*Proceedings of the 2025 Winter Simulation Conference*
*E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.*

# MULTI-AGENT MULTI-ARMED BANDIT WITH FULLY HEAVY-TAILED DYNAMICS

Xingyu Wang[1] and Mengfan Xu[2]

[1]Quantitative Economics, University of Amsterdam, Amsterdam, NETHERLANDS
[2]Mechanical and Industrial Engineering, University of Massachusetts Amherst, Amherst, MA, USA

## ABSTRACT

We study decentralized multi-agent multi-armed bandits, where clients communicate over sparse random graphs with heavy-tailed degree distributions and observe heavy-tailed reward distributions with potentially infinite variance. We are the first to address such fully heavy-tailed scenarios, capturing the dynamics and challenges of communication and inference among multiple clients in real-world systems, and provide regret bounds that match or improve upon existing results developed in even simpler settings. Under homogeneous rewards, we exploit hub-like structures unique to heavy-tailed graphs to aggregate rewards and reduce noises when constructing UCB indices; under $M$ clients and degree distributions with power-law index $\alpha > 1$, we attain a regret (almost) of order $O(M^{1-\frac{1}{\alpha}} \log T)$. Under heterogeneous rewards, clients synchronize by communicating with neighbors and aggregating exchanged estimators in UCB indices; by establishing information delay bounds over sparse random graphs, we attain a $O(M \log T)$ regret.

## 1 INTRODUCTION

Multi-armed Bandit (MAB) is an online sequential decision-making framework where a decision maker, or a client, pulls an arm from a finite set of arms at each time step, receives the reward of the pulled arm, and aims to maximize the cumulative received reward, or equivalently, minimize the regret compared to always pulling the optimal arm (Auer et al. 2002; Auer et al. 2002). Recent advancements have focused on its multi-agent variant, namely Multi-agent Multi-armed Bandit (MA-MAB), capturing the complexity of networks of decision makers in real-world scenarios. A notable focus in on cooperative MA-MAB, where $M$ clients aim to optimize the regret of the entire network via communication, with respect to a globally optimal arm defined by the global reward averaged across all clients. In this work, we consider a general and widely applicable setting with decentralization, where clients communicate on a graph without the stringent assumption of a central server.

Unique to decentralized MA-MAB is the coupling between graphs and time in the sequential regime, where two clients can communicate only when there is an edge between them on the graph. Many studies have focused on time-invariant graphs (i.e., with edges remaining constant over time). However, the emergence of examples such as ad-hoc wireless networks (Roman et al. 2013) has motivated the recent interest in time-varying graphs. In particular, the random-graph setting allows graphs to be redrawn from a distribution at each time step. For instance, Xu and Klabjan (2024) has recently explored the classical light-tailed Erdős–Rényi graphs in MA-MAB, where the edge between each pair of clients is sampled under a Bernoulli distribution with constant probability. Notably, this model is symmetric and dense, with each client having exactly the same expected degree (i.e., the count of connected clients) of order $O(M)$. However, real-world tasks often involve uneven and sparse communications: the degree—representing the communication resource assigned to each client—can be rather heterogeneous among clients, and the resource for each individual may not scale with $M$. Enabling collaboration over such asymmetric and sparse graphs facilitates system-wide performance in cooperative MA-MAB, which depends on the collective performance of all clients rather than any single individual, and has profound impact on promoting fairness and enhancing social good. This setting, however, remains unexplored and leaves a significant research gap.

This work focuses on sparse random graphs with power-law heavy-tailed degree distributions that capture the highly asymmetric network dynamics in real-world tasks, where a few vertices play hub-like roles with many connections to others, while most vertices have limited degrees. Such power-law heavy tails prevail in real-world networks across numerous contexts, including finance and economics, transportation networks, online retailing, supply chains, social communications, and epidemiology, to name a few; see, e.g., (da Cruz and Lind 2013; Clancy, David Jr 2021; Hearnshaw and Wilson 2013; Kunegis 2013; Clauset et al. 2009; Vázquez et al. 2002; Pastor-Satorras and Vespignani 2002). Therefore, it is essential to address this research gap and build solid theoretical and algorithmic foundations for MA-MAB over sparse and heavy-tailed random graphs, thus enabling efficient and robust decision-making via time-evolving, uneven, and limited communication and coordination over a wide range of real-world networks.

Two aspects of the reward distributions are central to multi-agent MAB problems. First, if the expected reward of an arm is the same for all clients, it is categorized into homogeneous settings; otherwise, it is heterogeneous. While homogeneous MA-MAB is generally well understood, heterogeneous MA-MAB has recently gained attention and presents additional challenges. This is particularly true in the decentralized setting, as inferring the global optimal arm requires the reward information from all clients. While Erdős–Rényi (E-R) graphs have been addressed in (Xu and Klabjan 2024), their approach does not apply to our sparse and asymmetric setting. Second, the intensity of randomness (more specifically, tail behaviors) of reward distributions significantly impacts the complexity of the problem. Aside from the long-standing focus on sub-Gaussian rewards in the bandit community, there has been a recent interest in distributions with heavier tails, including the sub-Exponential class and Exponential families (Korda et al. 2013; Jia et al. 2021), and those with even heavier tails and infinite $p^{\text{th}}$ order moments (Bubeck et al. 2013; Vakili et al. 2013; Dubey and Pentland 2019; Tao et al. 2022). Inferring the mean value using reward observations under such extreme randomness becomes quite challenging, often necessitating the use of robust estimators in the algorithms. Notably, such efforts are mostly limited to single-agent MAB. A recent work (Dubey and Pentland 2020) addresses heavy-tailed rewards with $(1+\varepsilon)^{\text{th}}$-order moments in MA-MAB, but only considers the homogeneous-rewards setting and with time-invariant connected graphs. Considering heavy-tailed rewards in a more challenging and general setting with heterogeneous rewards and time-varying random graphs remains unexplored, a gap we address herein as well.

In this paper, we focus on the following question: *Can we formulate and solve the multi-agent multi-armed bandit problem with heavy-tailed random graphs and heavy-tailed rewards in both homogeneous and heterogeneous settings?*

### 1.1 Main Contributions

We hereby provide an affirmative answer to the research question through our contributions, elaborated as follows. We formulate the multi-agent multi-armed bandit problem over sparse, asymmetric, and heavy-tailed random graphs, and under rewards with potentially infinite variance. Specifically, we consider rank-1 inhomogeneous random graphs (Boguná and Pastor-Satorras 2003; Chung and Lu 2002) with heavy-tailed degree distributions, a standard setting in literature (e.g., (van der Hofstad et al. 2017; van der Hofstad et al. 2020)). In this framework, the probability for having an edge between a pair of clients at each time step is dictated by (normalization of) attraction weights of clients; under heavy-tailed weight distributions, some clients consistently play hub-like roles and often connect to many other clients. Moreover, the graphs we consider are much more sparse (with $O(1)$ expected degree for each client) compared to Erdos-Renyi graphs (with $O(M)$ expected degree), which translates to significantly reduced communication costs and is of broad interest in large-scale multi-agent learning problems.

