we can obtain the properties of MoM described in the following proposition.

**Proposition 2** *Let $\delta \in (0,1), \epsilon \in (0,1]$ be positive parameters. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables sampling from fixed distribution $\mathcal{P}$ with finite mean $\mu$ and bounded $(1+\epsilon)$-th central moments, i.e., $\mathbb{E}_{X \sim \mathcal{P}}[|X - \mu|^{1+\epsilon}] \leq v$. Let $k = \lceil 8 \log(1/\delta) \rceil$, $N = \lceil n/k \rceil$,*

$$\widehat{\mu}_1 = \frac{1}{N} \sum_{t=1}^{N} X_t, \cdots, \widehat{\mu}_k = \frac{1}{N} \sum_{t=(k-1)N+1}^{kN} X_t , \tag{21}$$

*be $k$ empirical mean estimates, where each one is computed on $N$ rewards. Consider the MoM estimator*

$$\widehat{\mu}_M = \text{median}(\widehat{\mu}_1, \cdots, \widehat{\mu}_k) , \tag{22}$$

*then with probability at least $1-\delta$,*

$$\widehat{\mu}_M \geq \mu - (12v)^{\frac{1}{1+\epsilon}} \left( \frac{8 \log(e^{1/8}/\delta)}{n} \right)^{\frac{\epsilon}{1+\epsilon}} , \tag{23}$$

*and also, with probability at least $1-\delta$,*

$$\widehat{\mu}_M \leq \mu + (12v)^{\frac{1}{1+\epsilon}} \left( \frac{8 \log(e^{1/8}/\delta)}{n} \right)^{\frac{\epsilon}{1+\epsilon}} . \tag{24}$$

Through the concentration properties in Proposition 2, we can redefine a graph-based UCB index for RUNE by

$$\widehat{\mu}_i(t-1) + (12v)^{\frac{1}{1+\epsilon}} \left( \frac{8 \log(|N(i)|/\delta)}{n} \right)^{\frac{\epsilon}{1+\epsilon}} . \tag{25}$$

Finally, we obtain the regret upper bound of RUNE-MoM as following.

**Theorem 3** *Let $G = (V; E)$, $\epsilon \in (0, 1]$ and $v > 0$. Assume that the reward satisfy the distributions $\mathcal{P}_i$ with mean $\mu_i$ such that*

$$\mathbb{E}_{X\sim\mathcal{P}_i}\left[|X-\mu_i|^{1+\epsilon}\right] \le v\ (\forall i \in V)\ , \tag{26}$$

*then the expected regret of RUNE-MoM with $\delta = \frac{1}{t^4}$ after $T$ steps is upper bounded by*

$$\begin{aligned}\mathbb{E}[R(T)] \le \inf_{\mathcal{C}} \Bigg\{ & 64\left(\sum_{C\in\mathcal{C}} \frac{(24v)^{1/\epsilon}\Delta_C^{\max}}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}}\right)\log T \\ & + \sum_{C\in\mathcal{C}}\left[\left(\frac{64(24v)^{1/\epsilon}\Delta_C^{\max}}{(\Delta_C^{\min})^{(1+\epsilon)/\epsilon}}\right)\log N_C \right. \\ & \left. + \left(1+\frac{e^{1/8}\pi^2}{3}\right)\Delta_C^{\max}\right]\Bigg\},\end{aligned} \tag{27}$$

*where $\Delta_C^{\min} = \min_{i\in C\setminus\{i^\star\}} \Delta_i, \Delta_C^{\max} = \max_{i\in C} \Delta_i$, and $N_C = \max_{i\in C} |N(i)|^{\frac{1}{4}}$ is determined by the maximum degree of clique $C \in \mathcal{C}$.*

**Remark.** Note that the theoretical guarantee of RUNE-MoM is in the same order as RUEN-TEM. However, the regret bound of RUEN-TEM depends on the raw moment bound while the regret bound of RUNE-MoM depends on the central moment bound, which is translation invariant under a synchronized shift of all the reward distributions.

### 3.2 ROBUST ACTIVE ARM ELIMINATION WITH FEEDBACK GRAPH

Since the expected regret of RUNE is bounded by a sum of gap-based quantities over the *clique covering* of $G$, which may be unacceptable when the *clique covering* gets too large. Thus, a question arises here is whether it is possible to further improve the regret. We answer this question affirmatively by designing an elimination-based strategy

Inspired by AAE-AlphaSample [Cohen et al., 2016], we propose Robust Active Arm Elimination (RAAE), described in Algorithm 2, where the main idea is to sample each arm a minimal number of times and eliminate the "bad" arms one by one. To reduce the sampling times of each epoch, we firstly select a maximal independent set from the sub-graph induced by the active arm set and then play the arms in it once. Although Cohen et al. [2016] consider a similar setting, their theoretical results cannot be applied directly to stochastic GB with heavy-tailed rewards. To settle this issue, we adopt TEM or MoM estimator in RAAE and estimate the reward mean of each arm by a graph-based sampling mechanism.

