Open Multi-agent Multi-armed Bandit with Applications in Permissionless Blockchain

Mengfan Xu
Mechanical and Industrial Engineering
University of Massachusetts Amherst
mengfanxu@umass.edu

Diego Klabjan
Industrial Engineering and Management Sciences
Northwestern University
d-klabjan@northwestern.edu

Abstract

We study a multi-agent multi-armed bandit problem (MA-MAB) in open systems, where multiple agents can enter and leave at any time and face multiple bandit problems to minimize the group-wise cumulative regret. To our knowledge, this is the first work to consider a dynamic set of agents that arrive and depart according to stochastic processes, systematically evolving over time. We also extend to a permissionless blockchain-based MA-MAB (PB-MA-MAB) problem, where agents may behave either honestly or maliciously depending on compliance with the mechanism, and malicious agents may disrupt honest ones. These formulations pose new challenges, as regret grows with the increasing number of agents. To this end, we design new UCB-based methodologies for both MA-MAB and PB-MA-MAB, introducing information-integration rules for existing agents and information-access mechanisms for new agents to fully leverage available information. We derive regret bounds for our algorithms and characterize the complexity of the formulation via regret lower bounds in both settings. We establish regret upper bounds of order $\max\{O(M_0), O(\log T), O(\frac{\log^2 T}{(M_0)})1_{\{\lambda>0\}}\}$ (a significant improvement over the naïve bound $(M_0+T)\log T$), where M_0 is the initial number of agents and C reflects the arrival/departure rate. We also prove lower bounds of $O(\log T)$ and $O(M_0)$ for all consistent algorithms, and tighter bounds of $O(\log T + M_0)$ or $O(\log^2 T)$ for a subset including ours. These imply that our algorithm is nearly optimal in general and optimal in certain cases.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Multi-armed bandit (MAB) (Auer et al., 2002a,b) is a classical online learning paradigm in which, over a finite time horizon, a decision-maker (or agent) selects an arm at each time step, receives a reward, and adds it to their cumulative reward. The goal is to maximize the total reward by the end of the game, or equivalently, to minimize the regret compared to the total reward from always pulling the arm with the highest expected reward. It has become increasingly popular and is extensively applied in e-commerce (Xiang et al., 2022; Jiang et al., 2021), recommender systems (Barraza-Urbina and Glowacka, 2020; Zhu and Van Roy, 2023), healthcare (Zhou et al., 2023; Mate et al., 2020), and, more recently, large language models (Bouneffouf and Féraud, 2024; Behari et al., 2024). Recent distributed applications with multiple decision-makers motivate the cooperative multi-agent multi-armed bandit (MA-MAB) framework, which drives advances in network routing (Zhu et al., 2021; Yang et al., 2024), e-commerce (Agarwal et al., 2022; Feng et al., 2018), and biology (Azim et al., 2025; Lin et al., 2024), an active area of research. It is categorized into two settings by reward structure: stochastic, where rewards follow time-invariant distributions, and adversarial, where rewards are chosen arbitrarily by an adversary. We focus on the stochastic setting due to its wide applicability, referred to as MA-MAB. Here, multiple agents interact with multiple bandit problems and aim to maximize the global cumulative reward by selecting the arm with the highest average expected reward across all agents.

To date, research on MA-MAB has primarily focused on a static/bounded set of agents, assuming a fixed or bounded pool that does not change over time, or that has an upper bound on the number of agents (Landgren et al., 2016a,b, 2021; Zhu et al., 2020; Martínez-Rubio et al., 2019; Agarwal et al., 2022; Wang et al., 2022,

2020; Li and Song, 2022; Sankararaman et al., 2019; Chawla et al., 2020; Xu and Klabjan, 2023, 2024, 2025; Rosenski et al., 2016; Trinh and Combes, 2021). However, emerging applications often violate it, particularly with the rise of digital currencies and platforms (Helliar et al., 2020). A prominent example is blockchain systems: Bitcoin, one of the largest digital currency platforms, operates as an open-ended system where agents can freely join and leave at any time, namely permissionless blockchain (Deuber et al., 2019; Esmaili and Christensen, 2025; Li et al., 2023; Thai et al., 2019). Recently, Nakamura et al. (2023) studies the adversarial case but imposes strong assumptions on how agents enter and exit the system. In practice, yet, agent dynamics can be entirely random, contradicting these assumptions. Moreover, the stochastic rewards, despite their large community, have not yet been explored. Consequently, it remains an open question how to design a mechanism that captures the stochastic nature of agents and how to address it in MA-MAB, which we address herein.

A dynamic agent mechanism provides flexibility and attracts more agents to join, but it also raises a natural question: is the system robust to the new dynamics introduced by openness? Historically, the robust multi-agent multi-armed bandit framework has been widely studied, where some agents may behave maliciously (Vial et al., 2021, 2022; Zhu et al., 2023). Yet, this line of work focuses only on (1) risks from a fixed set of agents, and (2) primarily in the form of misleading information. In contrast, (1) a dynamic set of agents poses greater challenges, since new agents can join at any time, and (2) cyberattacks introduce more disruptive risks, including threats to the stability of the entire system. These gaps remain largely unexplored, motivating the work herein.

Stochastic processes are widely used to model randomness in complex systems (Cinlar, 2013; Cox, 2017). Queuing systems, for example, analyze how customers arrive at and are served by a facility (Simaiakis and Balakrishnan, 2016; Avi-Itzhak and Heyman, 1973). When the facility is busy, customers form a queue; these join/leave events are modeled as arrival and departure processes. This framework naturally aligns with the random agent dynamics in MA-MAB. In permissionless blockchains, it is often assumed that agents are randomly sampled from a candidate pool (binomial) (Esmaili and Christensen, 2025; Thai et al., 2019). Since the binomial distribution approaches a Poisson distribution as the pool size grows (Chen, 1974), the sequential arrivals naturally form a Poisson point process, one of the most widely used stochastic models. Thus, Poisson process can effectively model randomness in distributed and blockchain systems. Yet its role has not been understood, a gap we address herein.

Notably, blockchain systems offer potential for a robust multi-agent multi-armed bandit framework to defend against cyberattacks. In turn, the MA-MAB online learning framework holds great promise for intelligent blockchain systems, where agents can learn to make decisions autonomously. Xu and Klabjan (2024) integrates blockchain with MA-MAB by characterizing blockchain-based MA-MAB with a fixed set of agents, known as a permissioned blockchain. However, the case where agents can join or leave freely, corresponding to permissionless blockchains, remains largely unexplored, motivating the study of robust MA-MAB in open systems. In current permissionless blockchains, a common assumption is that the candidate pool of agents is finite. However, this is often unrealistic, as the global user base can be infinite. Whether sampling from an infinite pool in permissionless blockchains, together with MA-MAB, ensures robustness remains open.

The key question we answer is: *Can we design open MA-MAB that captures agent stochasticity motivated by permissionless blockchains and, in turn, apply this mechanism to blockchain systems to ensure robustness?*

1.1 Main Contributions

We answer the above question affirmatively through the following contributions. We formulate the open multi-agent multi-armed bandit (MA-MAB) problem with a dynamic agent set, considering both the absence and presence of agent departures, bridging MA-MAB and stochastic processes. The agent set evolves according to Poisson processes: arrivals follow a time-homogeneous Poisson process, and departures follow another. We further integrate permissionless blockchains with MA-MAB, formulating a robust PB-MA-MAB problem, addressing the new challenges induced by openness.

Methodologically, we design new information-integration rules for existing agents and information-access mechanisms for new agents to intelligently leverage available information. In MA-MAB, initial agents collect all reward information at each time step and maintain an aggregated reward estimator via simple averaging, thereby reducing sample complexity. To compensate for missing information, new agents follow the reward estimators maintained by existing agents. Agents employ Upper Confidence Bounds (UCB) based on their estimators. In PB-MA-MAB, we develop robust approaches given the presence of malicious agents. Specifically, initial agents construct robust estimators by placing less weight on historical data and more on recent information. New agents then rely on reward estimators verified by the blockchain after consensus among agents. Agents proceed according to blockchain operations.

Analytically, we derive regret upper bounds for our algorithms and characterize the complexity of the new formulation via lower bounds in both settings. Specifically, we show that the regret upper bound is of order $\max\{O(M_0), O(\log T), O(\frac{\log^2 T}{(M_0)})1_{\{\lambda>0\}}\}$, where M_0 is the size of the initial agent set and C depends on the arrival and departure rates. Notably, without careful handling, a trivial bound could be as large as $O((M_0+T)\cdot \log T)$ in expectation. We eliminate the linear T term by reducing the sample complexity. Moreover, we prove that the regret lower bound is of order $O(\log T)$ and $O(M_0)$ for all consistent algorithms, matching the upper bound up to a $\log T$ factor. We further establish that the lower bound can grow as large as $O(\log T + M_0)$ or $O(\log^2 T)$ when considering a new family of consistent algorithms, whose regret upper bounds scale polynomially in T based on M_0 and $\log T$. These results rely on analyzing the sample complexity in our new setting, adapting the seminal techniques from Lai and Robbins (1985) with a modified sample complexity that influences the sample log-likelihood.

2 PROBLEM FORMULATION

2.1 Open Multi-agent Multi-armed Bandit

We begin by introducing the notation; a full notation table is provided in Appendix E. Consistent with the traditional multi-armed bandit (MAB) framework, we consider K arms, the set of which is denoted by $[K] = \{1, 2, \ldots, K\}$. The time horizon is denoted by T, and each time step is indexed by $1 \le t \le T$. It is worth noting that unlike traditional multi-agent MAB, we allow the set of participating agents to change over time. We study a multi-agent MAB problem with M_t agents at time t, labeled 1 through M_t, where all agents face the same set of arms and share the same unknown reward dynamics. The agents are distributed over a fully connected graph $G_t = (V_t, E_t)$, where $V_t = \{1, 2, \ldots, M_t\}$ and E_t is the edge set. Specifically, $E_t = \{(m, n) : m \in V_t, n \in V_t\}$. We use $\mathcal{N}_m(t)$ to denote the neighbor set of agent m, where agent n is a neighbor of agent m (i.e., $n \in \mathcal{N}_m(t)$) if $(m, n) \in E_t$.

Next, we consider the reward dynamics in the stochastic setting. Specifically, at each time step $1 \le t \le T$, the reward observed by agent $1 \le m \le M_t$ for arm $1 \le i \le K$ is denoted by $r^i_m(t)$, which is drawn from a time-invariant distribution with mean μ_i. In other words, all agents share the same reward distribution (homogeneous setting). Let a^t_m denote the arm selected by agent m at time t, and let $n_{m,i}(t)$ denote the number of times agent m has pulled arm i up to time t. Let $N_{m,i}(t)$ be the total number of times arm i has been pulled by agent m's neighbors (i.e., all agents) by t.

Additionally, open systems exhibit agent dynamics. We model the agent population M_t as consisting of a fixed initial set of agents M_0 (which does not change over time), together with stochastic arrivals and departures governed by *Poisson point processes*. Formally, the agent population at time t is given by: $M_t = M_0 + M^A_t - M^D_t$ where M^A_t and M^D_t denote the number of agents that arrive and depart at time t, respectively.

The number of agents at time t, denoted M^A_t, follows a Poisson distribution, forming a Poisson point process. Specifically, for any integer $N \in \mathbb{Z}^+ \cup \{0\}$, $P(M^A_t = N) = \frac{(\lambda_A)^N e^{-\lambda_A}}{N!}$, where $\lambda_A \ge 0$ is the arrival rate.

Similarly, we model agent departures using a Poisson point process with parameter λ_D. The number of agents leaving the system at time t, denoted M^D_t, follows a Poisson distribution. Specifically, for any integer $N \in \mathbb{Z}^+ \cup \{0\}$, $P(M^D_t = N) = \frac{(\lambda_D)^N e^{-\lambda_D}}{N!}$, where $\lambda_D \ge 0$ is the departure rate.

The rationale for such an agent mechanism is as follows. Notably, if the total pool of potential agents is infinite but only a finite number are active at any given time (i.e. Binomial), this captures a realistic and practical scenario, also seen in permissionless blockchains. It is well known that the Poisson distribution arises as the limit of a Binomial distribution $\text{Bin}(n, p)$ as $n \to \infty$ with $n \cdot p = \lambda_A$. This connection motivates modeling the arrival process using a Poisson distribution with rate λ_A with an infinite agent pool. A similar argument applies to departures: if each agent leaves the system with a small probability during each time interval, then in the limit the total number of departing agents follows a Poisson distribution. Broadly, Poisson point processes are widely used to model real-world queuing and service systems, and our modeling provides a systematic way to incorporate dynamic agent sets through these stochastic processes. It reduces to existing MA-MAB when $\lambda_A = \lambda_D = 0$, implying consistency.

The objective is to minimize the group-wise regret, defined as the cumulative regret summed over all active agents: $R_T = \max_i \sum_{t=1}^T \sum_{m \in M_t} r^m_i(t) - \sum_{t=1}^T r_T$ and the group-wise pseudo regret defined as $\bar{R}_T = \max_i \sum_{t=1}^T \sum_{m \in M_t} \mu^m_i - E[r_T]$, with respect to the difference between the cumulative reward of always pulling the optimal arm and the total reward actually obtained r_T, reading as $r_T = \sum_{t=1}^T \sum_{m \in M^H_t} r^m_{a^t_m}(t)$.

2.2 Permissionless Blockchain Bandit

Motivated by the agent mechanism of randomly sampling from an infinite pool in permissionless blockchain, we next study an important application of this new agent mechanism in permissionless blockchain-based multi-agent multi-armed bandits (PB-MA-MAB). To the best of our knowledge, this is the first work to study bandits on permissionless blockchains, thereby

making a novel contribution to the robust multi-agent multi-armed bandit literature under corruptions.

Blockchain formulation. We denote the sets of honest and malicious agents at time step t by M_t^H and M_t^A, respectively, which are not known a priori. We assume that M_t^H and M_t^A evolve according to Poisson processes, with arrival and departure rates $(\lambda_A^H, \lambda_D^H)$ for honest agents and $(\lambda_A^A, \lambda_D^A)$ for malicious agents.

Let $S_V(t)$ denote the set of validators at time t, selected algorithmically, and $S_C(t)$ the set of commanders, also selected algorithmically. The total number of blocks is denoted by $B = T$. The process at each iteration follows the structure in Xu and Klabjan (2024), except that the agent sets M_t^H and M_t^A evolve over time.

Following Xu and Klabjan (2023), each agent m is associated with a public and secret key pair (PK_m, SK_m) for $1 \le m \le M_t$. The list of public keys is publicly available and ordered according to the agent set. Block approval at time t is denoted by $b_t^m \in \{0, 1\}$.

The objective is $R_T = \max_i \sum_{t=1}^T \sum_{m \in M_t^H} r_i^m(t) - r_T$ and pseudo regret $\bar{R}_T = \max_i \sum_{t=1}^T \sum_{m \in M_t^H} \mu_i^m - E[r_T]$, where $r_T = \sum_{t=1}^T \sum_{m \in M_t^H} r_{a_m^t}^m(t) 1_{b_t^m = 1}$.

Remark. *This formulation poses additional challenges beyond Section 2.1 and prior blockchain bandits (Xu and Klabjan, 2024). Compared to MA-MAB, it includes the term $1_{b_t^m=1}$, reflecting that unapproved blocks incur higher regret. The dynamic agent mechanism also complicates the setting relative to Xu and Klabjan (2024), since permissionless blockchains require accurate information: new agents must fetch data, and misleading information may propagate, causing regret to grow linearly with the number of agents.*

3 METHODOLOGY

In this section, we propose new methods to minimize the regret defined above in light of the newly introduced challenges. For the two problem settings (MA-MAB and PB-MA-MAB), we develop two algorithmic frameworks: one for MA-MAB in Section 3.1, and one for PB-MA-MAB in Section 3.2. The full pseudocode is provided in Appendix C.

3.1 Open Multi-agent Multi-armed Bandit

We begin with MA-MAB in the dynamic-agent setting, illustrating both the burn-in and learning periods. The detailed steps of the algorithm are described below.

Burn-in. During the burn-in period of length L, which will be specified in the regret statement, we treat the setting as a standard MA-MAB problem with a fixed set of agents in a homogeneous environment, except that agents may arrive and depart according to the stochastic processes. Specifically, for the first $L = O(\log T)$ steps, the initial set of agents interacts with the bandit game using an algorithm designed for homogeneous multi-agent MAB problems on fully connected graphs, as in Landgren et al. (2021) (used here for illustration purposes; in practice, any homogeneous multi-agent bandit algorithm could be applied). Newly arriving agents randomly select arms during this phase.

At the end of the burn-in period, the initial set of agents, i.e. for $m \in M_0$, it updates the number of arm pulls and their local reward estimators with respect to arm i as follows: $N_{m,i}(t) = \sum_{j \in \mathcal{N}_m(t)} n_{j,i}(t), \bar{\mu}_i^m(t) = \frac{\sum_{s: a_m^s = i} r_{a_m^s}^m(s)}{n_{m,i}(t)}, \tilde{\mu}_i^m(t) = \frac{\sum_m \sum_{s: a_m^s = i} r_{a_m^s}^m(s)}{N_{m,i}(t)}$.