Methodologically, we propose new algorithms for homogeneous and heterogeneous settings. In the homogeneous-reward setting, we characterize and exploit the notion of hubs exclusive to heavy-tailed graphs: clients over the hub communicate and aggregate rewards, achieving variance reduction proportional to the hub size, while other clients use delayed aggregation through a hub representative. This principle guides the design of our novel UCB index, which also incorporates the median-of-means estimator for robust

estimation under heavy-tailed rewards. In the heterogeneous setting, another challenge is asynchronization (differences in arm pulls) among clients due to variations in their reward distributions. To address this, clients use random sampling when asynchronization occurs, and deploy UCB-based strategies otherwise based on our newly constructed reward estimators. Specifically, we propose an aggregation method that integrates the most recent heavy-tailed reward information from all clients, introducing novel information update mechanisms.

We establish theoretical guarantees for the proposed algorithms via comprehensive regret analyses. In homogeneous settings, we obtain a regret upper bound that is (almost) of order $O(M^{1-\frac{1}{\alpha}}\log T)$, which is sublinear in $M$. This improves upon the potential linearity in (Dubey and Pentland 2020) and demonstrates sample complexity reduction even under sparse graph structures. Under heterogeneous rewards, we derive an upper bound of order $O(M\log T)$, extending the bound in (Xu and Klabjan 2024) to sparse and asymmetric graphs with heavy-tailed rewards. The results highlight the consistency and effectiveness of our approach.

### 1.2 Motivating Examples

We highlight that this work is motivated by several real-world applications that depart from traditional assumptions. For instance, we consider the sequential testing and development of advertising strategies across customers or centers in large-scale networks, where the underlying communication topology is often heavy-tailed (see, e.g., (Kunegis 2013)). Such network structures reflect the heterogeneous and highly skewed connectivity observed in practical marketing and social platforms.

In another example, within wireless ad-hoc networks in edge computing and digital infrastructure, coordination among multiple devices must occur under sparse and unevenly distributed communication patterns (see, e.g., (Besson and Kaufmann 2018)). These settings illustrate the challenges of decentralized decision-making in dynamic and bandwidth-limited environments, further motivating the need for topology-aware cooperative learning algorithms given heavy-tailed graphs herein.

## 2 PROBLEM SETUP

We start by setting notations that will be frequently used throughout this paper. Given a positive integer $k$, let $[k]=\{1,2,\ldots,k\}$. We adopt the convention that $[0]=\emptyset$. Given two sequences of non-negative real numbers $(x_n)_{n\geq 1}$ and $(y_n)_{n\geq 1}$, we say that $x_n=O(y_n)$ (as $n\to\infty$) if there exists some $C\in[0,\infty)$ such that $x_n\leq Cy_n\ \forall n\geq 1$. Besides, we say that $x_n=o(y_n)$ if $\lim_{n\to\infty}x_n/y_n=0$.

Let $M$ denote the number of clients, which are labeled by $[M]=\{1,2,\ldots,M\}$. At each time $t=1,2,\ldots$, the clients are distributed over an undirected graph $G_t=(V,E_t)$, where $V=[M]$, and $E_t$ is the set of edges (i.e. two clients $m,l$ communicate at time $t$ only if $(m,l)\in E_t$). Let $K$ be the number of arms. For each client $m\in[M]$, we denote the reward of arm $1\leq i\leq K$ at time $t$ by $r_i^m(t)$, which is an i.i.d. copy from a time-invariant distribution $F_i^m$ with mean value $\mu_i^m$. By a *homogeneous-reward setting*, we refer to the case where $\mu_i^m=\mu_i^l\ \forall m,l\in[M]$ holds for any arm $i$. Otherwise, we call it a *heterogeneous-reward setting*. Our Assumption 1 is sufficiently general to account for both heavy-tailed reward distributions (potentially with infinite variance) and light-tailed distributions (with finite moments of any order, e.g. sub-Gaussian and sub-exponential classes).

**Assumption 1** (Rewards with Uniformly Bounded $(1+\varepsilon)^{\text{th}}$ Central Moments) Given $m\in[M]$ and $i\in[K]$, rewards $(r_i^m(t))_{t\geq 1}$ are i.i.d. copies from the distribution $F_i^m$. Also, there exist $\varepsilon\in(0,1]$ and $\rho\in(0,\infty)$ such that $\sup_{m\in[M],\ i\in[K]}\mathbf{E}|r_i^m(1)-\mu_i^m|^{1+\varepsilon}\leq\rho$, where $\mu_i^m\triangleq\mathbf{E}r_i^m(1)=\int xF_i^m(dx)$ denotes expected rewards.

We use $a_m^t$ to denote the arm pulled by client $m$ at time $t$. We define the global reward of arm $i$ at each time step $t$ as $r_i(t)=\frac{1}{M}\sum_{m\in[M]}r_i^m(t)$, and the expected value of the global reward of arm $i$ by $\mu_i=\frac{1}{M}\sum_{m\in[M]}\mu_i^m$. We denote the *global optimal arm* by $i^*=\arg\max_{i\in[K]}\mu_i$, and consider the cooperative setting where all clients would, ideally, pull the globally optimal arm $i^*$. The optimality gap for arm $i$ is $\Delta_i=\mu_{i^*}-\mu_i$. This motivates the definition of the total regret of the system, which we call regret for

simplicity throughout the paper, by $R_T = M \cdot T \cdot \mu_{i^*} - \sum_{t=1}^{T} \sum_{m\in[M]} \mu_{a_m^t}$, measuring the difference in the cumulative expected reward between the global optimal arm and the actions.

The main objective of this paper is to develop a multi-agent MAB algorithm and minimize $R_T$ for clients given the sparse communications available on $(G_t)_{t\geq 1}$, and we are particularly interested in the case where the communication graphs $(G_t)_{t\geq 1}$ are time-varying with heavy-tailed degree distributions. More precisely, we consider the graph sequence $(G_t)_{t\geq 1}$ that are generated as follows. Independently for each time $t \geq 1$ and pair $m, l \in [M]$ with $m \neq l$, we have $(m,l) \in E_t$ with probability $P(h_m, h_l)$, and $(m,n) \notin E_t$ with probability $1 - P(h_m, h_n)$, where the kernel $P(\cdot,\cdot)$ is defined by

$$P(u,v) = \min\{1,\ uv/(\theta M)\}, \qquad \forall u, v \geq 0, \tag{1}$$

and $(h_m)_{m\geq 1}$ are i.i.d. copies of some random variable $h$ with mean $\theta = \mathbf{E}h$. Note that this is the standard rank-1 inhomogeneous random graphs; e.g., (Boguná and Pastor-Satorras 2003; Chung and Lu 2002). Intuitively speaking, $h_m$ is the weight assigned to the $m^{\text{th}}$ node, representing its *attraction* to the other nodes. The weight $h_m$ would not change with time $t$, and is close to the expected degree of the $m^{\text{th}}$ node (especially under large $M$) over the graphs $(G_t)_{t\geq 1}$. In particular, we consider the case where the weights $h_m$'s are heavy-tailed, and capture heavy tails using notion of regular variation. Given a measurable function $\phi : (0,\infty) \to (0,\infty)$, we say that $\phi$ is regularly varying as $x \to \infty$ with index $\beta$ (denoted as $\phi(x) \in \mathscr{RV}_\beta(x)$ as $x \to \infty$) if $\phi(x) = x^\beta \cdot l(x)$ for some function $l : (0,\infty) \to (0,\infty)$ with $\lim_{x\to\infty} l(tx)/l(x) = 1$ for all $t > 0$. That is, $\phi(x)$ roughly follows a power-law tail with index $\beta$. For a standard treatment on the properties of regularly varying functions, see, e.g., (Resnick 2007). We impose the following assumption on the attraction weight sequence $(h_m)_{m\geq 1}$.