We begin with RAAE equipped by TEM estimator. RAAE-TEM works in epochs $r = 1, 2, \cdots$. At each epoch $r$, the player maintains an active arm set $V_r$, initialized by $V_1 = V$, and selects a maximal independent set $I_r$ greedily from the subgraph induced by $V_r$. After that, the player pulls the arms

**Algorithm 2** RAAE-TEM

**Input:** Graph $G = (V, E)$ with $K$ nodes, $\epsilon \in (0, 1]$, $(1 + \epsilon)$-th raw moment bound $v$, number of rounds $T$
**Initialize:** $r \leftarrow 1, t \leftarrow 1, V_1 \leftarrow V, \varepsilon_1 \leftarrow 1/4^\epsilon$
**while** $|V_r| > 1$ **and** $t \le T$ **do**
  Select a maximal independent set $I_r$ greedily from the subgraph induced by set $V_r$
  Compute the sampling times $n_r$ by (30)
  **for** $s = 1$ **to** $n_r$ **do**
    **for** all $i \in I_r$ **do**
      Pull arm $i$ and receive reward $X_{i,t}$
      Observe rewards of all arms in $N(i)$
      Update $O_j(t)$ and $\widehat{\mu}_j(t)$ for all arms $j \in N(i)$ with the truncated level $B_{j,t}$ described in (28)
      $t \leftarrow t + 1$
    **end for**
  **end for**
  Compute $\widehat{\mu}_r^\star = \max_{i\in V_r} \widehat{\mu}_i(t-1)$
  Execute active arm elimination described by (29)
  $\varepsilon_{r+1} \leftarrow \varepsilon_r/2^\epsilon, r \leftarrow r + 1$
**end while**
Play the arm left in $V_r$ until $T$ rounds have passed

in $I_r$ once to update the average truncated rewards $\widehat{\mu}_i(t)$ of all the arms $i \in V_r$, with truncation levels

$$B_{i,t} = \left(\frac{vO_i(t)}{\log(2KT)}\right)^{\frac{1}{1+\epsilon}}\ . \tag{28}$$

Then, we will eliminate the arms in $V_r$ that are known to be sub-optimal with sufficient confidence:

$$V_{r+1} = \{i \in V_r : \widehat{\mu}_i(t-1) \ge \widehat{\mu}_r^\star - 2\varepsilon_r\}\ , \tag{29}$$

where $\widehat{\mu}_r^\star = \max_{i\in V_r} \widehat{\mu}_i(t-1)$ and $\varepsilon_r$ is the accuracy parameter, initialize by $1/4^\epsilon$. As the analysis will show, by repeating this process for sampling times

$$n_r = \left\lceil \frac{(5(5v)^{1/\epsilon}\log(2KT)}{\varepsilon_r^{(1+\epsilon)/\epsilon}} \right\rceil\ , \tag{30}$$

the mean rewards of all arms in $V_r$ can be estimated within $\varepsilon_r$ accuracy. As a result, each suboptimal arm $i$ with $\Delta_i > 4\varepsilon_r$ will be eliminated with high probability at each epoch $r$. Thus, we multiply $\varepsilon_r$ by $1/2^\epsilon$ after each epoch to increase the estimation accuracy. Finally, we obtain the following regret bound of RAAE-TEM.

**Theorem 4** *Assume $K \ge 2$ and $T \ge K$. Suppose that the independence number of feedback graph $G = (V, E)$ is at most $\alpha$, and the reward distributions $\mathcal{P}_i$ satisfy that,*

$$\mathbb{E}_{X\sim\mathcal{P}_i}\left[|X|^{1+\epsilon}\right] \le v\ (\forall i \in V)\ , \tag{31}$$

*then the expected regret of Algorithm 2 (RAAE-TEM) after $T$ steps is at most*

$$\mathbb{E}[R_T] \leq O\left(\sum_{i \in V^{(\alpha)}} \frac{v^{1/\epsilon}}{\Delta_i^{1/\epsilon}} \log T + \Delta_{\max} \log T\right), \quad (32)$$