Here, $n_{m,i}(t)$ and $\bar{\mu}_i^m(t)$ denote the local sample counts and reward estimates of arm i at agent m, while $N_{m,i}(t)$ and $\tilde{\mu}_i^m(t)$ denote the global sample counts and reward estimates of arm i at agent m.

Using the information accumulated during the exploration phase, we proceed to the learning period, where agents employ more informed strategies. Newly arriving agents are incorporated intelligently into the system, as the regret depends heavily on their participation.

Arm selection. As in the standard MAB framework, agents decide which arm to pull at each time step. Specifically, during the burn-in period, each agent m selects arm $a_m^t = t \bmod K$. During the learning period, each agent follows a UCB-like approach by constructing an Upper Confidence Bound (UCB) index for each arm, i.e., a score assigned to every arm i, and then selecting the arm with the highest score. Precisely, $a_m^t = \arg\max_i \tilde{\mu}_i^m(t-1) + F(m, i, t-1)$ where $\tilde{\mu}_i^m(t)$ is the aforementioned global estimator at agent m, and quantity $F(m, i, t) = (\frac{C_1 \log t}{N_{m,i}(t)})^\beta$ represents the uncertainty in the global estimator $\tilde{\mu}_i^m(t)$, with constant C_1, β being specified in the theorems.

Broadcasting. During broadcasting, the agents sent information to neighbors. Each agent m broadcast $n_{m,i}(t), N_{m,i}(t), \bar{\mu}_i^m(t), \tilde{\mu}_i^m(t)$ to agent $1 \le j \le M_t$.

Aggregation. For agents 1 through M_t, information is aggregated across all agents, i.e., $N_{m,i}(t) = \sum_{j \in \mathcal{N}_m(t)} n_{j,i}(t), \hat{\mu}_i^m(t) = (\sum_{j \in \mathcal{N}_m(t)} \bar{\mu}_i^j(t-1) \cdot n_{j,i}(t) + \sum_{j \in \mathcal{N}_m(t)} r_{a_j^t}^j(t) \cdot 1_{a_j^t = i}) / N_{m,i}(t)$. For newly arriving agents, i.e., $m \in M_{t+1} \setminus M_t$, they update their information as $N_{m,i}(t) = N_{j,i}(t), \hat{\mu}_i^m(t) = \hat{\mu}_i^j(t), j \in M_t$.

Update. The agents subsequently update their local estimators as follows: $n_{m,i}(t) = n_{m,i}(t-1) + 1_{b_t=1} \cdot 1_{a_m^t = i}, \bar{\mu}_i^m(t) = (\bar{\mu}_i^m(t-1) + r_{a_m^t}^m(t) \cdot 1_{a_m^t = i}) / n_{m,i}(t)$.

3.2 Permissionless Blockchain Bandit

Next, we present the algorithm steps for permissionless blockchains, following the structure of Xu and Klabjan

(2024) but highlighting key differences in each module.

Information fetch. Agents arrive according to a Poisson point process M_t, and new agents begin by fetching the available information on the chain upon arrival. This highlights a key innovation compared to permissioned blockchains, where all agents are assumed to remain synchronized. For each new agent m (i.e., $m \notin M_t$ and $m \in M_{t+1}$), the agent first updates its global estimators as $\tilde{\mu}_i^m(t+1) = \tilde{\mu}_i(t)$ and $N_{m,i}(t+1) = \tilde{N}_i(t)$ where $\tilde{\mu}_i(t)$ and $\tilde{N}_i(t)$ are the global reward and sample count estimators of arm i after the consensus step at time t. Then, the agent initializes its local sample counts and reward estimators for arm i as: $n_{m,i}(t) = 0, \bar{\mu}_i^m(t) = 0$.

For existing agents that have not departed at time t (i.e., $m \in M_{t-1}$ and $m \in M_t$), they proceed directly to the next stage of arm selection.

Arm selection. At each time step, agents choose which arm to pull, with the strategy depending on whether the agent is honest or malicious. This procedure builds upon Xu and Klabjan (2024), with the key modifications being in the construction of $\tilde{\mu}_i^m$ and $F(m, i, t-1)$. Honest agents adopt a UCB-style policy: during the burn-in phase, each honest agent m cycles through the arms via $a_m^t = t \bmod K$. In the learning phase, each arm i is assigned a score, and the agent selects the arm with the maximum score, i.e., $a_m^{t+1} = \arg\max_i \left(\tilde{\mu}_i^m(t) + F(m, i, t)\right)$, where $\tilde{\mu}_i^m(t)$ denotes the reward estimator maintained by agent m, and $F(m, i, t) = \left(\frac{C_1 \log t}{\tilde{n}_i(t)}\right)^\beta$, with constants C_1 and β newly designed in the theoretical statements.

By contrast, malicious agents j adopt arbitrary (adversarial) strategies, often referred to as Byzantine attacks, formalized as $a_j^t = h_j^t(\mathcal{F}_t) \in [K]$, where $\mathcal{F}_t$ represents the information history up to time t, including the blockchain state and any additional data disseminated by other agents.

Broadcasting. As in Xu and Klabjan (2024), agents enter a broadcasting phase to exchange information (see details therein). Each agent sends its local estimators to the validators (selection is referred to Xu and Klabjan (2024)) for aggregation: a malicious agent j may broadcast $\bar{\mu}_i^j(t)$ and $\tilde{\mu}_i^j(t)$ using black-box strategies (e.g., Byzantine or backdoor attacks), while an honest agent m truthfully reports $\bar{\mu}_i^m(t)$ and $\tilde{\mu}_i^m(t)$.

Aggregation. A key feature of the algorithm in permissioned blockchains is reduced reliance on historical data during aggregation, which improves robustness against adversarial attacks. Following this principle, we adopt a similar aggregation rule to that in Xu and Klabjan (2024), with the key difference that aggregation is now performed over all active agents $m \in [M_t]$, rather than a fixed set.

In the aggregation step, validators combine the information they receive. For each honest validator j, two sets, $\mathcal{A}_t^j$ and $\mathcal{B}_t^j$, are constructed as follows.

For $t > L$, the set $\mathcal{A}_t^j$ is defined by $m \in \mathcal{A}_t^j \Leftrightarrow N_{m,i}(t) > \frac{N_{j,i}(t)}{k_i(t)}$ for every i, where $k_i(t) \geq \max_{k \in M} \frac{N_{k,i}(t) K}{L}$ is a threshold parameter that can be securely computed via a multi-party computation protocol as in Asharov et al. (2012), ensuring privacy without revealing the exact values of $N_{m,i}(t)$. Specifically, each agent m submits $N_{m,i}(t)$ and $k_i(t)$ to the protocol, which outputs whether $m \in \mathcal{A}_t^j$. The set $\mathcal{B}_t^j$ is then determined based on the size of $\mathcal{A}_t^j$. If $|\mathcal{A}_t^j| > 2f$, where $f = M_t^A$, and $t > L$ (i.e., during the learning phase), then $\mathcal{B}_t^j = \bigcup_i \{(m, \bar{\mu}_i^m(t)) : \bar{\mu}_i^m(t)$ is smaller than the top f values in $\mathcal{A}_t^j$ and larger than the bottom f values in $\mathcal{A}_t^j\}$. O.w., during the burn-in phase, we set $\mathcal{B}_t^j = \{t \bmod K\}$ and $\mathcal{A}_t^j = \emptyset$.

As before, malicious agents may construct their own sets $\mathcal{A}_t$ and $\mathcal{B}_t$ arbitrarily in a black-box manner.

Consensus. The consensus protocol is central to the operation of the blockchain and ensures the security of the chain. Assuming that the public keys of all agents are known to all agents, we adopt the same consensus protocol used in the permissioned blockchain framework presented in Xu and Klabjan (2024). The rationale behind this assumption is as follows. It aligns with existing work on quasi-permissionless blockchains, such as Algorand, where public key availability is standard. Moreover, as noted in Lewis-Pye and Roughgarden (2023), for permissionless blockchains without a public key list, such as dynamically available or fully permissionless systems, it is currently infeasible to design provably secure consensus mechanisms.

Global update. The set $\mathcal{B}_t$ is forwarded to the validators, who then compute the global update by averaging the estimators contained in $\mathcal{B}_t$. Define $N_i(t) = \tilde{n}_i(t) = \sum_{m \in \mathcal{B}_t} n_{m,i}(t)$ as the total number of times arm i has been pulled by the agents in $\mathcal{B}_t$. Accordingly, for each arm i at time t, the estimator is given by: $\tilde{\mu}_i(t) = \frac{1}{2}(\hat{\mu}_i(t) + \tilde{\mu}_i(\tau)), \hat{\mu}_i(t) = \frac{\sum_{m \in \mathcal{B}_t} n_i^m(t-1) \bar{\mu}_i^m(t-1)}{\tilde{n}_i(t)}$ where $\tau = \max_{s<t} \{b_s = 1\}$ is the most recent time a block was approved. When $\mathcal{B}_t$ is empty: $\tilde{\mu}_i(t) = \infty, \hat{\mu}_i(t) = \infty$.

Block verification. Smart contracts support both permissionless and permissioned blockchains as in Liang et al. (2024). A block is approved (i.e., $b_t = 1$) if the condition $\tilde{\mu}_i(t) \leq 2$ is satisfied. Otherwise, the block is disapproved, and $b_t = 0$.

Block operation. The smart contract outputs the validated estimator $\tilde{\mu}_i(t)$, the set $\mathcal{B}_t$, and the approval indicator b_t, and sends this information to the environment. Based on the output, the environment performs the block operation, distributing rewards accordingly. Unlike Xu and Klabjan (2024), we remove the cost mechanism here, since both honest and malicious agents enter the system through the dynamic agent mechanism which may not necessarily follow a cost mechanism. This removes the need to incentivize malicious agents, since agents are free to depart at any time, and marks a key algorithmic improvement. Case 1: If $b_t = 1$, the environment allocates $r^m_{a^t_m}(t)$ to agent m ($1 \le m \le M_t$). Case 2: If $b_t = 0$, then no agent receives a reward.

Information update. After receiving information from the environment, agents update their estimators: For $\tilde{\mu}^m_i(t)$, a honest agent updates it upon receiving $\tilde{\mu}_i(t)$, i.e., $\tilde{\mu}^m_i(t) = \tilde{\mu}_i(t)$; o.w., it sets $\tilde{\mu}^m_i(t) = \bar{\mu}^m_i(t)$. The number of arm pulls and the local reward estimators are updated similarly: $n_{m,i}(t) = n_{m,i}(t-1) + 1_{b_t=1} \cdot 1_{a^t_m=i}, \bar{\mu}^m_i(t) = \frac{\bar{\mu}^m_i(t-1)+r^m_{a^t_m}(t)\cdot 1_{a^t_m=i}}{n_{m,i}(t)}$. Malicious agents update information arbitrarily: it is defined as $f^{t+1}_j = g_t(f^t_j, F_t)$, where f^t_j is the information maintained by the malicious agents at time t, a 4-dim message tuple $(n_{m,i}(t), N_{m,i}(t), \bar{\mu}^m_i(t), \tilde{\mu}^m_i(t))$. The term F_t is the information history (filtration) in $R^{4K\sum^t_{s=1} M_s}$ (i.e., K arms' message across M_s agents over t steps), and g_t is any mapping from $R^{4\times 4K\sum^t_{s=1} M_s}$ to message space R^4.

Remark. *Our method is fundamentally different from that of (Xu and Klabjan, 2024). First, the aggregations in A_t and B_t use info. from the currently active agents $N_{m,i}$, not the local counts $n_{m,i}$. This seemingly intuitive change introduces nontrivial analytical challenges. Second, newly arriving agents fetch info. from the chain, avoiding the synchronization required in permissioned cases and ensuring robustness, since all on-chain info. is verified. Third, our estimators and updates $\hat{\mu}_i(t)$ explicitly account for the changing agent set through weighted averages, unlike their vanilla average. Lastly and most importantly, a key distinction lies in the block operation: theirs relies on a cost mechanism agreed upon by all agents to maintain a fixed agent set. The env. must assign extra rewards for participation, depending on each agent's contribution, an extremely hard quantification in real-world. Their analysis also depends on assumptions on the mechanism, e.g., distance-based or constant costs. In contrast, our open system imposes no such cost mechanism or extra reward structure and allows agents to join and leave freely; malicious agents and the env. are not required to comply with any such mechanism. This removes a stringent assumption and leads to a fundamentally different algorithm and analysis.*

4 REGRET UPPER BOUNDS

In this section, we analyze the theoretical effectiveness of the proposed methods across different settings by establishing their regret upper bounds. Our evaluation metric, group-wise regret, aligns with existing work on multi-agent multi-armed bandits (except that we allow for a dynamic set of agents) and serves as the standard measure for validating algorithms. We present regret guarantees for both the MA-MAB setting in Section 4.1 and the permissionless blockchain-based MA-MAB setting (PB-MA-MAB) in Section 4.2. Proofs of all theoretical results, as well as remarks on the novelty and proof intuition, and Table 2 summarizing these technical results, are provided in Appendix F.

4.1 Open Multi-agent Multi-armed Bandit

We begin with MA-MAB involving a dynamic set of agents. By tuning the arrival and departure rates λ_A and λ_D, our framework offers flexibility and generality while being consistent with existing models under specific settings. The absence or presence of departures yields fundamentally different agent sets; accordingly, we present separate results in Section 4.1.1 and 4.2.1.

If there is no departure (i.e., $\lambda_D = 0$), the agent set becomes monotone non-decreasing over time. This aligns with the assumption made in Nakamura et al. (2023), effectively providing a solution under that assumption. However, we emphasize that our stochastic setting differs significantly and leads to more refined regret bound compared to $\sqrt{T}$, which underscores the novelty of our work. We study this case in Section 4.1.1.

In contrast, when a departure process is present (i.e., $\lambda_D > 0$), the agent set is no longer monotone non-decreasing, and a previously departed agent may rejoin the game at a later time. These behaviors directly contradict the assumption in Nakamura et al. (2023) and motivate the new agent mechanisms developed and analyzed in our work. The corresponding results are presented in Section 4.2.1.

4.1.1 Without Departures

When $\lambda_D = 0$, the number of agents grows over time, which can potentially cause regret to increase. Our proposed algorithm addresses this challenge, yielding the following sublinear regret upper bound.

Theorem 1. *Let us assume that the parameter of the Poisson process for departures is $\lambda_D = 0$, while that for arrivals satisfies $\lambda_A \ge 1$. Let us assume that $L = [\frac{\log T}{M_0}]$. Then the regret of the proposed algorithm satisfies* $E[R_T|A] \le (M_0 + \lambda_A \cdot M_0 \cdot [\frac{\log T}{M_0}])[\frac{\log T}{M_0}] + 1 + \sum_i [\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3} = O(M_0 + \log T + \lambda_A \frac{\log T^2}{M_0}) = \max\{O(M_0), O(\log T), O(\frac{\log T^2}{(M_0)}) \cdot 1_{\lambda_A > 0}\}$ *where* $A =$

$\{|M_L - M_0 - \lambda_A \cdot \log T| \leq \frac{1}{2}\}$ *captures the arrival randomness, and* $P(A) \geq 1 - \frac{2}{T^2}$, $C_1 = 2$ *and* $\beta = \frac{1}{2}$.

Remark (Comparison with the existing work)**.** *Unlike existing work on bandits with a dynamic agent set (Nakamura et al., 2023), our setting is stochastic, with rewards drawn from time-invariant distributions. This enables a more refined, instance-dependent regret upper bound of order* $\log T$*, rather than* $\sqrt{T}$*. We also introduce a novel stochastic agent mechanism without the stringent assumptions of Nakamura et al. (2023) and account for it in the regret bound, where complexity in the number of agents is governed by* M_0 *and* λ_A*. Relative to fixed-agent models (Martínez-Rubio et al., 2019), we further establish a* $\log T$ *regret bound when* $\lambda_A = 0$*, which uniquely incorporates dependence on both* M_0 *and* λ_A*, a factor previously overlooked.*

Remark. *It is worth noting that three dominant terms appear in our bound:* $\log^2 T \cdot 1_{\lambda_A > 0}$, $\log T$, *and* M_0*. When* $\lambda_A > 0$*, this differs from the existing regret bound of order* $\log T$ *in the setting of multi-agent multi-armed bandits with a fixed set of agents. The difference mainly arises because newly arriving agents contribute additional regret whenever* $\lambda_A > 0$*. Intuitively, while new agents provide extra information about the arms, they also increase the group regret by definition. This highlights a fundamental trade-off between information gain and regret growth in the number of agents, which may explain the extra* $\log T$ *factor when* M_0 *is small. By contrast, when* M_0 *is large, the regret is dominated by* M_0*, a phenomenon not captured in earlier studies, where* M_0 *is typically treated as a constant. In practice, however,* M_0 *may depend on the horizon. Importantly, when* $\lambda_A = 0$*, our regret bound is of order* $\max\{O(\log T), M_0\}$*, consistent with existing results in terms of* T*, but with an additional dependence on the number of agents* M_0*, which the existing literature does not account for.*

Remark (Improvement over the naive regret bound)**.** *Notably, it improves upon the naive upper bound, which is linear in the number of agents and arises when ignoring information aggregation or fetching across agents, even when each agent plays optimally and incurs the minimal* $\log T$ *regret. Precisely, if existing agents do not aggregate, or if new agents act from scratch without fetching, the total regret is the per-agent regret (*$\log T$*) multiplied by the agent size (*$\approx M_0 + T$ *by* $2T/\lambda$ *rounds; w.l.g.* $M_0 \leq T$*), yielding* $(M_0 + T)\log T$*. The numerical comparisons with existing methods (either a fixed or bounded agent set), further validate this: methods that do not properly handle new agents incur larger regret.*

4.1.2 With Departures

We now consider the general case with $\lambda_D \geq 0$, incorporating departures into the regret analysis.