**Assumption 2** (Heavy-Tailed Graph) $\mathbf{P}(h > x) \in \mathscr{RV}_{-\alpha}(x)$ for some $\alpha > 1$.

We impose the next assumption to exclude the pathological case that some clients are (almost) never connected to others at any time $t$.

**Assumption 3** (Lower Bound for $h$) There exists $c_h > 0$ such that $\mathbf{P}(h \geq c_h) = 1$.

We use $\mathscr{N}_m(t)$ to denote the neighborhood set of the $m^{\text{th}}$ node at time $t$, which includes all the other nodes that are connected to $m$ over the graph $G_t$. Equivalently, the graph $G_t$ can be represented by the adjacency matrix $(X_{m,l}^t)_{1\leq m,l\leq M}$ where $X_{m,l}^t = 1$ if there is an edge between nodes $m$ and $l$, and $X_{m,l}^t = 0$ otherwise. We set $X_{m,m}^t \equiv 1$ for any $1 \leq m \leq M$. We also define the empirical adjacency matrix by $P_t(m,l) \triangleq \sum_{s=1}^t X_{m,l}^s/t$ and $P_t = \big(P_t(m,l)\big)_{1\leq m,l\leq M}$. We note that each node $m$ only knows its own neighbors and can only observe the $m$-th row of $P_t$, i.e., node $m$ has access to $\{P_t(m,q)\}_{q=1}^M$ but not $\{P_t(n,q)\}_{q=1}^M$ for $n \neq m$.

## 3 ANALYSES OF RANDOM GRAPHS

In this section, we establish useful properties regarding the hub structures and information delay over random graphs $G_t$ defined in Section 2. The results lay the foundation for our subsequent analysis of multi-agent multi-armed bandits. Due to the page limit, we refer the readers to the online preprint (Wang and Xu 2025) for the rigorous proofs of all theoretical results in this paper.

### 3.1 Hubs on Heavy-Tailed Graphs

A feature exclusive to heavy-tailed graphs is the arise of hub-like nodes with disproportionately large degrees (i.e., being connected to a large number of nodes), which enables efficient communication among all clients through such hubs. We first consider a *deterministic* characterization of the hub. Let $\hat{m} = \arg\max_{m\in[M]} |\mathscr{N}_m(1)|$ be the client with the highest degree at time 1 (arbitrarily pick one if there are ties). Let $S_0^t \triangleq \{m \in [M] : (m,\hat{m}) \in E_t\}$ be the clients communicating with $\hat{m}$ at time $t$. Note that, for any $m \in [M]$ with $h_m > \theta M/h_{\hat{m}}$, by (1) we know that such $m$ must be *deterministically* (i.e., with probability 1) connected to $\hat{m}$ for all $t$. Lemma 1 confirms that, with high probability, such nodes $m$ are plenty. Specifically,

by standard techniques in extreme value theory for heavy-tailed variables, one can show that $h_{\hat{m}}$ is roughly of order $M^{1/\alpha}$, and (under $\alpha \in (1,2)$) the count of $m \in [M]$ with $h_m > \theta M/h_{\hat{m}} \approx \theta M^{1-1/\alpha}$ is roughly of order $M^{2-\alpha}$. By taking the $\zeta$-slackness in the power-law index (as demonstrated in Lemma 1), we are able to ensure the exponentially decaying bound for pathological cases.

**Lemma 1** Let Assumptions 2 and 3 hold with $\alpha \in (1,2)$. Given $\zeta \in (0, 2-\alpha)$, there exists $\gamma > 0$ such that $\mathbf{P}\big(|S_0| \le M^{2-\alpha-\zeta}\big) = o(\exp(-M^{\gamma}))$, as $M \to \infty$, where $S_0 = \cap_{t \ge 1} S_0^t$.

In fact, we can further improve upon Lemma 1 by considering the following *stochastic* characterization of hubs. Given $\zeta > 0$, let $\tau(t) \triangleq \max\{u \le t : \ |S_0^u| > M^{\frac{1}{\alpha}-\zeta}\}$ be the last time the size of the hub is large (w.r.t. threshold $M^{\frac{1}{\alpha}-\zeta}$) up until time $t$, under the convention that $\tau(t) = 0$ when taking maximum over the empty set. Lemma 2 bounds the time gap between the emergence of large hubs. The proof builds upon extreme value theory and a straightforward coupling between $\sup_{t \le T} t - \tau(t)$ and geometric random variables.

**Lemma 2** Let Assumptions 2 and 3 hold. Define the event $A_{\alpha,\zeta} \triangleq \{h_{\hat{m}} \ge M^{\frac{1}{\alpha}-\frac{\zeta}{2}}\}$, where $\alpha > 1$ is the heavy-tailed index stated in Assumption 2 for the degree distribution. For each $\zeta \in (0, 1-\frac{1}{\alpha})$, there exists $\gamma > 0$ such that $\mathbf{P}\big((A_{\alpha,\zeta})^{\complement}\big) = o\big(\exp(-M^{\gamma})\big)$. Furthermore, there exists $M_0 > 0$ such that $\mathbf{P}(\sup_{t \le T} t - \tau(t) > \log T \mid A_{\alpha,\zeta}) \le 1/MT, \qquad \forall M \ge M_0, \ T \ge 1.$

**Remark 1** (Comparison to dense and light-tailed graphs) Xu and Klabjan (2024) considers dense E-R graphs, where any pair of clients communicates on a regular basis (i.e., the expected degree of each graph $G_t$ is $O(M^2)$). In this work, we show that under the presence of heavy tails in degree distributions, clients can afford to collaborate over a much sparse $O(M)$-budget communication by sending messages to and receiving messages from the hub center $\hat{m}$ that integrates all information.

### 3.2 Information Delay over Sparse Graphs

The sparsity of graphs $G_t$ makes existing analysis in Xu and Klabjan (2024) obsolete and calls for a new approach to obtain detailed bounds regarding the information delay under sparse communication. Specifically, given some non-empty subset of clients $S \subseteq [M]$, let $\bar{S}^0 = S$, and $\bar{S}^t \triangleq \{m \in [M] : \ m \in \bar{S}^{t-1};$ or $\exists n \in \bar{S}^{t-1}$ s.t. $(m,n) \in E_t\}$ for each $t \ge 1$. That is, if all clients in $S$ send a piece of message at time 1, which will be passed to neighbors over graph $G_t$ at each time $t$, then $\bar{S}^t$ is the collection of clients that have received the message at time $t$. Lemma 3 shows that, with high probability, the information delay uniformly for any client $m \in [M]$ is at most $O\big((\log M)^2\big)$. Our proof strategy is to establish a coupling between the sizes of the graphs $(G_t)_{t \ge 1}$ and a branching process, whose size grows geometrically fast in expectation.