*where $\Delta_{\max} = \max_{i \in V} \Delta_i$, $V^{(\alpha)}$ denotes a subset of the first $\alpha$ suboptimal arms with ties broken arbitrarily.*

**Remark.** If $G$ is a complete graph, then $\alpha = 1$ and the above regret bound reduces to

$$O\left(\frac{\log T}{\Delta_{\min}^{1/\epsilon}} + \Delta_{\max} \log T\right), \quad (33)$$

which is independent of $K$. Inversely, if $G$ is an empty graph, which means that $E = \emptyset$, then $\alpha = K$ and the above regret is on the order of $O(\sum_{i:\Delta_i>0}(\frac{v}{\Delta_i})^{\frac{1}{\epsilon}} \log T)$ which has been proved optimal for heavy-tailed MAB [Bubeck et al., 2013]. Except for these two extremes, the regret bound of RUN-TEM can be improved significantly compared to Robust UCB [Bubeck et al., 2013] when $\alpha < K$.

We provide a discussion about the difference between RAAE-TEM and RUNE-TEM. Note that the regret bound of RUNE-TEM (16) is summed over the *clique covering* of $G$ and the gap-based quantities rely on the ratio of the maximum and minimum mean reward gaps within each clique, which can be quite large in the worst case. However, the regret bound of RAAE-TEM (32) is summed over the subset of $\alpha$ arms with the smallest nonzero gaps and the gap-based quantities only rely on the reciprocal of the mean reward gaps. As $\alpha$ is much smaller than the size of the *clique covering* for benign graphs, we conclude that the regret bound of RAAE-TEM is tighter than RUNE-TEM in this case.

Furthermore, we can employ RAAE with MoM estimator to process heavy-tailed rewards with bounded $(1+\epsilon)$-th central moments. By using Proposition 2, we reselect block number $k = \lceil 8 \log(2KT) \rceil$, block size $N = \lceil n/k \rceil$ and sampling times

$$n_r = \left\lceil \frac{(8(12v)^{1/\epsilon} \log(2e^{1/8}KT))}{\varepsilon_r^{(1+\epsilon)/\epsilon}} \right\rceil, \quad (34)$$

where $\varepsilon_r$ is the same as that in RAAE-TEM. Finally, we obtain the following regret bound of RAAE-MoM.

**Theorem 5** *Assume $K \geq 2$ and $T \geq K$. Suppose that the independence numbers of feedback graph $G = (V, E)$ is at most $\alpha$, and the reward distributions $\mathcal{P}_i$ with mean $\mu_i$ satisfy that,*

$$\mathbb{E}_{X \sim \mathcal{P}_i}\left[|X - \mu_i|^{1+\epsilon}\right] \leq v \ (\forall i \in V), \quad (35)$$

*then the expected regret of RAAE-MoM after $T$ steps is at most*

$$\mathbb{E}[R_T] \leq O\left(\sum_{i \in V^{(\alpha)}} \frac{v^{1/\epsilon}}{\Delta_i^{1/\epsilon}} \log T + \Delta_{\max} \log T\right), \quad (36)$$

*where $\Delta_{\max} = \max_{i \in V} \Delta_i$, $V^{(\alpha)}$ denotes a subset of the first $\alpha$ suboptimal arms with ties broken arbitrarily.*

# 4 EXPERIMENTS

In this section, we present numerical results to demonstrate the effectiveness of our algorithms. We compare our methods (RUNE-TEM, RAAE-TEM)[1] with UCB-N [Caron et al., 2012] and AAE-AlphaSample [Cohen et al., 2016].

**Setup.** We synthesize a stochastic GB problem with $K = 30$, there are 2 optimal arms assigned uniformly at random from $[K]$ and all other arms are sub-optimal. The means of the optimal rewards are set to $\mu^\star = 1.0$ and the means of sub-optimal rewards are restricted to $(0, 1.0)$. The time horizon is set as $T = 10000$ for all experiments, and we take the average of 10 independent runs of each algorithm.