Theorem 2. *Let us assume that the arrival and departure processes are Poisson with rates* $\lambda_A \geq 0$ *and* $\lambda_D \geq 0$*, respectively. Let us suppose that* $(\sqrt{(\lambda_A)} - \sqrt{(\lambda_D)})^2 \geq 1$*. Then we have* $E[R_T|A] \leq (M_0 + (\lambda_A - \lambda_D) \cdot M_0 \cdot [\frac{\log T}{M_0}])[\frac{\log T}{M_0}] + 1 + \sum_i [\frac{4C_1 \log T}{\Delta_i^2}] + \frac{\pi^2}{3} = O(M_0 + \log T + (\lambda_A - \lambda_D)\frac{\log T^2}{M_0}) = \max\{O(M_0), O(\log T), O(\frac{\log T^2}{(M_0)})1_{\lambda_A - \lambda_D > 0}\}$ *where* $A = \{|M_t - M_0 - (\lambda_A - \lambda_D) \cdot t| \leq \frac{1}{2}, \forall 1 \leq t \leq T\}$ *captures the randomness in the arrival/departure process, and* $P(A) \geq 1 - \frac{2}{T^2} - \frac{1}{T(\sqrt{(\lambda_A)} - \sqrt{(\lambda_D)})^2}$, $C_1 = 2$ *and* $\beta = \frac{1}{2}$.

Remark (Discussion on the order)**.** *The regret remains of the same order as in the case* $\lambda_D = 0$*. However, the complexity is now determined by* $\lambda_A - \lambda_D$*, indicating that departures reduce the regret upper bound. A stronger dominating term arises only when* $\lambda_A > \lambda_D$ *instead of when* $\lambda_A > 0$*. This again highlights the trade-off between gaining more information from additional agents and incurring less regret with fewer agents.*

4.2 Permissionless Blockchain Bandit

As introduced earlier, the open agent mechanism is inspired by permissionless blockchains. Hence, as an application of the open system, we now analyze the methods for robust MA-MAB with a permissionless blockchain that presents an open system and agents may behave maliciously in multiple ways. We first present the regret guarantee without departures for both honest and malicious agents (i.e., the departure rates $\lambda_D^H, \lambda_D^A = 0$) in Section 4.2.1, and then examine the general setting with departures (i.e., the departure rates $\lambda_D^H, \lambda_D^A \geq 0$) in Section 4.2.2.

4.2.1 Without Departures

Notably, even in the absence of a departure process, introducing malicious components into the system brings in two key parameters: λ_A^H and λ_A^H, which denote the arrival rates of honest and malicious agents, respectively. The relationship between these two parameters determines the ratio of honest to malicious agents, making it a crucial aspect of the blockchain structure and a key factor in the regret upper bounds of the algorithms.

Consequently, we show that the regret bound for the permissionless blockchain can be derived from that of the permissioned blockchain, under standard assumptions with appropriate modifications. To the best of our knowledge, this constitutes the first theoretical result in the context of permissionless blockchains. The formal statement is presented below.

Theorem 3. *Let us assume that* $\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)} \geq 1$*, and that* $M_0^H > 2M_0^A$*. Further, let us suppose that the malicious agents are capable of performing existential forgery on the signatures of honest agents under an adaptive chosen-message attack. Finally, let us assume that all agents operate within a standard*

universal composability framework when constructing $\mathcal{A}_t$ (as defined in the algorithm) for any t. We then obtain the following result on the regret: $E[R_T|A] \leq (c+1)\cdot(M_0^H + (\lambda_A^H)\cdot M_0^H \cdot L)\cdot L + K([\frac{4C_1 \log T}{\min_i \Delta_i}] + \frac{\pi^2}{3} C \max_i \Delta_i) + (M_0^H) + \lambda_A^H + 1 \leq O(M_0^H + \log T + \lambda_A^H + \lambda_A^H \cdot \frac{\log T^2}{(M_0^H)}) = \max\{O(M_0^H), O(\log T), O(\frac{\log T^2}{(M_0^H)})1_{\lambda_A^H > 0}\}$ where L is the length of the burn-in period $\frac{\log T}{M_0^H}$, $c = 1$, C_1 meets the condition that $\frac{C_1}{6|M_0^H|k_i\sigma^2} \geq 1$, $\sigma^2 \geq \frac{1}{M_0^H}$, Δ_i is the sub-optimality gap, l is the length of the signature of the agents, and k_i is the threshold parameter used in the construction of A_t. Here the high probability event A is defined as $A = \{\forall 1 \leq t \leq T, b_t = 1, M_t^H > 2M_t^A\}$; $P(A) \geq 1 - \frac{1}{Tl^{T-1}} - \frac{1}{T(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2}$.

It is worth noting that the dominant term in the regret bound is of the same order as in the bound without blockchains (i.e., Theorem 1). In addition, the non-dominant term $M_0^H + \lambda_A^H + 1$, capture the additional complexity introduced by permissionless blockchains.

4.2.2 With Departures

Next, we consider the PB-MA-MAB setting with departures ($\lambda_D^H, \lambda_D^A \geq 0$). The result for the case without departures extends here by incorporating the departure process into the analysis, while the dominant term remains unchanged. The statement is as follows.

Theorem 4. *Let us assume that the malicious agents perform existential forgery on the signatures of honest agents with an adaptive chosen message attack. Lastly, let us assume that the agents are in a standard universal composability framework when constructing $\mathcal{A}_t$ for any t. Then we have that $E[R_T|A] \leq (c+1)\cdot(M_0^H + (\lambda_A^H - \lambda_D^H)M_0 \cdot L)\cdot L + K([\frac{4C_1 \log T}{\min_i \Delta_i}] + \frac{\pi^2}{3} C \max_i \Delta_i) + (M_0^H) + (\lambda_A^H - \lambda_A^D) + 1 \leq O(M_0^H + \log T + (\lambda_A^H - \lambda_D^H) + (\lambda_A^H - \lambda_D^H)\cdot \frac{\log T^2}{(M_0^H)}) = \max\{O(M_0^H), O(\log T), O(\frac{\log T^2}{(M_0^H)})1_{\lambda_A^H - \lambda_D^H > 0}\}$ where L is the length of the burn-in period $\frac{\log T}{M_0^H}$, $c = 1$, C_1 meets the condition that $\frac{C_1}{6|M_0^H|k_i\sigma^2} \geq 1$, $\sigma^2 \geq \frac{1}{M_0^H}$, Δ_i is the sub-optimality gap, l is the length of the signature of the agents, and k_i is the threshold parameter used in the construction of $\mathcal{A}_t$. Here the set A is defined as $A = \{\forall 1 \leq t \leq T, b_t = 1, M_t^H > 2M_t^A\}$ which satisfies that $P(A) \geq 1 - 1/(Tl^{T-1}) - 1/(T(\sqrt{(\lambda_A^H + \lambda_D^A)} - \sqrt{(2\lambda_A^A + 2\lambda_D^H)})^2)$.*

5 REGRET LOWER BOUNDS

We first claim that the regret lower bounds for MA-MAB also apply to PB-MA-MAB by replacing M_0 with M_0^H, since the regret in the latter is always at least as large as in the former. The formal result, via information-theoretic arguments, is given below. All proof steps are provided in Appendix F.

Theorem 5. *For decentralized multi-agent problems with any number of malicious agents, we have $R_T \leq R_T^{PB}$ for all instances, where R_T^{PB} denotes the regret in PB-MA-MAB with the same initial number of honest agents as in MA-MAB.*

As a result, we focus on characterizing regret lower bounds for the MA-MAB setting and omit the corresponding statements for PB-MA-MAB.

For MA-MAB, we establish regret lower bounds for both the case without departures and the case with departures, since it is not a priori clear, nor mathematically straightforward, which one dominates in terms of regret. These two cases are presented in Section 5.1 and Section 5.2, respectively.

5.1 Without Departures

Traditionally, lower bounds characterize the regret in the worst-case scenario (with respect to the problem instance) for any consistent algorithm. Notably, this does not hold for all algorithms—for example, a deterministic algorithm may achieve zero regret in certain cases but incur regret of order T in others. Such algorithms tend to be unstable and fail to generalize across problem instances. By consistency, we refer to algorithms that achieve $o(T^a)$ regret for all instances and for any positive real number a, algorithms that consistently achieve sublinear regret.

Focusing first on the setting without departures (i.e., $\lambda_D = 0$), we analyze general consistent algorithms and establish two types of regret lower bounds, in terms of both M_0 and T. These capture two layers of problem complexity, extending beyond the traditional focus solely on T with a fixed set of agents. Notably, even deriving a regret lower bound matching existing MA-MAB results requires new techniques, as known bounds may not hold with an evolving agent set.

Theorem 6. *Let us assume that the algorithms we consider are consistent in the sense that $E[R_T|A] \leq (\lambda_A \cdot S_T + T \cdot M_0)^a$ for any $a > 0$ where $S_T = \frac{T(T+1)}{2}$. Then we have that $E[R_T|A] \geq O(\log T + \log(\lambda_A \cdot T + M_0))\}$.*

Theorem 7 (Corollary)**.** *Assuming the same condition as in Theorem 6, $E[R_T|A] \geq O(M_0)$.*

Our proofs adapt existing analyses by focusing on the information likelihood & the algorithm class, with a key novelty: *While existing work on lower bounds over the past decades has relied on the commonly used family of consistent algorithms, we prove the explicit connection between lower bounds and algorithm classes.*

The dominant terms in the two theorems above are M_0 and $\log T$, which align with the regret upper bounds up to a $\log T$ factor. This demonstrates the near-optimality of our algorithm but also suggests room for improvement through tighter lower bounds. To address this, we establish regret lower bounds that exactly

match the upper bounds *by modifying the assumptions on algorithm consistency*: instead of requiring $o(T^a)$ regret for all $a > 0$, we allow the exponent a to exceed a threshold. This new class of algorithms may include more suboptimal cases, enabling stronger lower bounds. Our approach provides a new perspective for regret lower bounds, even in the single-agent setting.

Theorem 8. *Assume the same condition as in Theorem 6 with* $a > \frac{M_0-2}{M_0-1}$. *It then follows that,* $E[R_T|A] \geq O(M_0) \cdot \log T \geq O(M_0) + O(\log T)$.

Theorem 9 (Corollary)**.** *Under Theorem 6's conditions with* $a > \frac{logT-2}{\log T-1}$, $E[R_T|A] \geq O(\log^2 T)$.

Remark. *The lower bound on the power index a depends on considering* M_0 *or* T, *thereby uniquely highlighting the connection between problem complexity (regret lower bounds) and the class of algorithms. In both cases, the regret lower bounds,* $M_0 + \log T$ *and* $\log^2 T$, *match the upper bounds of order at most* $\max\{O(\log^2 T), O(M_0)\}$. *Moreover, our algorithm is consistent under the new definition of consistency, the above lower bounds also apply, thereby establishing the (near-) optimality of our algorithm. Deriving matching bounds for all consistent algorithms remains open.*

5.2 With Departures

With $\lambda_D \geq 0$, the regret lower bounds remain unchanged, since departures do not affect sample complexity or their order. Theorems 10, 11, 12, and 13 present these bounds; they are mostly **corollaries** of the lower-bound results in Section 5.1: 10 and 11 for general consistent algorithms and 12 and 13 under the new consistency assumptions.

Theorem 10 (Corollary)**.** *Let us assume that* $\lambda_A \geq \lambda_D$. *Let us assume that the algorithms we consider are consistent in the sense that* $E[R_T|A] \leq ((\lambda_A - \lambda_D) \cdot S_T + T \cdot M_0)^a$ *for any* $a > 0$ *where* $S_T = \frac{T(T+1)}{2}$. *Then we have that* $E[R_T|A] \geq O(\log T + \log((\lambda_A - \lambda_D) \cdot T + M_0))$.

Theorem 11 (Corollary)**.** $E[R_T|A] \geq O(M_0)$.

Theorem 12 (Corollary)**.** *Assuming the same condition as in Theorem 10 with any* $a > \frac{M_0-2}{M_0-1}$, $E[R_T|A] \geq O(M_0) \cdot O(\log T) \geq O(M_0) + O(\log T)$.

Theorem 13 (Corollary)**.** *Under the same condition as in Theorem 10 with* $a > \frac{logT-2}{\log T-1}$, $E[R_T|A] \geq O(\log^2 T)$.

Remark. *Importantly, the regret upper bound established for our algorithm ensures that the policy is consistent under the new definition of consistency. As a result, the preceding lower bounds hold in this setting as well, confirming the (near-)optimality of our algorithm in both asymptotic and finite-time analyses given the newly added dimension of departure processes.*

For the established upper and lower bounds across various settings, we refer to Table 1 in Appendix A for a summary of the theorems, implying the comprehensiveness and tightness of the results.

6 NUMERICAL EXPERIMENTS

We compare our algorithm to the existing benchmarks. The simulation evaluated ours, SE-AAC-ODC (Wang et al., 2025), UCB-ODC (Chen et al., 2023), and Randomized KL-UCB (Trinh and Combes, 2021). Regarding scalability with respect to λ, we observe that even a moderate arrival rate $\lambda = 0.2$ can significantly disrupt the benchmark algorithms, while our method remains stable under the same conditions (our sensitivity study below even pushes λ to be 2.5). To ensure a fair comparison, we conducted the experiments using a relatively small value of λ, and we expect the performance improvement of our algorithm to become even more pronounced as λ increases.

A plot visualizing regret is in Appendix D; among all algorithms, ours exhibits substantially smaller regret and even tighter confidence intervals, indicating both superior performance and greater stability. In contrast, the benchmark algorithms display a clear linear trend. These observations demonstrate both the necessity of developing new algorithms for settings with agent dynamics and the sufficiency of our proposed method in addressing these challenges.

We also report the table (Appendix D) summarizing the regret of the methods, along with the improvement of our algorithm relative to each benchmark. We observe that the regret improvement (defined as the absolute regret difference divided by our algorithm's regret) increases as the time horizon grows and becomes highly significant, as the agent population expands over time. The average improvement over the second-best approach (UCB-ODC (Chen et al., 2023)) is 60% for $0 < t \leq 2.5K$, rising to 215% for $2.5K < t \leq 5K$. This highlights the increased benefits of accounting for agent dynamics and more precisely demonstrates the superiority of our algorithm compared to the benchmarks.

In addition, we conduct further experiments on real-world data beyond simulations, using a real-world e-commerce dataset where customers arrive and make purchases each day, naturally forming a dynamic agent model. An example of such data is the Brazilian E-Commerce Public Dataset by Olist on Kaggle. The results are also shown in the table using the real dataset (with a total of 609 time steps; Appendix D). Our algorithm demonstrates significant improvements over the existing benchmarks (at least 130%). We also evaluate on a simulator derived from the real-world dataset at a longer horizon ($T = 5000$), where our algorithm again outperforms benchmarks. These strongly support the practical effectiveness and significance of our method.

We also present a comprehensive study on the parameters T, M_0, and λ, and defer the details to Appendix D.

Acknowledgments

We deeply appreciate the valuable comments and constructive suggestions from the meta-reviewer and reviewers, which have greatly helped polish this paper and make this final version possible.

References

Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. Multi-agent multi-armed bandits with limited communication. *Journal of Machine Learning Research*, 23(212):1–24, 2022.

Mohammad Javad Amiri, Divyakant Agrawal, and Amr El Abbadi. Parblockchain: Leveraging transaction parallelism in permissioned blockchain systems. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1337–1347. IEEE, 2019.

Rodelio Arenas and Proceso Fernandez. Credenceledger: a permissioned blockchain for verifiable academic credentials. In *2018 IEEE international conference on engineering, technology and innovation (ICE/ITMC)*, pages 1–6. IEEE, 2018.

Gilad Asharov, Abhishek Jain, Adriana López-Alt, Eran Tromer, Vinod Vaikuntanathan, and Daniel Wichs. Multiparty computation with low communication, computation and interaction via threshold fhe. In *Advances in Cryptology–EUROCRYPT 2012: 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings 31*, pages 483–501. Springer, 2012.

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32 (1):48–77, 2002b.

Benjamin Avi-Itzhak and Daniel P Heyman. Approximate queuing models for multiprogramming computer systems. *Operations Research*, 21(6):1212–1230, 1973.