**Lemma 3** Under Assumption 3, there exists $\kappa \in (0,\infty)$ such that $\mathbf{P}\big(m \notin \bar{S}^{\gamma \cdot \kappa \cdot (\log M)^2} \text{ for some } m \in [M]\big) \le M^{-\gamma}$ holds for any $\gamma > 0$, $M \ge 1$, and any non-empty subset $S \subseteq [M]$.

## 4 HOMOGENEOUS REWARDS

This section considers the homogeneous-reward setting. The algorithmic framework will be extended to the heterogeneous-reward setting in the next section. Specifically, we propose the algorithm in Section 4.1, addressing the challenges from both heavy-tailed rewards and sparse graphs. Then, we establish the theoretical effectiveness of the proposed algorithm in Section 4.2.

### 4.1 Algorithm

Under homogeneous rewards, we propose a new algorithm called HT-HMUCB (**H**eavy-**T**ailed **H**o**M**ogeneous **U**pper **C**onfidence **B**ounds). See Algorithm 1 for the pseudo-code. The algorithm consists of several stages that proceed in the following order.

*Hub identification.* A novel step in our algorithm concerns the identification of hub center $\hat{m} = \arg\max_{m \in [M]} |\mathcal{N}_m(1)|$, i.e., the client with the highest degree at time $t = 1$. Due to space limitations, we

defer the details of this step to Algorithm 4 in the online supplementary material (Wang and Xu 2025), but stress that it is quite intuitive: each client transmits its degree information (at time $t = 1$) across the entire network, and this information is allowed to propagate sequentially over the graphs $(G_t)_{t\geq 1}$ for a sufficiently long period (more precisely, $O\big((\log M)^2\big)$ steps), so that we are high confidence that every client accurately identifies $\hat{m}$—the client with the highest degree at $t = 1$. Afterwards, all non-center clients will take actions based on the reward information processed by and sent from the hub center $\hat{m}$.

*Arm selection.* During this stage, the clients decide which arm to pull by executing a UCB-based strategy, where each arm $i$ is assigned a UCB index, formally expressed as $\hat{\mu}_i^m(t) + \rho^{\frac{1}{1+\varepsilon}}\big(\frac{c\log(t)}{N_{m,i}(t)}\big)^{\frac{\varepsilon}{1+\varepsilon}}$. Here, $\hat{\mu}_i^m(t)$ and $N_{m,i}(t)$ represent the global reward estimators and sample counts of arm $i$ by client $m$, respectively, defined in Rule 1 below. Constants $\rho$ and $\varepsilon$ are characterized in Assumption 1, and $c$ is specified in Theorem 1.

*Transmission.* We define an information filtration $\mathscr{F}_m(t)$ as the information available to $m$ up to time $t$, which reads as $\mathscr{F}_m(0) = \{(m,1), r^m_{a^1_m}(1), N_{m,i}(1), \hat{\mu}_i^m(1)\}, \mathscr{F}_m(t) = \cup_{l\in\mathscr{N}_m(t-1)}\mathscr{F}_l(t-1)$. Each client $m$ communicates with its neighbors $\mathscr{N}_m(t)$ by sending a message, composed of $(m,t), r^m_{a^t_m}(t)$, $N_{m,i}(t)$, $\hat{\mu}_i^m(t)$, and $\mathscr{F}_m(t)$, while collecting messages from its neighbors.

*Information update.* After pulling arms and receiving feedback from the environment, as well as receiving information from neighbors, the clients proceed to update their information based on **Rule 1** detailed below.

**Rule 1** (Information update rule for HT-HMUCB)

1) Local estimation $t_{m\leftarrow l} = \max\{s \leq t : \ (l,s) \in \mathscr{F}_m(t)\}$, Local sample counts: $n_{m,i}(t+1) = n_{m,i}(t) + \mathbb{1}_{a^t_m = i}$ 2) Global estimation

Sample counts: $N_{m,i}(t+1) = \sum_{l\in\mathscr{N}_m(t)} n_{l,i}(t+1)$

if $m$ is center (i.e., $m = \hat{m}$), set $\hat{\mu}_i^{\hat{m}}(t+1) = MoM_B(\{r_i^l(s) : r_i^l(s) \in \mathscr{F}_{\hat{m}}(t)\})$, o.w., $\hat{\mu}_i^m(t+1) = \hat{\mu}_i^{\hat{m}}(t_{m\leftarrow\hat{m}})$

Here, $MoM_B\big((X_i)_{i\in[n]}\big)$ denote the **median of mean** estimator with $B$ batches, i.e., the median of $\hat{\mu}_1, \ldots, \hat{\mu}_B$ defined by $\hat{\mu}_j = \frac{1}{N}\sum_{t=(j-1)N+1}^{jN} X_t$ with $N = \lfloor n/B \rfloor$. MoM estimators have been applied in Bubeck et al. (2013) for UBC algorithms in single-agent settings. Our work further demonstrates its use for robust estimation under heavy tails in multi-agent MAB problems.

**Remark 2** (Comparison to prior works) Our algorithm differs from the existing ones in Dubey and Pentland (2020) for homogeneous settings with heavy-tailed rewards in the following key aspects: (i) the clients identify the hub center to maximize information efficiency, which is computationally more tractable than the clique search in Dubey and Pentland (2020); and (ii) the clients take actions based solely on the information from hub center, rather than maintaining global estimators individually.

### 4.2 Regret Analyses

In this section, we demonstrate the effectiveness of the proposed algorithm through regret analyses. Notably, the tail index $\alpha$ for degree distributions in Assumption 2 plays a key role in our regret bound. First, we establish Theorem 1 for $\alpha \in (1,2)$ by utilizing the deterministic characterization of hubs in Lemma 1.