**Reward Distribution.** To generate heavy-tailed rewards, we consider Pareto random variable $X$ with shape parameters $\alpha$ and scale parameter $x_m$, whose probability density function can be written as following

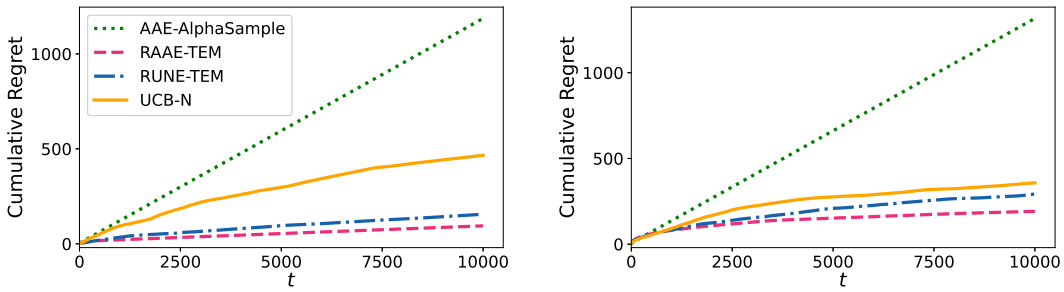
$$f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m \\ 0, & x < x_m \end{cases}. \quad (37)$$

In our experiments, we make $\alpha > 1$, such that the expectation exists and can be computed in the form $\mathbb{E}[X] = \frac{\alpha x_m}{\alpha - 1}$. Also, the $r$-th raw moments exists when $r < \alpha$ and can be calculated by the formula $\mathbb{E}[X^r] = \frac{\alpha x_m^r}{\alpha - r}$. We can verify that the smaller is $\alpha$, the heavier is the distribution tail. To guarantee that the $(1+\epsilon)$-th raw moment is bounded by some constant $v > 0$, we set $\alpha = 1.1 + \epsilon$, where $\epsilon = 0.3$. Each arm's rewards are sampling independently from a predetermined Pareto distribution with parameter $\alpha, x_m = \frac{\mu(\alpha-1)}{\alpha}$ and the rewards of any two arms in a given round are generated independently.

**Feedback Graph.** We conduct experiments on two fixed undirected graphs, One is a random graph, generated by the Erdős-Rényi model [Erdős and Rényi, 1960]. In details, we represent the edges by a random matrix $E \in \{0, 1\}^{K \times K}$, and assign $E_{ij} = 1$ $(i \neq j)$ with a fixed probability $p$ and $E_{ii} = 1$ for all $i \in [K]$. The other is a deterministic graph constructed by Lu et al. [2021] with $K = 30$, $\alpha = 10$ and $\bar{\chi} = 14$, which is illustrated in Fig. 2.

**Results.** We present two results in Fig. 1. As can be seen, the regret curves of elimination-based methods increase at the beginning and then maintain stable after some epochs, because they can find the best arms with high probability. Furthermore, RAAE-TEM performs better than AAE-AlphaSample in both settings, which is expected since it can use more refined robust estimators to improve the estimated accuracy. Also, RUNE-TEM suffers smaller regret

[1]Code will be made available at https://github.com/yutian-007/graphical-bandits-with-heavy-tailed-rewards/.




(a) Random Graph

(b) Deterministic Graph

Figure 1: Comparison of our algorithms (RUNE-TEM, RAAE-TEM) versus UCB-N and AAE-AlphaSample for stochastic GB with heavy-tailed rewards

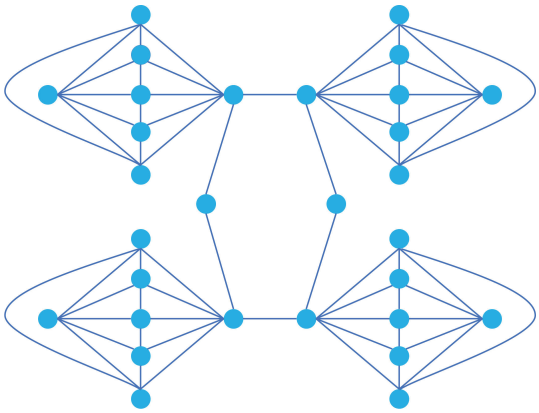


Figure 2: Illustration of the Deterministic Graph [Lu et al., 2021]

than UCB-N, since it has a preferable regret bound. Particularly, RAAE-TEM performs better than RUNE-TEM, which is consistent with our theoretical analysis.

## 5 CONCLUSION AND FUTURE WORK

We design two novel algorithms for stochastic graphical bandits with heavy-tailed rewards, which only require the existence of the $(1+\epsilon)$-th moments for some $\epsilon \in (0, 1]$. One of our algorithms is based on UCB strategy and obtains regret bounds depending on a sum of gap-based quantities over the *clique covering* of the feedback graph. The other one is based on successive elimination technique and enjoys an improved regret bound depending on a gap-based sum with size controlled by $\alpha$, which is smaller than the size of the *clique covering* for benign graphs. To the best of our knowledge, we provide the first regret bounds for stochastic GB with heavy-tailed rewards. Thus, a natural and challenging open problem is whether one can prove a lower bound for this setting. Obtaining lower bounds seems highly non-trivial even for stochastic GB under the sub-Gaussian setting [Marinov et al., 2022b], and we leave it as a future work.