Ehtesamul Azim, Dongjie Wang, Tae Hyun Hwang, Yanjie Fu, and Wei Zhang. Biological pathway guided gene selection through collaborative reinforcement learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 4250–4260, 2025.

Andrew D Barbour. Stein's method and poisson process convergence. *Journal of Applied Probability*, 25(A): 175–184, 1988.

Andrea Barraza-Urbina and Dorota Glowacka. Introduction to bandits in recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 748–750, 2020.

Nikhil Behari, Edwin Zhang, Yunfan Zhao, Aparna Taneja, Dheeraj Nagaraj, and Milind Tambe. A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health. *Advances in Neural Information Processing Systems*, 37:3964–4002, 2024.

Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

Djallel Bouneffouf and Raphaël Féraud. A tutorial on multi-armed bandit applications for large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6412–6413, 2024.

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. In *Conference on Learning Theory*, pages 122–134. PMLR, 2013.

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International conference on artificial intelligence and statistics*, pages 3471–3481. PMLR, 2020.

Louis HY Chen. On the convergence of poisson binomial to poisson distributions. *The Annals of Probability*, pages 178–180, 1974.

Yu-Zhen Janice Chen, Lin Yang, Xuchuang Wang, Xutong Liu, Mohammad Hajiesmaili, John CS Lui, and Don Towsley. On-demand communication for asynchronous multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3903–3930. PMLR, 2023.

Erhan Cinlar. *Introduction to stochastic processes*. Courier Corporation, 2013.

David Roxbee Cox. *The theory of stochastic processes*. Routledge, 2017.

Dominic Deuber, Bernardo Magri, and Sri Aravinda Krishnan Thyagarajan. Redactable blockchain in the permissionless setting. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 124–138. IEEE, 2019.

Molud Esmaili and Ken Christensen. Performance modeling of public permissionless blockchains: A survey. *ACM Computing Surveys*, 57(7):1–35, 2025.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour.

Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Ghareeb Falazi, Michael Hahn, Uwe Breitenbücher, Frank Leymann, and Vladimir Yussupov. Process-based composition of permissioned and permissionless blockchain smart contracts. In *2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC)*, pages 77–87. IEEE, 2019.

Jun Feng, Heng Li, Minlie Huang, Shichen Liu, Wenwu Ou, Zhirong Wang, and Xiaoyan Zhu. Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning. In *Proceedings of the 2018 World Wide Web Conference*, pages 1939–1948, 2018.

Ben Fisch, Arthur Lazzaretti, Zeyu Liu, and Lei Yang. Permissionless verifiable information dispersal (data availability for bitcoin rollups). In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 1983–2001. IEEE, 2025.

Shafi Goldwasser, Silvio Micali, and Ronald L Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM Journal on computing*, 17(2):281–308, 1988.

Christine V Helliar, Louise Crawford, Laura Rocca, Claudio Teodori, and Monica Veneziani. Permissionless and permissioned blockchain diffusion. *International Journal of Information Management*, 54: 102136, 2020.

Garud Iyengar, Fahad Saleh, Jay Sethuraman, and Wenjun Wang. Economics of permissioned blockchain adoption. *Management Science*, 69(6):3415–3436, 2023.

Svante Janson. Poisson convergence and poisson processes with applications to random graphs. *Stochastic processes and their Applications*, 26:1–30, 1987.

Daniel Jiang, Haipeng Luo, Chu Wang, and Yingfei Wang. Multi-armed bandits and reinforcement learning: Advancing decision making in e-commerce and beyond. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4133–4134, 2021.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pages 203–226. 2019.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference*. 243–248. IEEE, 2016a.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control*. 167–172. IEEE, 2016b.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125: 109445, 2021.

Andrew Lewis-Pye and Tim Roughgarden. Permissionless consensus. *arXiv preprint arXiv:2304.14701*, 2023.

Jianhao Li, Hui Ma, Jiabei Wang, Zishuai Song, Wenhan Xu, and Rui Zhang. Wolverine: A scalable and transaction-consistent redactable permissionless blockchain. *IEEE Transactions on Information Forensics and Security*, 18:1653–1666, 2023.

Tan Li and Linqi Song. Privacy-preserving communication-efficient federated multi-armed bandits. *IEEE Journal on Selected Areas in Communications*, 40(3):773–787, 2022.

Wei Liang, Yaqin Liu, Ce Yang, Songyou Xie, Kuanching Li, and Willy Susilo. On identity, transaction, and smart contract privacy on permissioned and permissionless blockchain: a comprehensive survey. *ACM Computing Surveys*, 56(12):1–35, 2024.

Jiabin Lin, Karuna Anna Sajeevan, Bibek Acharya, Shana Moothedath, and Ratul Chowdhury. Distributed stochastic contextual bandits for protein drug interaction. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7160–7164. IEEE, 2024.

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Tomoki Nakamura, Naoki Hayashi, and Masahiro Inuiguchi. Cooperative learning for adversarial multi-armed bandit on open multi-agent systems. *IEEE Control Systems Letters*, 7:1712–1717, 2023.

Raghu Pasupathy. Generating homogeneous poisson processes. *Wiley Encyclopedia of Operations Research and Management Science*, 2010.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach.

In *International Conference on Machine Learning*, pages 155–163. PMLR, 2016.

Sara Rouhani, Rafael Belchior, Rui S Cruz, and Ralph Deters. Distributed attribute-based access control system using permissioned blockchain. *World Wide Web*, 24(5):1617–1644, 2021.

SM Samuels. A characterization of the poisson process. *Journal of Applied Probability*, 11(1):72–85, 1974.

Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

Ioannis Simaiakis and Hamsa Balakrishnan. A queuing model of the airport departure process. *Transportation Science*, 50(1):94–109, 2016.

Quang Tung Thai, Jong-Chul Yim, and Sun-Me Kim. A scalable semi-permissionless blockchain framework. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 990–995. IEEE, 2019.

Cindy Trinh and Richard Combes. A high performance, low complexity algorithm for multi-player bandits without collision sensing information. *arXiv preprint arXiv:2102.10200*, 2021.

Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2021.

Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent bandits over undirected graphs. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(3):1–57, 2022.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020.

Xuchuang Wang, Lin Yang, Yu-Zhen Janice Chen, Xutong Liu, Mohammad Hajiesmaili, Don Towsley, and John CS Lui. Achieving near-optimal individual regret & low communications in multi-agent bandits. In *The Eleventh International Conference on Learning Representations*, 2022.

Xuchuang Wang, Yu-Zhen Janice Chen, Xutong Liu, Lin Yang, Mohammad Hajiesmaili, Don Towsley, and John CS Lui. Asynchronous multi-agent bandits: Fully distributed vs. leader-coordinated algorithms. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 9(1):1–39, 2025.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Ding Xiang, Rebecca West, Jiaqi Wang, Xiquan Cui, and Jinzhou Huang. Multi armed bandit vs. a/b tests in e-commerce-confidence interval and hypothesis test power perspectives. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4204–4214, 2022.

Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. In *Advances on Neural Information Processing Systems*, 2023.

Mengfan Xu and Diego Klabjan. Decentralized blockchain-based robust multi-agent multi-armed bandit. *Coordination and Cooperation in Multi-Agent Reinforcement Learning (Honorable Mention)*, 2024.

Mengfan Xu and Diego Klabjan. Multi-agent multi-armed bandit regret complexity and optimality. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Mengfan Xu, Liren Shan, Fatemeh Ghaffari, Xuchuang Wang, Xutong Liu, and Mohammad Hajiesmaili. Heterogeneous multi-agent multi-armed bandits on stochastic block models. In *Proceedings of the ACM on Measurement and Analysis of Computer Systems*, 2025.

Meiyi Yang, Deyun Gao, Chuan Heng Foh, Wei Quan, and Victor CM Leung. Multi-agent reinforcement learning-based joint caching and routing in heterogeneous networks. *IEEE Transactions on Cognitive Communications and Networking*, 10(5):1959–1974, 2024.

Tongxin Zhou, Yingfei Wang, Lu Yan, and Yong Tan. Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. *Information Systems Research*, 34(4):1493–1512, 2023.

Jingxuan Zhu, Romeil Sandhu, and Ji Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *59th IEEE Conference on Decision and Control*. 3078–3083. IEEE, 2020.

Jingxuan Zhu, Alec Koppel, Alvaro Velasquez, and Ji Liu. Byzantine-resilient decentralized multi-armed bandits. *arXiv preprint arXiv:2310.07320*, 2023.

Kun Zhu, Lujiu Li, Yuanyuan Xu, Tong Zhang, and Lu Zhou. Multi-connection based scalable video

streaming in udns: A multi-agent multi-armed bandit approach. *IEEE Transactions on Wireless Communications*, 21(2):1156–1169, 2021.

Zheqing Zhu and Benjamin Van Roy. Scalable neural contextual bandit for recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3636–3646, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix:
Open Multi-agent Multi-armed Bandit with Applications in Permissionless Blockchain

A CONCLUSION AND FUTURE WORK

		Without departure	**With departure**
MA-MAB	Lower Bound	$O(\log T)$, $O(M_0)$ $O(\log T + M_0)$, $O(\log T^2)$ Theorems 6, 7, 8, 9 consistency; & modified consistency	$O(\log T)$, $O(M_0)$ $O(\log T + M_0)$, $O(\log T^2)$ Theorems 10, 11, 12, 13 consistency; & modified consistency
	Upper Bound	$\max\{O(M_0), O(\log T),$ $O(\frac{\log T^2}{(M_0)})1_{\lambda_A>0}\}$ Theorem 1 nearly tight in general; tight under assumptions	$\max\{O(M_0), O(\log T),$ $O(\frac{\log T^2}{(M_0)})1_{\lambda_A-\lambda_D>0}\}$ Theorem 2 nearly tight in general; tight under assumptions
PB-MA-MAB	Lower Bound	$O(\log T)$, $O(M_0)$ $O(\log T + M_0)$, $O(\log T^2)$ No smaller than MA-MAB consistency; & modified consistency	$O(\log T)$, $O(M_0)$ $O(\log T + M_0)$, $O(\log T^2)$ No smaller than MA-MAB consistency; & modified consistency
	Upper Bound	$O(\log^2 T)$ $\max\{O(M_0^H), O(\log T),$ $O(\frac{\log T^2}{(M_0^H)})1_{\lambda_A^H>0}\}$ Theorem 3 nearly tight in general; tight under assumptions	$O(\log^2 T)$ $\max\{O(M_0^H), O(\log T),$ $O(\frac{\log T^2}{(M_0^H)})1_{\lambda_A^H-\lambda_D^H>0}\}$ Theorem 4 nearly tight in general; tight under assumptions
Relationship:		MA-MAB serves as the baseline, while PB-MA-MAB introduces robustness challenges; departures further complicate both settings.	
Comparison:		MA-MAB has a smaller lower bound than PB-MA-MAB, since PB-MA-MAB adds robustness constraints. Theorem 5; the sufficiency of proving lower bounds for MA-MAB.	

Table 1: Regret upper and lower bounds for MA-MAB and PB-MA-MAB under different agent dynamics, along with their relationship and comparison.

In this paper, we present the first study of the open multi-agent multi-armed bandit (MA-MAB) problem with a randomly evolving set of agents, motivated by and also applied to applications in permissionless blockchains (PB-MA-MAB) where agents typically join the game randomly from a candidate pool. Unlike existing MA-MAB formulations that assume a fixed set of agents, we allow agents to arrive and depart according to a systematic and probabilistic mechanism, namely a Poisson point process. This choice is natural since the asymptotic distribution of arrivals from an infinite pool is Poisson, thereby capturing the dynamics of permissionless blockchains more accurately. This setting necessitates a new problem formulation, as the regret now grows with the number of agents, introducing significant challenges. To address these, we design new methodologies inspired by the Upper Confidence Bound (UCB) principle. Each agent selects an arm that maximizes its UCB index based on empirical reward estimators. Importantly, newly arriving agents do not start from scratch: they either adopt the estimators of agents who remain in the system or use estimators stored on the blockchain, depending on whether the context is blockchain-based or not. Meanwhile, existing agents collaborate to accelerate the learning process, reduce sample complexity, and thereby lower regret. We provide theoretical guarantees for the proposed algorithms in terms of both upper and lower bounds on regret. Specifically, we show that the regret upper bound is of order $\max\{M_0, \log T, \frac{\log T^2}{M_0} 1_{\lambda_A>0}\}$, and the lower bound is $\max\{M_0^H, \log T\}$ for consistent algorithms, implying near-optimality. Moreover, for a new class of consistent algorithms, we establish matching lower bound $\log T^2$, demonstrating optimality in this refined setting.

We summarize the statements in Table 1 for both MA-MAB and PB-MA-MAB, with and without departures, and with respect to both lower and upper bounds. The table also indicates the tightness of our theoretical results.

Looking ahead, an important direction is to explore alternative stochastic models for agent dynamics, which can further broaden the applicability of this framework to real-world multi-agent systems. Another promising avenue is to adapt existing MA-MAB techniques to handle open systems.

B RELATED WORK

Multi-agent Multi-armed Bandit with fixed agents Recently, numerous studies on MA-MAB with a fixed set of agents have gained attention, driven by the rapid development of networked systems in online learning (Landgren et al., 2016a,b, 2021; Zhu et al., 2020; Martínez-Rubio et al., 2019; Agarwal et al., 2022; Wang et al., 2022, 2020; Li and Song, 2022; Sankararaman et al., 2019; Chawla et al., 2020; Xu and Klabjan, 2023, 2024, 2025). Importantly, if the reward distributions for the same arm are identical across agents, the setting is referred to as homogeneous MA-MAB (Landgren et al., 2016a,b, 2021; Zhu et al., 2020; Martínez-Rubio et al., 2019); otherwise, it is heterogeneous MA-MAB (Xu and Klabjan, 2023). Homogeneous MA-MAB is relatively well understood and is the focus of our work. Depending on whether there is an underlying communication topology among agents, the setting can be further categorized into centralized and decentralized variants. In the centralized case, agents communicate with a central server; in the decentralized case, agents communicate directly with one another based on a graph structure. Notably, decentralized MA-MAB is more general and introduces new challenges due to varying graph topologies. Examples include fully connected graphs (Landgren et al., 2016a) (where all agents are connected), connected graphs (where a path exists between any two agents) (Wang et al., 2022; Chawla et al., 2020; Sankararaman et al., 2019), random graphs (where edges exist independently with some fixed probability) (Xu and Klabjan, 2023), and stochastic block models (where edge probabilities depend on a cluster structure) (Xu et al., 2025). Nevertheless, all of this prior work assumes a fixed set of agents—a key distinction from our setting. This difference significantly impacts both the problem formulation and the definition of regret.

Multi-agent Multi-armed Bandit with unbounded sets of agents To the best of our knowledge, only one prior work has considered a dynamic and unbounded set of agents (Nakamura et al., 2023). They recently studied the adversarial setting, while our work focuses on the stochastic setting—highlighting a key difference. More importantly, their model imposes strong assumptions, such as agents never returning once they leave the system and a monotonic number of agents over time. While these assumptions may be mathematically convenient, they fail to reflect many real-world scenarios where randomness is inherent, and such deterministic conditions often break down. In contrast, we address this challenge by modeling agent arrivals and departures as stochastic processes, allowing us to better handle noise and align with practical applications.

Multi-agent Multi-armed Bandit with bounded sets of agents Overall, compared to (Wang et al., 2025; Chen et al., 2023; Rosenski et al., 2016; Trinh and Combes, 2021), our formulation is novel in capturing real-world applications in which the platform is open to anyone worldwide without capacity constraints, the global

population evolves according to branching processes, and thus there is no precise upper bound independent of T on the total number of agents. Our regret bounds (both upper and lower) reflect this complexity, and addresses the open problem. We also address malicious scenarios. This formulation further opens the door to incorporating richer stochastic processes from queueing theory into decision-making platforms (where, instead of merely being served, agents can make decisions) as the world becomes increasingly digital and agentic.

More precisely, the main difference between our agent model and the models in (Wang et al., 2025) and (Chen et al., 2023) is that, in their settings, the active agents are subsets of a fixed set whose size is bounded by a constant M independent of T (i.e., no more than M agents participate). Moreover, (Wang et al., 2025) and (Chen et al., 2023) focus on communication cost and regret without capturing dynamic agent activation patterns, nor do they consider malicious agents. In contrast, in our setting any agent can enter the game, potentially from an infinite pool, and the number of agents grows with t according to a Poisson process, yielding roughly λT agents in expectation by the end of the game. This introduces new challenges, as the regret is defined with respect to all active agents and may grow linearly in expectation. And our focus is on the regret and its scaling under a dynamic agent model. Deriving algorithms with guarantees that (1) grow sublinearly and (2) reflect the complexity of the arrival/departure process is therefore novel; we additionally incorporate malicious agents and establish robust guarantees. Compared to (Rosenski et al., 2016), that paper also considers a changing set of agents but explicitly assumes that "Generally, we assume K, N, N_t are all much smaller than T," where N_t denotes the dynamic agent count and N the fixed set size. In (Rosenski et al., 2016), the regret scales with N and N^2, and the setting involves collisions in which two agents selecting the same arm both receive zero reward. In our setting, however, N_t can be comparable to T because arrivals follow a Poisson process, and in expectation $E[N_t] = \lambda t$ can exceed T. This directly impacts the regret, yet we show that it does not scale linearly with $E[N_t]$; instead, it scales as $\lambda \log^2 T$ or M_0. We also consider both cooperative and malicious agents, where disruptions go far beyond collisions: if a block is not approved by the disruptive behaviors of agents, no agent receives reward. Compared to (Trinh and Combes, 2021), that paper assumes a fixed, bounded set of agents and analyzes collisions without collision sensing, further creatively extending to a dynamic agent setting only where a queue with capacity K is imposed for numerical experiments. In other words, the total number of agents will not exceed K, which directly constrains the regret. Also, the analysis for this setting is pointed out as an open problem. In contrast, our work employs a fully dynamic agent model with no queue-capacity limitation on the number of agents and hence no such constraint on regret. We further theoretically provide both upper and lower bounds that capture the inherent complexity of this setting that addresses this open problem. We also conduct experiments under the dynamic regime, and additionally incorporate malicious agents.