**Theorem 1** ($\alpha \in (1,2)$) Let Assumptions 1–3 hold with $1 < \alpha < 2$. Let Algorithm 1 run under Rule 1. Given $\zeta \in (0, 2-\alpha)$, there exists $\eta > 0$ such that, for any $T$ and $M$, the event $A_{\zeta,\delta}$ holds with probability at least $(1 - 2\eta/M - \eta/TM)$, and we have $\mathbf{E}[R_T|A_{\zeta,\delta}] \leq L + M\sum_{i\in[K]}\big(2c\Delta_i \log T/M^{2-\alpha-\zeta}\cdot(\Delta_i/2C\rho^{\frac{1}{1+\varepsilon}})^{\frac{1+\varepsilon}{\varepsilon}} + \frac{\pi^2}{3}\Delta_i\big) = O((1 + 2M^{\alpha-1+\zeta})\cdot\rho^{\frac{1}{\varepsilon}}\sum_{i\in[K]}\Delta_i^{-\varepsilon}\cdot\log T)$, where $L = 2\kappa K(\log M)^2\log T$, $\kappa$ is the constant characterized in Lemma 3, $c = (16\log 2e^{1/8})^{\frac{\varepsilon}{1+\varepsilon}}$, $C = (12)^{\frac{1}{1+\varepsilon}}$, $B = 8\log(e^{1/8}T)$ is the count of batches for MoM estimators, $|S_0|$ is the size of $S_0$ (see Lemma 1), and the event is defined by $A_{\zeta,\delta} = A^1_{\zeta,\delta}\cap A^2_{\zeta,\delta}\cap A^3_{\zeta,\delta}$ with $A^1_{\zeta,\delta} = \{|S_0| \geq M^{2-\alpha-\zeta}\}$, $A^2_{\zeta,\delta} = \{n_{m,i}(t_{m\leftarrow l}) \geq n_{m,i}(t) - \kappa\log M\log T \ \ \forall t \leq T, \ \forall i,m,l\}$, and $A^3_{\zeta,\delta} = \{\hat{m}(m) \neq \hat{m} \text{ for some } m \in [M]\}$ (see Algo. 4 in Online Material). In particular, $\mathbf{E}[R_T|A_{\zeta,\delta}] \leq O(M^{\alpha-1+\zeta}\cdot\log T) = o(M)\cdot O(\log T)$.

**Algorithm 1** HT-HMUCB (Heavy-tailed Homogeneous UCB)

**Initialization**: For each client $m$ and arm $i \in \{1,2,\ldots,K\}$, we set $N_{m,i}(L+1) = n_{m,i}(L)$; all other values at $L+1$ are initialized as 0; let $L$ be specified as in Theorem 1

**for** $t = 1,2,\ldots,L$ **do**

  Each client transmits the degree information at time $t=1$ (i.e., $|\mathscr{N}_m(1)|$) to their neighbors, so that after $L$ steps, it holds with high probability that each client accurately identifies $\hat{m} = \arg\max_{m\in[M]} |\mathscr{N}_m(1)|$;

**end**

**for** $t = L+1, L+2, \ldots, T$ **do**

  **for** *each client* $m$ **do** // UCB

1  $a_m^t = \arg\max_{i\in[K]} \hat{\mu}_i^m(t) + \rho \frac{1}{1+\varepsilon}\left(\frac{c\log(t)}{N_{m,i}(t)}\right)^{\frac{\varepsilon}{1+\varepsilon}}$ and pull arm $a_m^t$ and receive reward $r_{a_m^t}^m(t)$

2  **end**

3  The environment generates the graph $G_t$; // Env

4  Each client $m$ sends $(m,t)$, $r_{a_m^t}^m(t)$, $N_{m,i}(t)$, $\hat{\mu}_i^m(t)$, $\mathscr{F}_m(t)$ to each client in $\mathscr{N}_m(t)$ // Transmission

5  **for** *each client* $m$ **do**

6  **for** $i = 1,\ldots,K$ **do**

7  Update $n_{m,i}(t), N_{m,i}(t)$ and $\hat{\mu}_i^m(t)$ by Rule 1 // Update

8  **end**

9  **end**

**end**

*Proof Sketch.* The proof hinges on the key observation that, with high probability, the hub size $|S_0|$ is no less than $M^{2-\alpha-\zeta}$; see Lemma 1. The communication delay between the clients is then bounded by Lemma 3, hence the clients in the hub enjoy the reduction in sample complexity. This is achieved by utilizing a concentration inequality with respect to $\sum_{m\in S_0} n_{m,i}(t)$ instead of $n_{m,i}(t)$, resulting in an individual regret of order $1/|S_0| \cdot \log T$. Consequently, the total regret is of order $M-|S_0|/|S_0| \cdot \log T$. □

We stress that, in Theorem 1, the dependency on $\alpha$ characterizes the relationship between the regret and the heavy-tailed graph dynamics, and also reflects the reduction of complexity that is unique to our heavy-tailed-graph setting. Additionally, the regret bound depends on $\rho$ and $\varepsilon$ (see Assumption 1), capturing the influence of the heavy-tailed rewards. The dependency on optimality gaps $\Delta_i$'s is standard in MAB, but in our case there is an additional $\varepsilon$-polynomial factor due to heavy-tailed rewards.

In fact, we can obtain even stronger regret bound—and even without the requirement of $\alpha \in (1,2)$ in Theorem 1—by considering the the stochastic characterization of hubs $S_0^t$. This is demonstrated in Theorem 2.

**Theorem 2** ($\alpha > 1$) Let Assumptions 1–3 hold. Let Algorithm 1 run under Rule 1. Given $\zeta \in (0, 2-\alpha)$ there exists $\eta > 0$ such that, for any $T$, $M$, the event $A_{\alpha,\delta,\zeta}$ holds with probability at least $(1-\eta/M - \eta/TM)$, and $\mathbf{E}[R_T | A_{\alpha,\delta,\zeta}] \le L + M\sum_{i\in[K]}\left(2c\log T/M^{\frac{1}{\alpha}-\zeta}(\Delta_i/2C\rho)^{\frac{1}{1+\varepsilon}}\right)^{\frac{1+\varepsilon}{\varepsilon}} + \pi^2/3)\Delta_i = O(M^{1-\frac{1}{\alpha}+\zeta}\cdot\log T)$, where $A_{\alpha,\delta,\zeta} = A_{\zeta,\delta} \cap A_{\alpha,\zeta}$, with the event $A_{\alpha,\zeta}$ defined in Lemma 2, the event $A_{\zeta,\delta}$ defined in Theorem 1, and parameters $L, \kappa, c, C, B$ specified as in Theorem 1.

*Proof Sketch.* The information delay characterized in Lemma 3 for sparse graphs still holds here. Meanwhile, Lemma 2 proves that hub size $|S_0^t|$, despite being time-varying, is often times at least of order $M^{\frac{1}{\alpha}-\frac{\zeta}{2}}$, which is even larger than the $O(M^{2-\alpha-\zeta})$ lower bound for the hub size developed in Lemma 1. This tighter lower bound for the hub size leads to further variance reduction as clients can efficiently collect information sent by an (often times) even larger hub $S_0^t$. □

**Remark 3** (Comparison to Theorem 1) Given $\alpha > 1$, the index $1 - \frac{1}{\alpha}$ for regret bound in Theorem 2 is always smaller than the index $2 - \alpha$ in Theorem 1, due to the preliminary inequality $-\frac{1}{x} \leq 1 - x$ for any $x > 1$. This implies that, by considering a dynamic and time-dependent hub, the power-law index in our regret bound is further improved upon the results in Theorem 1. This highlights the advantage of leveraging the tighter characterization for notion of time-varying hub as in Lemma 2, thus enabling a regret bound of even smaller order under more relaxed assumptions (i.e., without the requirement of $\alpha < 2$).