### Acknowledgements

This work was partially supported by NSFC (62122037), and JiangsuSF (BK20200064), and the Fundamental Research Funds for the Central Universities (2023300246).


### References


Rajeev Agrawal. Sample mean based index policies by $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

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

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

Christian Brownlees, Emilien Joly, and Gábor Lugosi. Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics*, 43(6), 2015.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. In *Proceedings of the 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, page 289–300, 2014.

Swapna Buccapatnam, Fang Liu, Atilla Eryilmaz, and Ness B. Shroff. Reward maximization under uncertainty: Leveraging side-observations on networks. *Journal of Machine Learning Research*, 18(216):1–34, 2017.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, page 142–151, 2012.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré. Probabilités et statistiques*, 48(4):1148–1185, 2012.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 811–819, 2016.

P. Erdős and A Rényi. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Serguei Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, 2011.

Benjamín Gutiérrez, Loïc Peter, Tassilo Klein, and Christian Wachinger. A multi-armed bandit to smartly select a training set from big medical data. In *Proceedings of the 20th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 38–45, 2017.

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

Daniel J. Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17:18:1–18:40, 2016.

Bingshan Hu, Nishant A. Mehta, and Jianping Pan. Problem-dependent regret bounds for online learning with feedback graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, pages 852–861, 2019.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems 28*, pages 1450–1458, 2015.

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

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge university press, 2020.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 477–486, 2014.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, page 661–670, 2010.

Fang Liu, Swapna Buccapatnam, and Ness B. Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 3643–3650, 2018a.

Fang Liu, Zizhan Zheng, and Ness B. Shroff. Analysis of thompson sampling for graphical bandits without the graphs. In *Proceedings of the 34th Conference on Uncertainty in Artificial*, pages 13–22, 2018b.

Keqin Liu and Qing Zhao. Multi-armed bandit problems with heavy-tailed reward distributions. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, pages 485–492, 2011.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for Lipschitz bandits with heavy-tailed rewards. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4154–4163, 2019.

Shiyin Lu, Yao Hu, and Lijun Zhang. Stochastic bandits with graph feedback in non-stationary environments. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 8758–8766, 2021.

Thodoris Lykouris, Éva Tardos, and Drishti Wali. Feedback graph regret bounds for Thompson Sampling and UCB. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pages 592–614, 2020.

Aniket Mahanti, Niklas Carlsson, Anirban Mahanti, Martin F. Arlitt, and Carey Williamson. A tale of the tails: Power-laws in internet measurements. *IEEE Network*, 27 (1):59–64, 2013.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems 24*, pages 684–692, 2011.

Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. *CoRR*, abs/2206.10022, 2022a.

Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Open problem: Finite-time instance dependent optimality for stochastic online learning with feedback graphs. In *Proceedings of 35th Conference on Learning Theory*, pages 5644–5649, 2022b.

Andres Muñoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1642–1650, 2016.

Jaimie Yejean Park, Kyoung-Won Lee, Sangyeon Kim, and Chin-Wan Chung. Ads by whom? ads about what?: exploring user influence and contents in social advertising. In *Proceedings of the 1st ACM Conference on Online Social Networks*, pages 155–164, 2013.

Svetlozar T. Rachev. Handbook of heavy tailed distributions in finance. In *Elsevier Monographs*, 2003.

Herbert E. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

Eric M. Schwartz, Eric T. Bradlow, and Peter S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4): 500–522, 2017.

Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. In *Advances in Neural Information Processing Systems 31*, pages 8430–8439, 2018.

Min Shao and Chrysostomos L. Nikias. Signal processing with fractional lower order moments: stable processes and their applications. *Proceedings of the IEEE*, 81(7): 986–1010, 1993.

Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 1546–1574, 2022.

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

Aristide C. Y. Tossou, Christos Dimitrakakis, and Devdatt P. Dubhashi. Thompson sampling for stochastic bandits with graph feedback. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 2660–2666, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science*, 30 (2):199 – 215, 2015.

Douglas B West. *Introduction to Graph Theory (2nd Edition)*, volume 2. Prentice Hall, Upper Saddle River, 2001.

Min Xu, Tao Qin, and Tie-Yan Liu. Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems 26*, pages 2400–2408, 2013.

Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 2936–2942, 2020.

Lijun Zhang and Zhi-Hua Zhou. $\ell_1$-regression with heavy-tailed distributions. In *Advances in Neural Information Processing Systems 31*, pages 1076–1086, 2018.