Stochastic/Queuing systems Stochastic systems are those that involve significant randomness and uncertainty in their operations, giving rise to a broad area of research on stochastic processes, which rely on mathematical modeling and probabilistic tools (Cinlar, 2013; Cox, 2017; Simaiakis and Balakrishnan, 2016; Avi-Itzhak and Heyman, 1973). Common examples include queuing models (Simaiakis and Balakrishnan, 2016; Avi-Itzhak and Heyman, 1973) and Markov decision processes (Bellman, 1957; Wiesemann et al., 2013; Even-Dar et al., 2009). Among these, queuing models are particularly important—they describe systems where "customers" (i.e., entities in the system) arrive and are then served by "servers". In such models, both the arrival and service processes follow probabilistic patterns, with the Poisson process and compound Poisson processes being the most commonly used. Traditionally, queuing theory has focused on metrics such as customer waiting time and server utilization. However, queuing models have broader potential due to their general structure and flexibility. In particular, they have not been widely explored in settings where the "customers" are intelligent agents making online decisions—enabled by artificial intelligence—rather than passively receiving service. This opens up a new opportunity: modeling agent mechanisms in online learning systems using queuing theory. Motivated by this, we explore the use of queuing models—specifically, Poisson processes (Samuels, 1974; Pasupathy, 2010; Barbour, 1988; Janson, 1987)—to model agent arrival and departure dynamics in the multi-agent multi-armed bandit framework. This perspective could pave the way for incorporating more advanced queuing models in future research.

Robust Multi-agent Multi-armed Bandit Another key aspect of the multi-agent multi-armed bandit (MA-MAB) framework is robustness to the presence of malicious agents, which has been gaining increasing attention in digital platforms. To this end, researchers have proposed robust MA-MAB algorithms under various types of attacks, including noisy or distribution-shifted rewards, or adversarial rewards generated by malicious agents (Vial et al., 2021, 2022; Zhu et al., 2023; Xu and Klabjan, 2024). However, to the best of our knowledge, existing work does not consider cyberattacks—such as cases where the entire reward or game mechanism is

compromised by malicious agents, or where malicious agents disrupt or obscure the system itself. Recently, Xu and Klabjan (2024) explored the role of blockchain in ensuring robustness against such cyberattacks. However, their work focuses solely on permissioned blockchains, and does not address the more challenging permissionless blockchain setting. In this paper, we address this gap.

Permissioned and Permissionless Blockchain Blockchains are decentralized, secure ledgers where a group of agents (possibly malicious) perform transactions and receive feedback only if the transactions are verified as secure. Blockchains can be categorized as permissioned (Iyengar et al., 2023; Amiri et al., 2019; Arenas and Fernandez, 2018; Rouhani et al., 2021) or permissionless (Esmaili and Christensen, 2025; Thai et al., 2019; Falazi et al., 2019; Deuber et al., 2019; Liang et al., 2024; Fisch et al., 2025), where the former has a fixed, pre-selected set of agents, and the latter allows a dynamic set of agents—anyone can join at any time. While permissioned blockchains are relatively well understood in terms of both applications and methodologies, they are mostly limited to consortium-based use cases. In contrast, widely used blockchains such as Bitcoin and Ethereum are permissionless—yet remain largely unexplored in terms of agent mechanisms and learning methodologies. In the context of intelligent decision-making, Xu and Klabjan (2024) recently studied how permissioned blockchains can be integrated with a multi-agent multi-armed bandit (MA-MAB) framework. This integration not only enables robust decision-making in the presence of malicious attacks but also supports agent-level intelligence—crucial in the era of AI. However, the corresponding concept in permissionless blockchains has not yet been explored. In this work, we aim to close two gaps: (1) we show that MA-MAB with a dynamic set of agents naturally enables the characterization of agent mechanisms in permissionless blockchains, and (2) we demonstrate that permissionless blockchains can also provide robustness, in a way similar to Xu and Klabjan (2024).

The following table further highlights the contributions of this work relative to the existing literature.

Section	Existing work	Open-MA-MAB
Agent set	fixed/subset;	dynamic from an infinite pool;
Agent size	at most M	completely Poisson
Regret bounded by	M agents' regret	$E[M_T] = \lambda T$ agents' regret
Algorithm	aggregation	fetching/aggregation
Upper bound	$\log T$	$\{\log T, M_0, \lambda\frac{\log^2 T}{M_0}\}$
Lower bound	$\log T$	$\{\log T, M_0, \log^2 T\}$
Complexity of Experiments	T benchmarks	T, λ, M_0 significant improvements
References	(Wang et al., 2025; Chen et al., 2023)	this paper

Section	Existing work	PB-MA-MAB
Agent set	dynamic **bounded**;	dynamic with malice;
Agent size	at most K	completely Poisson
Regret bounded by	K agents' regret	λT agents + from corruptions
Algorithm	MC; KL-UCB	plus blockchains
Upper bound	$\sqrt{T}$ or NA	$\{\log T, M_0, \lambda\frac{\log^2 T}{M_0}\}$
Lower bound	NA	$\{\log T, M_0, \log^2 T\}$
Complexity of	T or NA	T, λ^H, λ^A, M_0^H, M_0^A
Experiments	benchmarks	in progress
References	(Rosenski et al., 2016; Trinh and Combes, 2021)	this paper

C ALGORITHM PSEUDOCODE

C.1 Open Multi-agent Multi-armed Bandit

Algorithm 1 Open-MA-MAB

Input: Horizon T, number of arms K, burn-in length L, parameters C_1, β

Initialize the initial active set M_0 and local/global estimators **for** $t = 1, 2, \ldots, T$ **do**
  Update the active set M_t **if** $t \leq L$ **then**
    `OpenBurnIn`(t, M_t, M_0, K)
  else
    `OpenSelect`(t, M_t, K, C_1, β)
  `OpenBroadcastAndAggregate`(t, M_t)
  `OpenUpdate`(t, M_t)

Algorithm 2 Open-MA-MAB subroutines

Function `OpenBurnIn`(t, M_t, M_0, K)

- **foreach** *agent* $m \in M_t$ **do**
  - **if** $m \in M_0$ **then**
    - Run a homogeneous MA-MAB algorithm on the fully connected graph
  - **else**
    - Select an arm randomly

Function `OpenSelect`(t, M_t, K, C_1, β)

- **foreach** *agent* $m \in M_t$ **do**
  - **foreach** *arm* $i \in [K]$ **do**
    - Compute
$$F(m, i, t-1) = \left(\frac{C_1 \log t}{N_{m,i}(t-1)} \right)^{\beta}$$
  - Select
$$a_t^m \leftarrow \arg\max_{i \in [K]} \left(\tilde{\mu}_i^m(t-1) + F(m, i, t-1) \right)$$
  - Pull arm a_t^m and observe reward $r_{a_t^m}^m(t)$

Algorithm 3 Open-MA-MAB aggregation and update

Function `OpenBroadcastAndAggregate`(t, M_t)

- **foreach** *agent* $m \in M_t$ **do**
  - Broadcast $\{n_{m,i}(t), N_{m,i}(t), \bar{\mu}_i^m(t), \tilde{\mu}_i^m(t)\}_{i \in [K]}$
- **foreach** *existing agent* $m \in M_t \cap M_{t-1}$ **do**
  - **foreach** *arm* $i \in [K]$ **do**
    - Update
$$N_{m,i}(t) \leftarrow \sum_{j \in \mathcal{N}_m(t)} n_{j,i}(t)$$
    - Update aggregated estimator $\hat{\mu}_i^m(t)$ using neighbors' counts, local estimators, and current rewards
- **foreach** *new agent* $m \in M_t \setminus M_{t-1}$ **do**
  - **foreach** *arm* $i \in [K]$ **do**
    - Pick an existing agent $j \in M_{t-1}$ and set
$$N_{m,i}(t) \leftarrow N_{j,i}(t), \qquad \hat{\mu}_i^m(t) \leftarrow \hat{\mu}_i^j(t)$$

Function `OpenUpdate`(t, M_t)

- **foreach** *agent* $m \in M_t$ **do**
  - **foreach** *arm* $i \in [K]$ **do**
    - Update
$$n_{m,i}(t) \leftarrow n_{m,i}(t-1) + \mathbf{1}\{a_t^m = i\}$$
    - **if** $a_t^m = i$ **then**
      - Update the local estimator $\bar{\mu}_i^m(t)$ using $r_{a_t^m}^m(t)$
    - Set $\tilde{\mu}_i^m(t) \leftarrow \hat{\mu}_i^m(t)$

C.2 Permissionless Blockchain Bandit

Algorithm 4 PB-MA-MAB

Input: Horizon T, number of arms K, burn-in length L, parameters C_1, β

Initialize the blockchain state and the honest/malicious active sets **for** $t = 1, 2, \ldots, T$ **do**

  Update the active set M_t `PBFetch`(t, M_t) `PBSelect`$(t, M_t, K, L, C_1, \beta)$ `PBAggregate`(t, M_t) `PBConsensusUpdate`(t) `PBLocalUpdate`(t, M_t)

Algorithm 5 PB-MA-MAB: fetch and arm selection

Function `PBFetch`(t, M_t)

  foreach *new honest agent* $m \in M_t \setminus M_{t-1}$ **do**

    foreach *arm* $i \in [K]$ **do**

      Fetch on-chain information and set

$$\tilde{\mu}_i^m(t) \leftarrow \tilde{\mu}_i(t-1), \qquad N_{m,i}(t) \leftarrow \tilde{N}_i(t-1)$$

      Initialize

$$n_{m,i}(t) \leftarrow 0, \qquad \bar{\mu}_i^m(t) \leftarrow 0$$

Function `PBSelect`$(t, M_t, K, L, C_1, \beta)$

  foreach *honest agent* $m \in M_t^H$ **do**

    if $t \le L$ **then**

      Select

$$a_t^m \leftarrow t \bmod K$$

    else

      foreach *arm* $i \in [K]$ **do**

        Compute

$$F(m, i, t) = \left(\frac{C_1 \log t}{\tilde{n}_i(t)} \right)^{\beta}$$

      Select the arm with the largest UCB index

  foreach *malicious agent* $j \in M_t^A$ **do**

    Choose an arm adversarially according to the maintained information filtration

  foreach *agent* $m \in M_t$ **do**

    Pull arm a_t^m

Algorithm 6 PB-MA-MAB: broadcasting, trusted aggregation, and consensus

Function `PBAggregate`(t, M_t)

- **foreach** *honest agent* $m \in M_t^H$ **do**
  - Broadcast truthful local/global estimators to validators
- **foreach** *malicious agent* $j \in M_t^A$ **do**
  - Broadcast arbitrary messages
- **foreach** *honest validator* $j \in S_V(t)$ **do**
  - **if** $t > L$ **then**
    - Construct the trusted set $\mathcal{A}_t^j$ **if** $|\mathcal{A}_t^j| > 2f$ **then**
      - Trim the largest f and smallest f values and obtain $\mathcal{B}_t^j$
    - **else**
      - Set $\mathcal{B}_t^j \leftarrow \emptyset$
  - **else**
    - Set $\mathcal{A}_t^j \leftarrow \emptyset$ and $\mathcal{B}_t^j \leftarrow \{t \bmod K\}$

Function `PBConsensusUpdate`(t)

- Run the blockchain consensus protocol among validators and commanders
- Produce the agreed set $\mathcal{B}_t$
- **foreach** *arm* $i \in [K]$ **do**
  - Compute $\tilde{n}_i(t)$ from $\mathcal{B}_t$ **if** $\mathcal{B}_t \neq \emptyset$ **then**
    - Compute $\hat{\mu}_i(t)$ and update $\tilde{\mu}_i(t)$
  - **else**
    - Set $\hat{\mu}_i(t) \leftarrow \infty$ and $\tilde{\mu}_i(t) \leftarrow \infty$
- Verify the block and determine b_t

Algorithm 7 PB-MA-MAB: block operation and local information update

Function `PBLocalUpdate`(t, M_t)

- Output $\tilde{\mu}_i(t)$, $\mathcal{B}_t$, and b_t via the smart contract **if** $b_t = 1$ **then**
  - Allocate rewards to all active agents
- **else**
  - No agent receives reward
- **foreach** *honest agent* $m \in M_t^H$ **do**
  - **foreach** *arm* $i \in [K]$ **do**
    - Update the maintained global estimator using the chain output if received; otherwise use the local estimator Update
      $$n_{m,i}(t) \leftarrow n_{m,i}(t-1) + \mathbf{1}\{b_t = 1\}\mathbf{1}\{a_t^m = i\}$$
    - **if** $a_t^m = i$ *and* $b_t = 1$ **then**
      - Update the local estimator using the received reward
- **foreach** *malicious agent* $j \in M_t^A$ **do**
  - Update information arbitrarily

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 Comparison with benchmarks

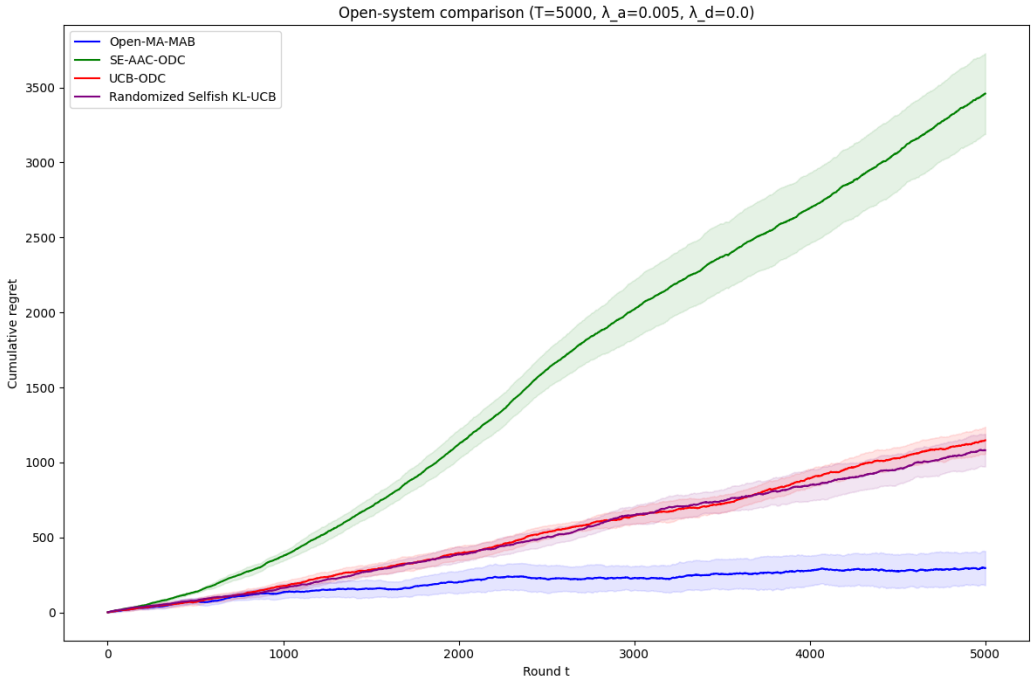

Figure 1: Synthetic Dataset.