**Remark 4** (Comparison with existing literature on homogeneous MA-MAB) We begin by focusing on the perspective of regret bounds and emphasizing the order w.r.t. $M$, as naive UCB already leads to a regret of order $\log T$. First, the regret bound in Dubey and Pentland (2020) is $O(\alpha(G)\rho^{\frac{1}{\varepsilon}}(\sum_i \Delta_i^{-\varepsilon})\log T)$ and their algorithm relies on solving an NP-hard problem to find the clique (i.e., the largest independent set). Here, the quantity $\alpha(G)$—the independence number of graph $G$—may not admit an explicit form and is still an active research topic. For connected graphs, some known bounds on $\alpha(G)$ are (Willis 2011) $\alpha(G) \leq M - \frac{M-1}{\Delta}$, where $\Delta$ is the maximum degree and $\alpha(G) \geq \frac{M}{1+\Delta}$, thus implying $R_T \leq O(M \cdot \log T)$. In contrast, we obtain an improved regret bound that is sub-linear in $M$. Specifically, since the slackness parameter $\zeta$ in Theorem 2 can be set arbitrarily close to 0, our regret bound is almost of order $O(M^{1-\frac{1}{\alpha}+\zeta})$. In particular, our bound is established for sparse graphs with expected degree (i.e., count of edges over the graph) of order $O(M)$. While Yang et al. (2023) establishes a regret bound of order $O(\log T)$ independent of $M$ , the authors assume that the graph is connected or complete and only consider sub-Gaussian rewards.

Regarding assumptions, we emphasize: (i) we do not require the graph to be connected (or $l$-periodically connected, as in Zhu and Liu (2023)); (ii) we allow the graph to change over time; and (iii) we address sparse graphs. Points (i) and (iii) address significant gaps in existing works on multi-agent multi-armed bandit problems, to the best of our knowledge, and point (ii) resolves an open problem identified in Dubey and Pentland (2020), which suggested time-varying network analysis and tested this case numerically. Our heavy-tailed reward assumption is in the same spirit as in Dubey and Pentland (2020). To the best of our knowledge, our framework is the most general to date for homogeneous MA-MAB.

## 5 HETEROGENEOUS REWARDS

This section addresses the more general heterogeneous setting. Specifically, we present the algorithm in Section 5.1 and the regret analysis in Section 5.2. The results are well beyond the scope of the existing work on MA-MAB with random graphs and heterogeneous rewards (Xu and Klabjan 2024), the scope of which is limited to Erdős–Rényi graphs with light-tailed, dense dynamics.

### 5.1 Algorithm

Under heterogeneous rewards, we propose a new algorithm, namely HT-HTUCB (**H**eavy-**T**ailed **H**e**T**erogeneous UCB), as the presence of heterogeneity in rewards necessitates different approaches to the communication and updates of information across clients. The pseudo-code is provided in Algorithm 2, with the details of the burn-in period (i.e., the first $L$ steps) collected in Algo. 3 of Online Material (Wang and Xu 2025) (which is identical to that of Xu and Klabjan (2024)). The burn-in period prepares the clients with initial rewards and graph estimators, enabling them to communicate and integrate information in the subsequent learning stage. Specifically, during the burn-in period, clients pull each arm $1 \leq i \leq K$ sequentially and update the average reward of each arm as local reward estimators. Simultaneously, clients observe the graph and update the edge frequency and average degree of clients. At the end of the burn-in period, the clients output an initial global estimator, calculated as the weighted average of the local reward estimators using the edge frequencies as weights.

Moving onto the learning period, clients employ UCB-based strategies to pull arms, communicate with each other, and integrate the information collected from neighbors. The steps are executed in the following order.

*Arm Selection.* We still employ a UCB-based strategy, but the global estimator is constructed differently, with an additional condition during the execution of the UCB-based strategy: $n_{m,i}(t) \leq N_{m,i}(t) - 2\kappa(\log M)^2 \log T$. This novel condition ensures that clients remain synchronized and accounts for the longer information delay caused by heavy-tailed graph dynamics, which differs from (Xu and Klabjan 2024; Zhu and Liu 2023).

*Transmission.* This step is almost identical to that of Section 4.1, except for the message components. Here, an information filtration $\mathscr{F}_m(t)$ reads as $\mathscr{F}_m(0) = \{(m,1), r^m_{a^t_m}(1), N_{m,i}(1), \bar{\mu}^m_i(1), \hat{\mu}^m_i(1)\}$ and $\mathscr{F}_m(t) = \cup_{l \in \mathscr{N}_m(t-1)} \mathscr{F}_l(t-1)$. Each client $m$ communicates with its neighbors $\mathscr{N}_m(t)$ by sending an message, including $(m,t), r^m_{a^t_m}(t)$, $N_{m,i}(t)$, $\bar{\mu}^m_i(t)$, $\hat{\mu}^m_i(t)$ and $\mathscr{F}_m(t)$, while collecting messages from its neighbors.

*Information update.* Since the heterogeneity in rewards necessitates obtaining reward information from all clients, we propose a new information update step to aggregate information, represented by **Rule 2**.

**Rule 2** (Information update rule for HT-HTUCB)

1) Local estimation

Local sample counts: $n_{m,i}(t+1) = n_{m,i}(t) + \mathbb{1}_{a^t_m = i}$, Local estimator: $\bar{\mu}^m_i(t+1) = MoM_B(\{r^m_i(s)\}_{1 \leq s \leq t})$

2) Global estimation with $N = (12^{\frac{1}{1+\varepsilon}})^{\frac{1+\varepsilon}{\varepsilon}} + 1$

$$t_{m \leftarrow l} = \max\{s \leq t : \ (l,s) \in \mathscr{F}_m(t)\}, \quad N_{m,i}(t+1) = \max\{n_{m,i}(t+1), \{N_{l,i}(t)\}_{l \in \mathscr{N}_m(t)}\};$$

$$\hat{\mu}^m_i(t+1) = \textstyle\sum_{l \in [M]} P'_t \hat{\mu}^l_i(t_{m \leftarrow l}) + d_{m,t} \sum_{l \in [M]} \bar{\mu}^l_i(t_{m \leftarrow l}), \quad P'_t = {}^{(N - M2^{1/(\varepsilon+1)})}/_{MN2^{1/\varepsilon+1}}, \quad d_{m,t} = {}^{(1-MP'_t)}/_{M}.$$

**Remark 5** (Comparison to existing work and Section 4.1) Compared to Section 4.1, the arm selection step imposes an extra requirement that $n_{m,i}(t) \leq N_{m,i}(t) - 2\kappa(\log M)^2 \log T$ in UCB, which balances exploitation and exploration given the noise in the global reward estimator $\hat{\mu}$ in the heterogeneous case. Secondly, we remove the hub estimation step in the heterogeneous setting, and each client now must collect information from all clients by message passing with through neighbors sets $\mathscr{N}_m(t)$. Lastly, we run Rule 2 instead of Rule 1 for information update.