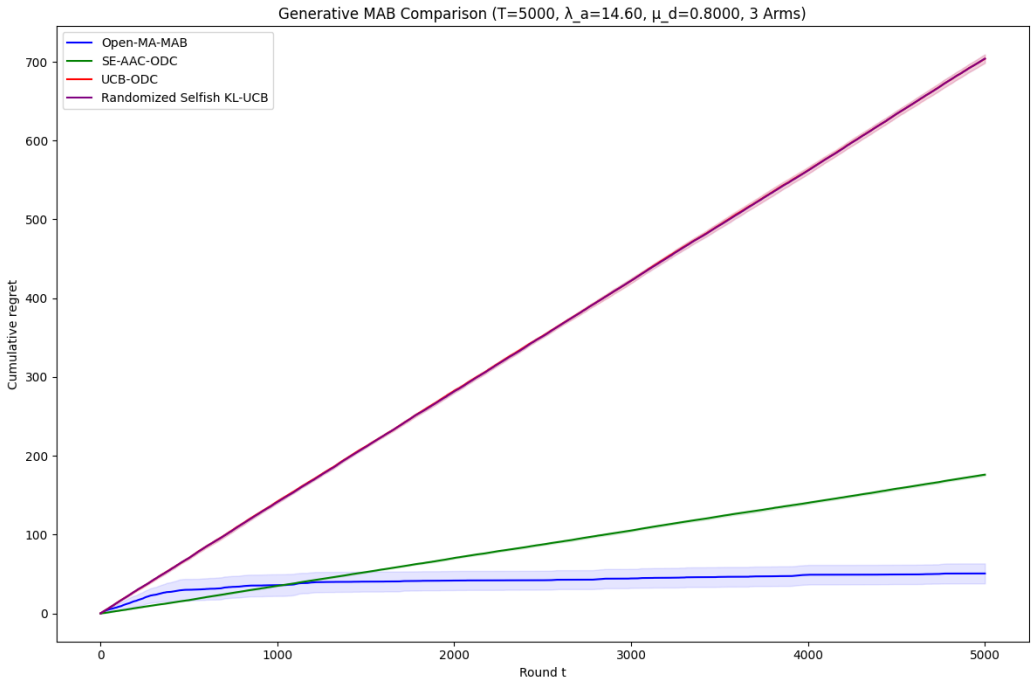

Figure 2: Synthetic Dataset.

Synthetic dataset

time steps	1,000	2,000	5,000
Regret of Open-MA-MAB	134	203	296
Regret of SE-AAC-ODC (Wang et al., 2025)	376	1122	3459
Regret of UCB-ODC (Chen et al., 2023)	177	396	1148
Regret of Randomized KL-UCB (Trinh and Combes, 2021)	175	380	1031
Improvement vs. (Wang et al., 2025)	181%	454%	1069%
Improvement vs. (Chen et al., 2023)	32%	96%	288%
Improvement vs. (Trinh and Combes, 2021)	31%	87%	249%

Real-world dataset

time steps	609
Regret of Open-MA-MAB	694
Regret of SE-AAC-ODC (Wang et al., 2025)	1655
Regret of UCB-ODC (Chen et al., 2023)	2639
Regret of Randomized KL-UCB (Trinh and Combes, 2021)	2738
Improvement vs. (Wang et al., 2025)	138%
Improvement vs. (Chen et al., 2023)	280%
Improvement vs. (Trinh and Combes, 2021)	294%

D.2 Validation of T

The regret upper bound of order $\max(\frac{\log^2 T}{M_0}, \log T, M_0)$ suggests that the regret should be comparable to $\log^2 T$ (small M_0), thereby dominating the $\log T$ term while still being better than T and $\sqrt{T}$. To validate this, we conduct the following numerical experiments; we do not validate the $\log T$ regime for $M_0 = O(\log T)$, as this order also appears in the single-agent case.

The simulation was run for 5,000 rounds with 20 runs, using $M_0 = 5$ initial agents (commonly used in MA-MAB) and a randomly chosen arrival rate of $\lambda = 0.2$.

We generated a plot visualizing the mean regret with a 95% CI band over time. The chart shows a sublinear trend with smaller confidence intervals, indicating effective learning and stability over time. We fit a linear model with respect to $\log^2 T$: coefficient is 8.36, with a Relative MSE of 0.1%.

More precisely, we also report a table showing the regret and its ratios (round to integers) with various functions of t. Both regret$/t$ and regret$/\sqrt{t}$ decrease significantly over time, confirming sublinear behavior. In contrast, regret$/\log t$ increases noticeably over time, indicating that the regret grows faster than $\log t$. Meanwhile, regret$/\log^2 t$ increases more slowly and approaches flat, consistent with the regret order in our theorem.

time	regret	regret$/t$	regret$/\sqrt{t}$	regret$/\log t$	regret$/\log^2 t$
100	46	0	5	10	2
500	195	0	9	31	5
1,000	286	0	9	41	6
2,000	370	0	8	49	6
5,000	469	0	7	55	6

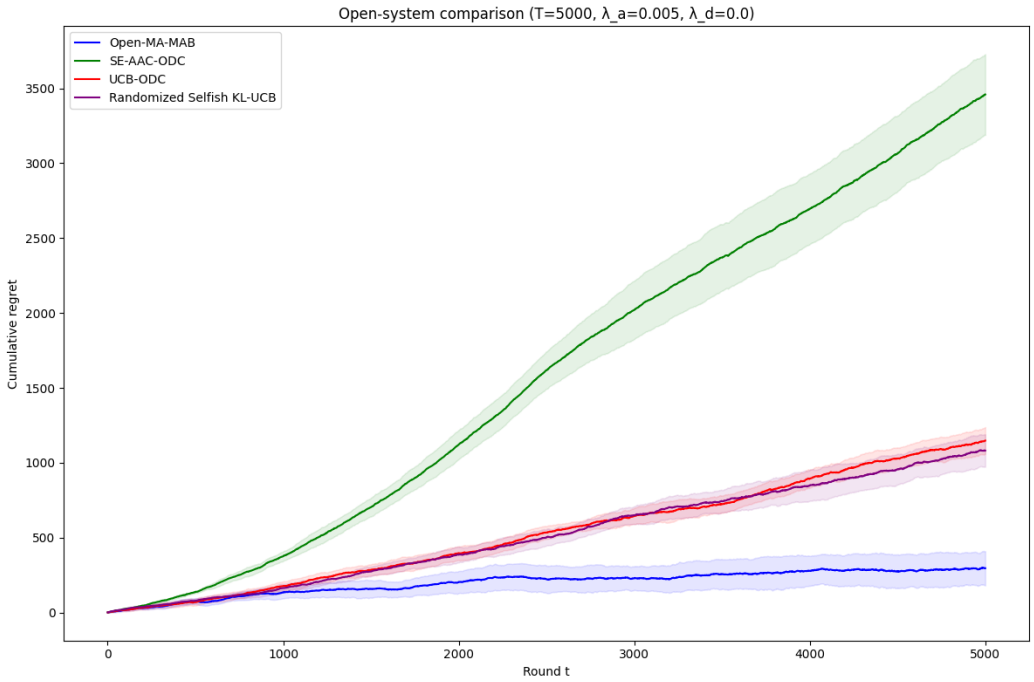

Figure 3: Synthetic Dataset.

D.3 Validation of M_0

We also examine the dependence of exact regret on M_0. As the regret depends on the maximum of $O(M_0)$ and $O(\frac{\log^2 T}{M_0})$, we expect the regret to **first decrease** with M_0 when M_0 is small (so that $\frac{\log^2 T}{M_0}$ is the dominating term), and then **increase** with M_0 when M_0 becomes large enough for the linear M_0 term to dominate. The numerical results below validate this. The simulation evaluated $M_0 = [2, 5, 10, 20]$.

We generated a plot of regret versus M_0, which shows a **monotonic decrease** in regret as M_0 increases initially, followed by a **monotonic increase** for larger M_0, matching the upper bound. We fit a linear model with respect to $\max(M_0, \frac{\log^2 T}{M_0})$, obtaining a coefficient of 1.53 and a relative MSE of 0.1%.

We also report the table summarizing the regret values for different M_0. Again, regret first **decreases** with M_0 (consistent with the dominance of the $\frac{\log^2 T}{M_0}$ term) and then **increases** once M_0 becomes the dominating term.

regret	M_0	$\log^2 T/M_0$
305	2	36
256	5	15
274	10	7
279	20	3

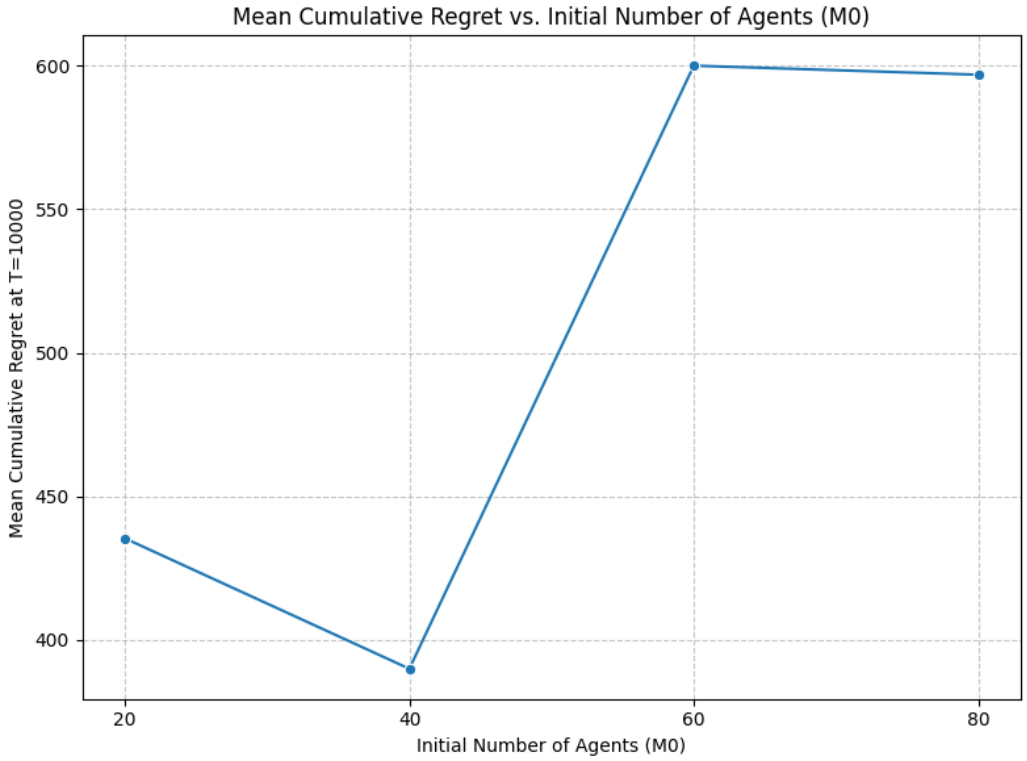

Figure 4: Synthetic Dataset.

D.4 Validation of λ

Additionally, if we examine the small factors in the regret upper bound more precisely, rather than just their orders, the dependence on $\lambda \frac{\log^2 T}{M_0}$ suggests that the regret should be piecewise linear and monotone in λ. Accordingly, we conduct an additional sensitivity study to examine the effect of λ: The simulation evaluated $\lambda = [0.3, 0.8, 1.5, 2, 2.5]$.

We generated a plot showing the dependence of regret on λ, which is piecewise linear and monotone increasing trend as λ increases, consistent with the theoretical bound. We also fit a linear model and obtain a coefficient of 216, with a relative MSE of 0.3%.

We also report a table displaying the regret for different values of λ. As predicted by our theorem, regret increases with λ (since M_0 is small, the $\lambda \log^2 T$ term dominates).

λ	regret
0.3	223
0.8	300
1.5	317
2.0	339
2.5	402

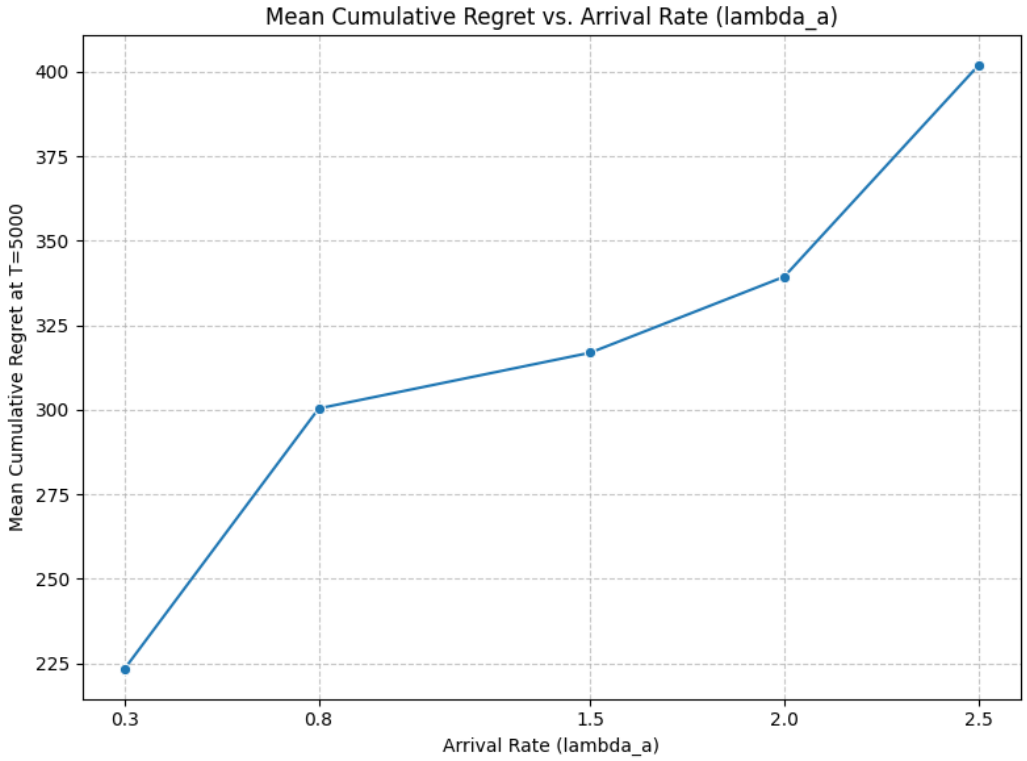

Figure 5: Synthetic Dataset.

The code is available at this link https://github.com/mengfanxu1997/aistats_26.

E NOTATIONS

Notation	Description
K	number of arms
$[K]$	the arm set $\{1, 2, \ldots, K\}$
T	time horizon
t	time index, $1 \leq t \leq T$
M_0	the size of initial agent set
M_t, M_t^A, M_t^D	number of active, arrival, departure agents at time t
m	agent index, $1 \leq m \leq M_t$
$G_t = (V_t, E_t)$	agent graph at time t
$\mathcal{N}_m(t)$	neighbor set of agent m (all agents in fully connected graph)
$r_i^m(t)$	reward for agent m selecting arm i at time t
μ_i	mean reward of arm i (homogeneous across agents)
a_t^m	arm selected by agent m at time t
$n_{m,i}(t)$	number of times agent m has pulled arm i up to t
$N_{m,i}(t)$	total number of times arm i has been pulled by all active agents up to t
λ_A	arrival rate
λ_D	departure rate
$\bar{\mu}_i^m(t)$	local reward estimator of arm i for agent m
$\tilde{\mu}_i^m(t)$	aggregated reward estimator of arm i for agent m
$F(m, i, t)$	uncertainty estimator of $\tilde{\mu}_i^m(t)$
C_1, β	parameters in UCB index that can be tuned
M_t^H, M_t^A	honest and malicious agents at time t
λ_A^H, λ_A^A	arrival rate of honest and malicious agents
λ_D^H, λ_D^A	departure rate of honest and malicious agents
$S_V(t), S_C(t)$	set of validators and commanders
B	the total number of approved blocks by end of the game
b_t^m	an indicator of whether a block is approved at time t for agent m
L	the length of the burn-in period
f_j^t	information of malicious agent j at time t
$\mathcal{F}_t$	information filtration at time t
$h(\cdot)$	arbitrary mapping governing malicious agent decisions
$\mathcal{A}_i(t)$, $\mathcal{B}_i(t)$	set of trusted agents and estimators at agent i
$\mathcal{B}(t)$	the set $\mathcal{B}_i(t)$ that achieves consensus among agents
$\hat{\mu}_i(t)$	instantaneous estimator of arm i based on $\mathcal{B}(t)$ at time t
$\tilde{n}_i(t)$	the number of pulls of arm i based on $\mathcal{B}(t)$ at time t
$\tilde{\mu}_i(t)$	overall estimator of arm i at time t

F PROOF OF THEORETICAL RESULTS

F.1 Proof intuition and novelty

Remark on proof intuition. The key idea behind the upper bounds is to turn openness from a source of extra regret into a source of shared information. In the MA-MAB setting, after the burn-in phase, the initial agents have already accumulated informative reward estimates, and the aggregation mechanism allows the active agents to learn from the *system-level* sample counts rather than only their own local pulls. Thus, newly arriving agents do not restart the learning process from scratch; instead, they inherit the information already accumulated in the system. This is what removes the naive linear-in-T penalty that would appear if every new agent had to relearn independently.

More specifically, the proofs of Theorems 1 and 2 decompose the regret into a burn-in term and a post-burn-in learning term, and then control the latter using UCB concentration with respect to aggregated sample counts. When departures are present, the same proof structure continues to hold after replacing the gross population growth by the *net* growth of the active set, which explains why the dominant term depends on $\lambda_A - \lambda_D$.

For the permissionless-blockchain setting in Theorems 3 and 4, the proof follows a similar decomposition but must

additionally control robustness events. The central high-probability event is that honest agents remain sufficiently dominant and that the on-chain aggregation and verification mechanism filters out harmful information, so that the blockchain estimator behaves as a robustified analogue of the open-system estimator. This preserves the dominant learning term while adding only lower-order overheads due to consensus, verification, and robustness constraints.

Result	**Main contribution**	**Status**	**Reason**
Theorem 1	Upper bound for open MA-MAB without departures	New	First stochastic regret upper bound for MA-MAB with a dynamically growing agent set under the proposed aggregation and information-sharing mechanism.
Theorem 2	Upper bound for open MA-MAB with departures	Carefully adapted	Extends Theorem 1 to the setting with departures, but the analysis must account for the nontrivial effect of departures through the net growth of the active population.
Theorem 3	Upper bound for PB-MA-MAB without departures	Carefully adapted	Uses the permissioned-blockchain proof architecture at a high level, but requires substantial new arguments for open participation, dynamic arrivals, and robust on-chain aggregation.
Theorem 4	Upper bound for PB-MA-MAB with departures	New	Extends Theorem 3 to the setting with departures, with additional work needed to control robustness and regret under a dynamically changing active population.
Theorem 5	PB-MA-MAB lower bound reduction to MA-MAB lower bound	New	Uses a tailored information-set comparison between the two formulations rather than a direct import of a standard theorem.
Theorem 6	Logarithmic lower bound for MA-MAB without departures	New	Adapts Lai–Robbins-type likelihood arguments to the open-system sample complexity induced by the evolving agent set.
Theorem 7 (Corollary)	$\Omega(M_0)$ lower bound for MA-MAB without departures	Carefully adapted	Captures the unavoidable exploration cost of the initial agent population, which does not appear in the same form in fixed-agent settings.
Theorem 8	Stronger lower bound under refined consistency without departures	New	Relies on the newly introduced refined consistency class and shows stronger dependence on both M_0 and T.
Theorem 9 (Corollary)	Stronger lower bound under refined consistency without departures	Carefully adapted	Further sharpens the lower bound in the refined class and helps match the corresponding upper-bound regime more tightly.
Theorem 10 (Corollary)	Logarithmic lower bound with departures	Directly adapted	Departure analogue of Theorem 6 with the same lower-bound logic under net population growth.
Theorem 11 (Corollary)	$\Omega(M_0)$ lower bound with departures	Directly adapted	Departure analogue of Theorem 7; the initial exploration obstruction remains the same.
Theorem 12 (Corollary)	Stronger lower bound under refined consistency with departures	Directly adapted	Departure counterpart of Theorem 8 under the same refined-consistency argument.
Theorem 13 (Corollary)	Stronger lower bound under refined consistency with departures	Directly adapted	Departure counterpart of Theorem 9 with the same proof logic under the modified agent dynamics.

Table 2: Summary of the role of each main theoretical result. Here, "new" means that the result or phenomenon is specific to the open-system formulation introduced in this paper; "carefully adapted" means that the proof uses prior machinery but requires nontrivial analytical adaptation for dynamic agents or robustness; and "directly adapted" means that the result is a relatively clean extension of an earlier theorem in the paper under a modified agent-dynamics regime.

For the lower bounds, the main proof idea is to reinterpret classical likelihood-ratio arguments in an open system. In particular, the relevant "sample size" is no longer only T, but the cumulative opportunity for information acquisition generated by the evolving agent set. This yields the logarithmic lower bounds in Theorems 6 and 10

after adapting the Lai–Robbins argument to the modified sample complexity. The $\Omega(M_0)$ terms in Theorems 7 and 11 arise from the unavoidable initial exploration cost of the starting agents. Finally, Theorems 8–9 and 12–13 strengthen the lower bounds by considering a refined algorithm class; under this stronger notion, the lower bound tightens to the same order as the upper bound, thereby explaining the near-optimality, and in some regimes optimality, of the proposed algorithms.

Remark on technical novelty. The technical novelty is not merely that existing MA-MAB or blockchain arguments are applied in a new setting; rather, the openness of the system changes the mathematical object being analyzed. First, the evolving agent set creates a new notion of effective sample complexity, since the amount of information available by time T is governed by cumulative agent-time rather than by the horizon alone. This change propagates throughout the analysis: it affects the estimator design, the confidence radii, the regret decomposition, and the lower-bound likelihood calculations.

Second, the algorithmic mechanisms are genuinely open-system mechanisms. In the MA-MAB setting, the aggregation rule and the estimator transfer to newly arriving agents prevent repeated relearning. In the permissionless-blockchain setting, this must be combined with a robust on-chain filtering and verification procedure that works with a *changing* set of honest and malicious agents, rather than a fixed synchronized population. This requires weighted updates and high-probability control of the honest-majority event, which are absent from prior permissioned formulations.

Third, the lower-bound analysis introduces a new viewpoint on algorithm classes. Under the standard consistency notion, one obtains logarithmic and M_0-type lower bounds. Under the refined consistency notion introduced here, one obtains stronger lower bounds such as $\Omega(M_0 \log T)$ and $\Omega(\log^2 T)$. This explicit connection between openness, effective sample complexity, and algorithm-class-dependent lower bounds is itself a conceptual contribution of the paper.

Lastly, we summarize the technical novelty of each theorem in Table 2.

Next, we present the complete proofs of the theoretical results from the main body.

The proofs for the theoretical results proceed in two stages: we establish the lemmas first and then the theorems.

F.2 Lemmas

We first present a lemma that characterizes the number of agents at the end of the burn-in period in the absence of departures.

Lemma 1. *Let us denote* $A = \{|M_L - M_0 - 2\lambda_A \cdot \log T| \leq \frac{1}{2}\}$. *Let us further assume that* $\lambda_A \geq 1$. *Then we have that* $P(A) \geq 1 - \frac{2}{T^2}$.

Proof. The proof of this results follow from the Chernoff-Hoeffding's inequality applied to Poisson distribution.

Specifically, note that $M_L - M_0 \sim \text{Pois}(\lambda_A L)$. By the Chernoff bound, for any $\epsilon > 0$, we have

$$\Pr\big(\,\big|M_L - M_0 - 2\lambda_A \log T\big| \geq \epsilon\,\big) \leq 2\exp\left(-\frac{4(\lambda_A \log T + \epsilon)^2}{2(\lambda_A \log T + \epsilon)}\right).$$

Taking $\epsilon = \frac{1}{2}$ and noting the assumption that $\lambda_A \geq 1$, it follows that

$$\Pr(A^c) \leq \frac{2}{T^2}, \qquad \text{and hence} \qquad \Pr(A) \geq 1 - \frac{2}{T^2}.$$

□

We then present a lemma that characterizes the relationship between the arrival and departure processes.

Lemma 2. *Let us assume that the arrival process and departure process of agents follows a Poisson process with arrival rate* λ_A *and departure rate* λ_D, *respectively. Let us further assume that* $(\sqrt{(\lambda_A)} - \sqrt{(\lambda_D)})^2 \geq 1$. *Then we have the following result holds, for any* $t > 0$, $P(|(M_L - M_0) - 2(\lambda_A - \lambda_D) \cdot \log T| \leq \frac{1}{2}) \geq 1 - \frac{2}{T^2}$.

Proof. The proof of this result follows from applying the Chernoff–Hoeffding inequality to the Skellam distribution, which represents the difference between two Poisson random variables, $\text{Pois}(\lambda_A \cdot \frac{\log T}{M_0})$ and $\text{Pois}(\lambda_D \cdot \frac{\log T}{M_0})$.

Let $A_L \sim \text{Pois}(\lambda_A L)$ and $D_L \sim \text{Pois}(\lambda_D L)$ denote the numbers of arrivals and departures over horizon $L > 0$, independently of each other. Then

$$M_L - M_0 \;=\; A_L - D_L \;=\; Z_L,$$

where Z_L follows a Skellam distribution:

$$Z_L \sim \text{Skellam}(\lambda_A L,\ \lambda_D L), \qquad \mathbb{E}[Z_L] = (\lambda_A - \lambda_D)L.$$

The moment generating function (mgf) of Z_L is

$$\mathbb{E}\big[e^{\theta Z_L}\big] = \exp\Big(-(\lambda_A + \lambda_D)L + \lambda_A L\, e^{\theta} + \lambda_D L\, e^{-\theta}\Big), \quad \theta \in \mathbb{R}.$$

Fix $\varepsilon = \frac{1}{2}$. By the Chernoff method, for any $\theta > 0$,

$$\Pr\big(Z_L - (\lambda_A - \lambda_D)L \le -\varepsilon\big) \;\le\; \exp\Big(-(\lambda_A + \lambda_D)L + \lambda_A L\, e^{-\theta} + \lambda_D L\, e^{\theta} + \theta\varepsilon\Big).$$

Optimizing the exponent over $\theta > 0$ yields the choice $\theta^\star = \frac{1}{2}\log\big(\frac{\lambda_A}{\lambda_D}\big)$, for which $\lambda_A e^{-\theta^\star} + \lambda_D e^{\theta^\star} = 2\sqrt{\lambda_A \lambda_D}$, and hence

$$\Pr\big(Z_L \le (\lambda_A - \lambda_D)L - \varepsilon\big) \;\le\; \exp\Big(-L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2 + \theta^\star\varepsilon\Big) \;\le\; \exp\big(-L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2\big),$$

where the last inequality uses $\varepsilon = \frac{1}{2}$ and absorbs the constant $\theta^\star\varepsilon$ into the exponent (the same bound is standard for the one–sided Skellam tail; see also the monotonicity of the bound in ε for fixed L). An analogous argument with $-\theta < 0$ gives the upper tail:

$$\Pr\big(Z_L \ge (\lambda_A - \lambda_D)L + \varepsilon\big) \;\le\; \exp\big(-L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2\big).$$

By a union bound,

$$\Pr\big(\,|Z_L - (\lambda_A - \lambda_D)L| \ge \tfrac{1}{2}\,\big) \;\le\; 2\,\exp\big(-L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2\big).$$

Then we have that

$$\Pr\big(\,|Z_L - 2(\lambda_A - \lambda_D)M_0 L| \ge \tfrac{1}{2}\,\big) \;\le\; 2\,\exp\big(-2M_0 L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2\big).$$

Now take

$$L \;=\; \frac{\log T}{M_0} \quad \text{and} \quad (\sqrt{\lambda_A} - \sqrt{\lambda_D})^2 \;\ge\; 1.$$

Then

$$2\,\exp\big(-2M_0 L(\sqrt{\lambda_A} - \sqrt{\lambda_D})^2\big) \;\le\; 2\,e^{-2\log T} \;\le\; \frac{2}{T^2}.$$

Therefore,

$$\Pr\big(\,\big|(M_L - M_0) - 2(\lambda_A - \lambda_D)L\big| \le \tfrac{1}{2}\,\big) \;\ge\; 1 - \frac{2}{T^2}.$$

Substituting $L = \frac{\log T}{M_0}$ completes the proof.

□

We then present a lemma that characterizes the relationship between the honest and malicious agents when there is no departure process.

Lemma 3. *Let us assume that the arrival process of honest agents and malicious agents follows a Poisson process with arrival rate λ_A^H and λ_A^A, respectively. Let us further assume that $\sqrt{(\lambda_A^H)} - \sqrt{(\lambda_A^A)} \ge 1$ and $M_0^H > 2M_0^A$. Then we have the following result about the number of honest and malicious agents hold, for any $t > 0$,*

$$\begin{aligned}
&P(M_t^H - 2M_t^A > 0)\\
&\ge 1 - e^{-(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2 t} \sum_{k=-\infty}^{-1} (c_0)^{\frac{k}{2}} I_{(k)}(2\sqrt{(\lambda_A^H)(2\lambda_A^A)}t)\\
&\ge 1 - e^{-(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2 t}
\end{aligned}$$

where $I_{(k)}(\cdot)$ denotes the modified Bessel function of the first kind associated with integer k and $c_0 = \frac{\lambda_A^H}{2\lambda_A^A}$.

Proof. Let $M_t^H \sim \text{Pois}(\lambda_A^H t)$ and $M_t^A \sim \text{Pois}(\lambda_A^A t)$ denote the numbers of honest and malicious agents by time t, respectively. Since M_t^H and M_t^A are independent Poisson random variables, their difference

$$Z_t = M_t^H - 2M_t^A$$

follows a Skellam distribution:

$$Z_t \sim \text{Skellam}(\lambda_A^H t, 2\lambda_A^A t).$$

The probability mass function of Z_t is then given by

$$\Pr(Z_t = k) = e^{-(\lambda_A^H + 2\lambda_A^A)t} \left(\frac{\lambda_A^H}{2\lambda_A^A}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H 2\lambda_A^A)}\, t\right), \quad k \in \mathbb{Z},$$

where $I_k(\cdot)$ denotes the modified Bessel function of the first kind.

We are interested in the event $M_t^H - 2M_t^A > 0$, i.e., $\Pr(Z_t > 0)$. From the definition of the Skellam distribution,

$$\Pr(Z_t \le 0) = \sum_{k=-\infty}^{0} e^{-(\lambda_A^H + 2\lambda_A^A)t} \left(\frac{\lambda_A^H}{2\lambda_A^A}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H 2\lambda_A^A)}\, t\right).$$

By applying Chernoff–Hoeffding's inequality for the Skellam distribution, and using the assumption

$$\sqrt{\lambda_A^H} - \sqrt{2\lambda_A^A} \ge 1 \quad \text{and} \quad M_0^H > M_0^A,$$

we obtain the bound

$$\Pr(Z_t > 0) \ge 1 - e^{-(\sqrt{\lambda_A^H} - \sqrt{2\lambda_A^A})^2 t} \sum_{k=-\infty}^{-1} \left(\frac{\lambda_A^H}{2\lambda_A^A}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H 2\lambda_A^A)}\, t\right).$$

Finally, the inequality

$$\Pr(Z_t > 0) \ge 1 - e^{-(\sqrt{\lambda_A^H} - \sqrt{2\lambda_A^A})^2 t}$$

follows by bounding the Bessel summation term related to I_k. This establishes the desired result.

□

Remark. *The above lemma essentially states that if the arrival rate of honest agents exceeds that of malicious agents, and the initial ratio of honest to malicious agents is greater than one, then with high probability the number of honest agents remains larger at every time step. This guarantees that the blockchain can proceed with approved blocks with high probability, which is crucial for establishing the regret bound, analogous to the permissioned case.*

Lastly, we establish a lemma that characterizes the relationship between the honest and malicious agents when the departure process exists.

Lemma 4. *Let us assume that the arrival process of honest agents and malicious agents follows a Poisson process with arrival rate λ_A^H and λ_A^A, respectively, and the corresponding departure rate λ_D^H and λ_D^A, respectively. Let us further assume that $\sqrt{(\lambda_A^H + \lambda_D^A)} - (\sqrt{(2\lambda_D^H + 2\lambda_A^A)} \ge 1$ and $M_0^H > 2M_0^A$. Then we have the following result about the number of honest and malicious agents hold, for any $t > 0$,*

$$\begin{aligned}
&P(M_t^H - 2M_t^A > 0) \\
&\ge 1 - e^{-(\sqrt{(\lambda_A^H + \lambda_D^A)} - \sqrt{(2\lambda_D^H + 2\lambda_A^A)})^2 t} \sum_{k=-\infty}^{-1} (c_0)^{\frac{k}{2}} I_{(k)}(2\sqrt{(\lambda_A^H + \lambda_D^A)(2\lambda_D^A + 2\lambda_A^A)}t) \\
&\ge 1 - e^{-(\sqrt{(\lambda_A^H + \lambda_D^A)} - \sqrt{(2\lambda_D^H + 2\lambda_A^A)})^2 t}
\end{aligned}$$

where $I_{(k)}(\cdot)$ denotes the modified Bessel function of the first kind associated with integer k, and $c_0 = \frac{\lambda_A^H + \lambda_D^A}{2\lambda_D^H + 2\lambda_A^A}$.

Proof. Consider the number of honest and malicious agents by time t, denoted by

$$M_t^H \sim \text{Pois}\big((\lambda_A^H + \lambda_D^A)t\big), \quad M_t^A \sim \text{Pois}\big((\lambda_A^A + \lambda_D^H)t\big).$$

Since these two processes are independent, their difference

$$Z_t = M_t^H - 2M_t^A$$

follows a Skellam distribution:

$$Z_t \sim \text{Skellam}\big((\lambda_A^H + \lambda_D^A)t,\, 2(\lambda_A^A + \lambda_D^H)t\big).$$

The probability mass function of Z_t is

$$\Pr(Z_t = k) = e^{-((\lambda_A^H+\lambda_D^A)+(2\lambda_A^A+2\lambda_D^H))t} \left(\frac{\lambda_A^H + \lambda_D^A}{2\lambda_A^A + 2\lambda_D^H}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H + \lambda_D^A)(2\lambda_A^A + 2\lambda_D^H)}\, t\right),$$

where $I_k(\cdot)$ is the modified Bessel function of the first kind.