Notably, our UCB index $\hat{\mu}^m_i(t) + \rho^{\frac{1}{1+\varepsilon}} \left(c \log(t) / N_{m,i}(t)\right)^{\varepsilon/1+\varepsilon}$ is able to address heavy-tailed rewards, while Xu and Klabjan (2024) considers $\hat{\mu}^m_i(t) + F(m,i,t)$ where $F(m,i,t) = \sqrt{C_1 \ln t / n_{m,i}(t)}$ (for sub-Gaussian rewards) and $F(m,i,t) = \sqrt{C_1 \ln T / n_{m,i}(t)} + C_2 \ln T / n_{m,i}(t)$ (for sub-exponential rewards). Also, we propose new the information update rule due to the differences in reward and graph dynamics.

### 5.2 Regret Analysis

Importantly, we next demonstrate the effectiveness of Algorithm 2 by investigating the regret upper bound. In particular, we stress that Theorem 3 does not rely on Assumption 2, meaning that the results address *both light-tailed and heavy-tailed degree distributions* in the random graph model (1).

**Theorem 3** Let Assumptions 1 and 3 hold. Let Algorithm 2 run under Rule 2. Then, for any $T$ and $M$, we have $\mathbf{P}(A_{\zeta,\delta}) \geq 1 - 7/T$ and $\mathbf{E}[R_T | A_{\zeta,\delta}] \leq 2\kappa K(\log M)^2 \log T + \sum_{i \in [K]} M\Delta_i \cdot (\max\{\frac{2cN \log T}{(\Delta_i / 2\rho^{\frac{1}{1+\varepsilon}})^{\frac{1+\varepsilon}{\varepsilon}}}, 2\kappa \log M \log T\}) + \sum_{i \in [K]} M\Delta_i \cdot (\frac{2\pi^2}{3} + 2\kappa(\log M)^2 \log T) = O(M \log T)$, where $K$ is the count of arms, the event is defined by $A_{\zeta,\delta} = \{n_{m,i}(t_{m,j}) \geq n_{m,i}(t) - \kappa(\log M)^2 \log T \ \ \forall t \leq T, \ \forall m,i,j\}$, $N = (12^{\frac{1}{1+\varepsilon}})^{\frac{1+\varepsilon}{\varepsilon}} + 1$, and parameters $c, B, \kappa, \rho, \varepsilon$ are specified as in Theorem 1.

*Proof Sketch.* Note that we do not exploit the hub structure in this setting, as clients need to collect information from all other clients rather than relying solely on a hub that contains only a subset of information. Nevertheless, the information delay bounds in Lemma 3 for sparse graphs still hold. Using the estimators constructed in Rule 2, which leverage neighbor information to collect and integrate global information, we prove a concentration inequality for the global estimator $\hat{\mu}$ with respect to the global mean values: $|\hat{\mu}^m_i(t) - \mu_i| \leq 2\rho^{1/1+\varepsilon} (\frac{Mc \log(T)}{\min_m n_{m,i}(t)})^{\frac{\varepsilon}{1+\varepsilon}}$. This ensures that clients can identify the globally optimal arm using UCB after ${}^{2cM \log T} / {(\Delta_i / 2\rho^{1/1+\varepsilon})^{1+\varepsilon/\varepsilon}}$ steps with high probability. Another possible scenario is that, when clients are not synchronized, they randomly select arms instead of using UCB (see Step 14 of Algo. 2), the regret

**Algorithm 2** HT-HTUCB (Heavy-Tailed Heterogeneous UCB): Learning period

**Initialization**: For each client $m$ and arm $i \in \{1,2,\ldots,K\}$, we have $\hat{\mu}_i^m(L+1)$, $N_{m,i}(L+1) = n_{m,i}(L)$; all other values at $L+1$ are initialized as 0 let $L$ be specified as in Theorem 1

**for** $t = L+1, L+2, \ldots, T$ **do**

  **for** *each client m* **do** // UCB

    **if** *there is no arm i such that* $n_{m,i}(t) \leq N_{m,i}(t) - 2\kappa(\log M)^2 \log T$ **then**

      $a_m^t = \arg\max_i \hat{\mu}_i^m(t) + \rho^{\frac{1}{1+\varepsilon}} \left(c\log(t)/N_{m,i}(t)\right)^{\frac{\varepsilon}{1+\varepsilon}}$

    **else**

      Randomly sample an arm from set $\{i : n_{m,i}(t) \leq N_{m,i}(t) - 2\kappa(\log M)^2 \log T\}$ and set it as $a_m^t$

    **end**

    Pull arm $a_m^t$ and receive reward $r_{a_m^t}^m(t)$

  **end**

  The environment generates the graph $G_t = (V, E_t)$; // Env

  Each client $m$ sends $(m,t)$, $r_{a_m^t}^m(t)$, $N_{m,i}(t)$, $\bar{\mu}_i^m(t)$, $\hat{\mu}_i^m(t)$, $\mathscr{F}_m(t)$ to $\mathscr{N}_m(t)$ // Transmission

  **for** *each client m* **do**

    **for** $i = 1, \ldots, K$ **do**

      Update $n_{m,i}(t), N_{m,i}(t)$ and $\bar{\mu}_i^m(t), \hat{\mu}_i^m(t)$ by Rule 2

    **end**

  **end**

**end**

of which is proved to be upper bounded by the last term of the regret bound in Theorem 3. As a result, the total number of pulls of sub-optimal arms $n_{m,i}(t)$ can be bounded by $O(\log T)$. Lastly, regret decomposition shows that the regret is dominated by $n_{m,i}(t)$, and thus has the upper bound. □

**Remark 6** (Extension) We emphasize that Theorem 3 holds for both light-tailed and heavy-tailed random graphs as it does not rely on Assumption 2. This also highlights a key difference between the homogeneous and heterogeneous settings: in the homogeneous case, heavy-tailed $h_m$'s lead to sample complexity reduction compared to the light-tailed case (with regret of order $O(M\log T)$), whereas the heterogeneous result holds true for both light-tailed and heavy-tailed distributions for the attraction weight $h_m$'s.

**Remark 7** (Comparison to existing work) We analyze the order of the regret with respect to $T$, as the absence of communication can lead to regret of order $O(T)$ (Xu and Klabjan 2025a). The most relevant work (Xu and Klabjan 2024) establishes a regret upper bound of $O(\log T)$ specifically for Erdős–Rényi graph with light-tailed and dense dynamics. In particular, the dense connectivity in (Xu and Klabjan 2024) requires any pair of clients connects with a rather high probability (of order $O(1)$) at any time step, which limits practical applicability. Besides, their work covers some reward distributions that are heavier than the sub-Gaussian class but still have finite moment-generating functions (MGFs). In contrast, we impose no such assumptions on graph connectivity, consider sparse graphs with only $O(M)$ total degree (instead of the $O(M^2)$ total degree in (Xu and Klabjan 2024)), and allow for a significantly more general class of reward distributions, potentially with infinite variance. On the other hand, Zhu and Liu (2023) achieves a regret bound of the same order, but assumes a connected or periodically connected graph under sub-Gaussian rewards, which may not hold for sparse graphs. Last but not least, our results close an open problem in Dubey and Pentland (2020) regarding homogeneous rewards under heavy-tailed settings (Section 4.2), thus bridging the gap in existing literature and addressing challenges posed by heavy-tailed graphs.

## 6 DISCUSSION ON OPTIMALITY

Notably, our regret upper bounds in these new settings are optimal in terms of their order in $T$. In (Xu and Klabjan 2025b), the authors show that in stochastic settings with time-invariant rewards, the regret lower bound for any multi-agent multi-armed bandit with random graphs is of order $\log T$. This matches the upper bounds established here, confirming the $T$-optimality of our results. However, the lower bounds are not known to be optimal in $M$, as no such result exists in the literature or in our paper—this remains an open direction. We also note that the dependence on $M$ in our bounds is smaller than in existing works that rely on stronger assumptions (see above remark), indicating that our results improve upon prior work and offer a practical advantage.

## REFERENCES


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

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. 2002. "The nonstochastic multiarmed bandit problem.". *SIAM Journal on Computing* 32(1):48–77.

Besson, L., and E. Kaufmann. 2018, 07–09 Apr. "Multi-Player Bandits Revisited". In *Proceedings of Algorithmic Learning Theory*, edited by F. Janoos, M. Mohri, and K. Sridharan, Volume 83 of *Proceedings of Machine Learning Research*, 56–92: PMLR.

Boguná, M., and R. Pastor-Satorras. 2003. "Class of correlated random networks with hidden variables". *Physical Review E* 68(3):036112.

Bubeck, S., N. Cesa-Bianchi, and G. Lugosi. 2013. "Bandits with heavy tail". *IEEE Transactions on Information Theory* 59(11):7711–7717.

Chung, F., and L. Lu. 2002. "The average distances in random graphs with given expected degrees". *Proceedings of the National Academy of Sciences* 99(25):15879–15882 https://doi.org/10.1073/pnas.252631999.

Clancy, David Jr 2021. "Epidemics on critical random graphs with heavy-tailed degree distribution".

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data". *SIAM Review* 51(4):661–703 https://doi.org/10.1137/070710111.

da Cruz, J. P., and P. G. Lind. 2013. "The bounds of heavy-tailed return distributions in evolving complex networks". *Physics Letters A* 377(3):189–194 https://doi.org/https://doi.org/10.1016/j.physleta.2012.11.047.

Dubey, A., and A. Pentland. 2019. "Thompson Sampling on Symmetric $\alpha$-Stable Bandits". *arXiv preprint arXiv:1907.03821*.

Dubey, A., and A. Pentland. 2020. "Cooperative multi-agent bandits with heavy tails". In *International Conference on Machine Learning*.

Hearnshaw, E. J., and M. M. Wilson. 2013. "A complex network approach to supply chain network theory". *International Journal of Operations & Production Management* 33(4):442–469.

Jia, H., C. Shi, and S. Shen. 2021. "Multi-armed bandit with sub-exponential rewards". *Operations Research Letters* 49(5):728–733 https://doi.org/https://doi.org/10.1016/j.orl.2021.08.004.

Korda, N., E. Kaufmann, and R. Munos. 2013. "Thompson Sampling for 1-Dimensional Exponential Family Bandits". In *Advances in Neural Information Processing Systems*, edited by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Volume 26: Curran Associates, Inc.

Kunegis, J. 2013. "KONECT: the Koblenz network collection". In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, 1343–1350. New York, NY, USA: Association for Computing Machinery https://doi.org/10.1145/2487788.2488173.

Pastor-Satorras, R., and A. Vespignani. 2002, Mar. "Epidemic dynamics in finite size scale-free networks". *Phys. Rev. E* 65:035108 https://doi.org/10.1103/PhysRevE.65.035108.

Resnick, S. I. 2007. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Roman, R., J. Zhou, and J. Lopez. 2013. "On the features and challenges of security and privacy in distributed internet of things". *Computer networks* 57(10):2266–2279.

Tao, Y., Y. Wu, P. Zhao, and D. Wang. 2022, 28–30 Mar. "Optimal Rates of (Locally) Differentially Private Heavy-tailed Multi-Armed Bandits". In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, edited by G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Volume 151 of *Proceedings of Machine Learning Research*, 1546–1574: PMLR.

Vakili, S., K. Liu, and Q. Zhao. 2013. "Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems". *IEEE Journal of Selected Topics in Signal Processing* 7(5):759–767 https://doi.org/10.1109/JSTSP.2013.2263494.

van der Hofstad, R., A. J. E. M. Janssen, J. S. H. van Leeuwaarden, and C. Stegehuis. 2017, Feb. "Local clustering in scale-free networks with hidden variables". *Phys. Rev. E* 95:022307 https://doi.org/10.1103/PhysRevE.95.022307.

van der Hofstad, R., P. van der Hoorn, N. Litvak, and C. Stegehuis. 2020. "Limit theorems for assortativity and clustering in null models for scale-free networks". *Advances in Applied Probability* 52(4):1035–1084 https://doi.org/10.1017/apr.2020.42.

Vázquez, A., R. Pastor-Satorras, and A. Vespignani. 2002, Jun. "Large-scale topological and dynamical properties of the Internet". *Phys. Rev. E* 65:066130 https://doi.org/10.1103/PhysRevE.65.066130.

Xingyu Wang and Mengfan Xu 2025. "Multi-agent Multi-armed Bandit with Fully Heavy-tailed Dynamics (Online Supplementary Material)". https://arxiv.org/pdf/2501.19239.

Willis, W. 2011. "Bounds for the independence number of a graph".

Xu, M., and D. Klabjan. 2024. "Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards". *Advances in Neural Information Processing Systems* 36.

Xu, M., and D. Klabjan. 2025a. "Multi-agent Multi-armed Bandit Regret Complexity and Optimality". *International Conference on Artificial Intelligence and Statistics*.

Xu, M., and D. Klabjan. 2025b. "Multi-agent Multi-armed Bandit Regret Complexity and Optimality". In *The 28th International Conference on Artificial Intelligence and Statistics*.

Yang, L., X. Wang, M. Hajiesmaili, L. Zhang, J. C. Lui, and D. Towsley. 2023. "Cooperative Multi-agent Bandits: Distributed Algorithms with Optimal Individual Regret and Communication Costs". In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

Zhu, J., and J. Liu. 2023. "Distributed Multi-Armed Bandits". *IEEE Transactions on Automatic Control*.

## AUTHOR BIOGRAPHIES

**XINGYU WANG** is a postdoc researcher in the Department of Quantitative Economics at the University of Amsterdam. He received his Ph.D. in Industrial Engineering and Management Sciences from Northwestern University. His research interests include applied probability, stochastic simulation, and machine learning. His email address is x.wang4@uva.nl and his website is https://joshwang0322.github.io.

**MENGFAN XU** is an assistant professor in the Department of Mechanical and Industrial Engineering and an adjunct assistant professor in the Manning College of Information & Computer Sciences at University of Massachusetts Amherst. She received her Ph.D. in Industrial Engineering and Management Sciences from Northwestern University. Her research interests include online learning especially multi-armed bandit, stochastic modeling and simulation, statistical learning, and multi-agent systems. Her e-mail address is mengfanxu@umass.edu and her website is https://mengfanxu1997.github.io.