We are interested in bounding $\Pr(M_t^H - M_t^A > 0)$, i.e., $\Pr(Z_t > 0)$. By definition,

$$\Pr(Z_t \le 0) = \sum_{k=-\infty}^{0} e^{-((\lambda_A^H+\lambda_D^A)+(2\lambda_A^A+2\lambda_D^H))t} \left(\frac{\lambda_A^H + \lambda_D^A}{2\lambda_A^A + 2\lambda_D^H}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H + \lambda_D^A)(2\lambda_A^A + 2\lambda_D^H)}\, t\right).$$

Applying Chernoff–Hoeffding's inequality to the Skellam distribution, and using the assumption

$$\sqrt{\lambda_A^H + \lambda_D^A} - \sqrt{2\lambda_A^A + 2\lambda_D^H} \ge 1, \quad \text{and} \quad M_0^H > 2M_0^A,$$

we obtain

$$\Pr(Z_t > 0) \ge 1 - e^{-(\sqrt{\lambda_A^H+\lambda_D^A}-\sqrt{2\lambda_A^A+2\lambda_D^H})^2 t} \sum_{k=-\infty}^{-1} \left(\frac{\lambda_A^H + \lambda_D^A}{2\lambda_A^A + 2\lambda_D^H}\right)^{k/2} I_{|k|}\left(2\sqrt{(\lambda_A^H + \lambda_D^A)(2\lambda_A^A + 2\lambda_D^H)}\, t\right).$$

Finally, bounding the summation term yields

$$\Pr(Z_t > 0) \ge 1 - e^{-(\sqrt{\lambda_A^H+\lambda_D^A}-\sqrt{2\lambda_A^A+2\lambda_D^H})^2 t},$$

which completes the proof.

□

F.3 Proof of Regret Upper Bounds in Section 4

F.3.1 Proof of Theorem 1

Proof. Let us assume that the length of the burn-in period is $L = \frac{\log T}{M_0}$. Then after L steps, the number of participants in the game is M_L. Based on Lemma 1, we have that with probability $1 - \frac{1}{T^2}$

$$|(M_L - M_0) - \lambda_A \cdot \tfrac{\log T}{M_0}| \le \tfrac{1}{2},$$

which we denote as event A. This also implies that $M_L \le M_0 + \lambda_A \cdot \frac{\log T}{M_0} + 1$. Notably, after L steps, we also have that there are total $M_0 \cdot \lambda_A \cdot \frac{\log T}{M_0} = \lambda_A \cdot \log T$ number of samples. Based on the Chernoff Hoeffding's inequality, we have that with probability $1 - \epsilon$

$$P(\tilde{\mu}_{m,i}(t) - \mu_i \ge \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}}|A) \le \tfrac{1}{P(A)} \tfrac{1}{M_t t^2},$$
$$P(\mu_i - \tilde{\mu}_{m,i}(t) \ge \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}}|A) \le \tfrac{1}{P(A) M_t t^2}.$$

We consider the number of pulls of arms resulting from the UCB strategies as follows.

We claim that what lead to pulling an sub-optimal arm i are explicit by the decision rule of the method, meaning that the result $a_t^m = i$ holds when any of the following conditions is met:

- Case 1: $\tilde{\mu}_{m,i} - \mu_i > \sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}}$,
- Case 2: $-\tilde{\mu}_{m,i^*} + \mu_{i^*} > \sqrt{\frac{C_1 \log t}{N_{m,i^*}(t-1)}}$,
- Case 3: $\mu_{i^*} - \mu_i < 2\sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}}$.

Then we formally consider the number of pulling arms $N_{m,i}(T)$ starting from $L+1$. For any $l > 1$, we have that based on the above listed conditions

$$\begin{aligned}
N_{m,i}(T) &\leq l + \sum_{t=L+1}^{T} 1_{\{a_t^m = i, N_{m,i}(t) > l\}} \\
&\leq l + \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_i^m - \sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_i, N_{m,i}(t-1) \geq l\}} \\
&\quad + \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_{i^*}^m + \sqrt{\frac{C_1 \log t}{N_{m,i^*}(t-1)}} < \mu_{i^*}, N_{m,i}(t-1) \geq l\}} \\
&\quad + \sum_{t=L+1}^{T} 1_{\{\mu_i + 2\sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l\}}.
\end{aligned}$$

Consequently, the expected value of $n_{m,i}(t)$ conditional on A reads as

$$\begin{aligned}
&E[n_{m,i}(T)|A] \\
&= \tfrac{l}{M_L} + \sum_{t=L+1}^{T} P(\tilde{\mu}_i^m - \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_i, n_{m,i}(t-1) \geq \tfrac{l}{M_t}|A) \\
&\quad + \sum_{t=L+1}^{T} P(\tilde{\mu}_{i^*}^m + \sqrt{\tfrac{C_1 \log t}{N_{m,i^*}(t-1)}} < \mu_{i^*}, n_{m,i}(t-1) \geq \tfrac{l}{M_t}|A) \\
&\quad + \sum_{t=L+1}^{T} P(\mu_i + 2\sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, n_{m,i}(t-1) \geq \tfrac{l}{M_t}|A) \\
&= \tfrac{l}{M_L} + \sum_{t=L+1}^{T} P(Case2, N_{m,i}(t-1) \geq l|A) + \sum_{t=L+1}^{T} P(Case1, N_{m,i}(t-1) \geq l|A) \\
&\quad + \sum_{t=L+1}^{T} P(Case3, N_{m,i}(t-1) \geq l|A)
\end{aligned} \tag{1}$$

where $l = [\frac{4C_1 \log T}{M_L \Delta_i^2}]$.

Subsequently, we obtain that

$$\begin{aligned}
&E[N_{m,i}(T)|A] \\
&= l + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case2, N_{m,i}(t-1) \geq l|A) + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case1, N_{m,i}(t-1) \geq l|A) \\
&\quad + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case3, N_{m,i}(t-1) \geq l|A)
\end{aligned}$$

For the last term in (1), we have

$$\sum_{t=L+1}^{T} P(Case3 : \mu_i + 2\sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l) = 0 \tag{2}$$

since the choice of l satisfies $l \geq [\frac{4C_1 \log T}{\Delta_i^2}]$ with $\Delta_i = \mu_{i^*} - \mu_i$.

Henceforth, we obtain that

$$\sum_{t=1}^{T} \sum_{m=1}^{M_t} P(Case3 : \mu_i + 2\sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l) = 0$$

where M_t denotes the total number of participants at time t.

For the first two terms, we have on event A

$$\begin{aligned}
&\sum_{t=L+1}^{T} P(Case2, N_{m,i}(t-1) \geq l|A) + \sum_{t=1}^{T} P(Case3, N_{m,i}(t-1) \geq l|A) \\
&\leq \sum_{t=L+1}^{T} P(\tilde{\mu}_{m,i} - \mu_i > \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}}|A) + \sum_{t=1}^{T} P(-\tilde{\mu}_{m,i^*} + \mu_{i^*} > \sqrt{\tfrac{C_1 \log t}{N_{m,i^*}(t-1)}}|A) \\
&\leq \sum_{t=1}^{T} (\tfrac{1}{M_t t^2}) + \sum_{t=1}^{T} (\tfrac{1}{M_t t^2})
\end{aligned} \tag{3}$$

where the first inequality holds by the property of the probability measure when removing the event $n_{m,i}(t-1) \geq l$ and the second inequality holds by the Chernoff–Hoeffding inequalities (C-H), which holds by the assumption that $\delta < c$.

Subsequently, we obtain that

$$\begin{aligned}
&\sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case2, N_{m,i}(t-1) \geq l|A) + \sum_{t=1}^{T} P(Case3, N_{m,i}(t-1) \geq l|A) \\
&\leq \sum_{t=1}^{T} \sum_{m=1}^{M_t} (\tfrac{1}{M_t t^2}) + \sum_{t=1}^{T} \sum_{m=1}^{M_t} (\tfrac{1}{M_t t^2}) \\
&\leq \tfrac{\pi^2}{3}
\end{aligned}$$

Consequently, we obtain that

$$\begin{aligned}
&E[N_{m,i}(T)|A] \\
&= l + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case2, N_{m,i}(t-1) \geq l|A) + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case1, N_{m,i}(t-1) \geq l|A) \\
&\quad + \sum_{t=L+1}^{T} \sum_{m=1}^{M_t} P(Case3, N_{m,i}(t-1) \geq l|A) \\
&\leq l + \tfrac{\pi^2}{3} = [\tfrac{4C_1 \log T}{\Delta_i^2}] + \tfrac{\pi^2}{3}
\end{aligned}$$

Subsequently, we derive the following regret upper bound. Precisely, for the proposed regret, we have that for any

constant L,

$$\begin{aligned}
R_T &= max_i \sum_{t=1}^{T}\sum_{m=1}^{M_t} \mu_i^m - \sum_{t=1}^{T}\sum_{m=1}^{M_t} \mu_{a_t^m}^m) \\
&= \sum_{t=1}^{T}\sum_{m=1}^{M_t} \mu_{i^*}^m - \sum_{t=1}^{T}\sum_{m=1}^{M_t} \mu_{a_t^m}^m \\
&\leq M_L \cdot L + \sum_{t=L+1}^{T} (\sum_{m=1}^{M_t} \mu_{i^*}^m - \sum_{m=1}^{M_t} \mu_{a_t^m}^m) \\
&= M_L \cdot L + \sum_{t=L+1}^{T} (M_t \mu_{i^*} - \sum_{m=1}^{M_t} \mu_{a_t^m}^m) \\
&= M_L \cdot L + (\sum_{t=L+1}^{T} M_t \cdot \mu_{i^*} - \sum_{m=1}^{M_t}\sum_{i=1}^{K} n_{m,i}(T)\mu_i^m) \\
&= M_L \cdot L + \sum_{i \neq i*} N_{m,i}(T)(\mu_{i^*}^m - \mu_i^m)
\end{aligned}$$

where the first inequality is by taking the absolute value and the second inequality results from the assumption that $0 < \mu_i^j < 1$ for any arm i and agent j.

Subsequently, we obtain

$$\begin{aligned}
E[R_T|A] &\leq M_L + \sum_i [\tfrac{4C_1 \log T}{\Delta_i^2}] + \tfrac{\pi^2}{3} \\
&\leq (M_0 + \lambda_A \cdot \tfrac{\log T}{M_0}) \cdot \tfrac{\log T}{M_0} + 1 + \sum_i [\tfrac{4C_1 \log T}{\Delta_i^2}] + \tfrac{\pi^2}{3}.
\end{aligned}$$

This implies the final regret upper bound on the regret and thus completes the proof.

□

F.3.2 Proof of Theorem 2

Proof. Here again, let us denote the length of the burn-in period is $L = \frac{\log T}{M_0}$. Based on Lemma 2, we again derive that with probability $1 - \frac{1}{T^2}$

$$|(M_L - M_0) - (\lambda_A - \lambda_D) \cdot \tfrac{\log T}{M_0}| \leq \tfrac{1}{2},$$

Then the remainder of the proof follows from that of Theorem 1, by replacing λ_A with $\lambda_A - \lambda_D$ in the corresponding steps, which represents the net increase in the number of agents per time step. This net increase also reflects the growth of information about the unknown bandit problems, thereby determining the corresponding regret upper bound.

□

F.3.3 Proof of Theorem 3

Proof. Notably, in this setting, the arrival is further categorized into the arrival of honest agents and the arrival of malicious agents. Dealing with the mixture of honest and malicious agents require the following new proof steps.

Based on Lemma 3, we have that the probability of $M_t^H > 2M_t^A$ holding for any $L \leq t \leq T$ is at least

$$\begin{aligned}
&1 - \sum_{t=L}^{T} e^{-(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2 t} \\
&\geq 1 - \tfrac{1}{T(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2}
\end{aligned}$$

Then, related to the new regret that is defined with respect to honest agents, we have the following regret decomposition using the proof framework in (Xu and Klabjan, 2024).

Define b_t to be an indicator variable that specifies if the block at time t has been approved. Similarly, let h_t indicate whether, at time t, the estimators contributed by malicious participants are incorporated into the aggregated estimator. Denote the burn-in horizon by L. Here the choice of L is different from the previous ones, and it depends on the mechanisms.

Based on the proof steps in (Xu and Klabjan, 2024) but with modifications since we have a dynamic set of honest and malicious agents, we have

$$R_T = \max_i \sum_{m\in M_t^H}\sum_{t=1}^{T}\mu_i - \sum_{m\in M_t^H}\sum_{t=1}^{T}\mu_{a_m^t}1_{b_t=1} + \sum_{m\in M_t^H}\sum_{t=1}^{T}c1_{h_t=1}$$

and

$$\begin{aligned} R_T &\le L + c\cdot L + \sum_{t=L+1}^{T}\sum_{m\in M_t^H}(\mu_{i^*} - \mu_{a_m^t}1_{b_t=1}) + \sum_{m\in M_t^H}\sum_{t=L+1}^{T}c1_{h_t=1} \\ &\doteq (c+1)\cdot L + T_1 + T_2 \end{aligned} \tag{4}$$

It is worth noting that based on Lemma 1, we have that with probability $1-\frac{1}{T^2}$

$$|(M_L^H - M_0^H) - \lambda_A^H\cdot\tfrac{\log T}{M_0^H}| \le \tfrac{1}{2},$$

We begin by analyzing the second term T_2. Observe that $h_t = 1$ holds precisely when the set $\{m \mid m \in \mathcal{B}_t \text{ and } m \notin M_t^H\}$ is nonempty. Since the cost is strictly positive, it follows that $\mathcal{B}_t$ cannot be empty.

Taking expectation on T_2 gives

$$\begin{aligned} E[T_2|A] &= \sum_{m\in M_t^H}\sum_{t=L+1}^{T}cE[1_{h_t=1}] \\ &= \sum_{m\in M_t^H}\sum_{t=L+1}^{T}cE[1_{\{m:m\in B_t\cap m\notin M_t^H\}\neq\emptyset}] \end{aligned}$$

With Lemma 2 from (Zhu et al., 2023), we derive that

$$1_{\{m:m\in\mathcal{B}_t\cap m\notin M_t^H\}\neq\emptyset} = 1_{|A_t|<2f}$$

and subsequently,

$$\begin{aligned} E[T_2|A] &= \sum_{m\in M_t^H}\sum_{t=L+1}^{T}cE[1_{h_t=1}] \\ &= \sum_{m\in M_t^H}\sum_{t=L+1}^{T}cE[1_{\{m:m\in\mathcal{B}_t\cap m\notin M_t^H\}\neq\emptyset}] \\ &= \sum_{m\in M_t^H}\sum_{t=L+1}^{T}cE[1_{|A_t|<2f}]. \end{aligned}$$

Note that after the burn-in phase, the definition of A_t guarantees inclusion of all honest validators. Specifically, for each $j \in M_t^H$ we have,

$$m \in A_t \Leftrightarrow k_i n_{m,i}(t) > n_{j,i}(t) \Leftrightarrow m \in M_t^H$$

where $1 < k_i < 2$.

This condition is satisfied at the conclusion of the burn-in phase, which is immediate since all participants are honest. Beyond this phase, the honest validators follow the same decision rule.

$$a_m^t = argmax_i \tilde{\mu}_i^m(t) + F(m, i, t)$$

where $\tilde{\mu}_i^m(t) = \tilde{\mu}_i^b(t)$. In other words, each honest participant uses the validated estimator $\tilde{\mu}_i^b(t)$. Since both $n_{m,i}(t)$ and $n_{j,i}(t)$ are larger than $\frac{L}{K}$, then we have that there exists $k_i = \frac{n_{j,i}(t)K}{L}$, such as $k_i n_{m,i}(t) > n_{j,i}(t)$ for every $m \in M_t^H$.

This implies that

$$A_t > |M_t^H| \geq 2f \tag{5}$$

according to Lemma 3 (with high probability the fraction of honest participants is no less than $\frac{2}{3}$ of M_t).

In other words, we derive that

$$E[1_{|A_t|>2f}] = 1$$

and subsequently, we have

$$\begin{aligned} E[T_2|A] &= \sum_{m \in M_t^H} \sum_{t=L+1}^{T} cE[1_{|A_t|<2f}] \\ &= 0 \end{aligned}$$

Observe that A_t is constructed without access to the number of arm pulls made by other participants. This is achieved through the homomorphic result stated in Theorem 5.2 of (Asharov et al., 2012), within the universal composability framework.

Our next step is to establish a bound on the first term T_1. In particular, we note that

$$\begin{aligned} E[T_1|A] &\leq \sum_{t=L+1}^{T} \sum_{m \in M_t^H} (\mu_{i^*} - \mu_{a_m^t} 1_{b_t=1}) \\ &= (T-L) \cdot |M_t^H| \cdot \mu_{i^*} - \sum_{t=L}^{T} \sum_{m \in M_t^H} E[\mu_{a_m^t} | b_t = 1] P(b_t = 1) \end{aligned}$$

Meanwhile, we note that

$$\begin{aligned} & E[\mu_{a_m^t} | b_t = 1] \\ &= E[\sum_{k=1}^{K} \mu_k \cdot 1_{a_m^t=k} | b_t = 1] \\ &= \sum_{k=1}^{K} E[\mu_k 1_{a_m^t=k} | b_t = 1] \\ &\geq \sum_{k=1}^{K} \mu_k \cdot \tfrac{1}{P(b_t=1)} \cdot (E[1_{a_m^t=k}] - P(b_t = 0)). \end{aligned}$$

Consequently, we obtain that

$$\begin{aligned}
&E[T_1|A] \\
&\leq \sum_{t=L}^{T} |M_t^H|\mu_{i^*} - \sum_{t=L}^{T}(\sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k \cdot \tfrac{1}{P(b_t=1)} \cdot (E[1_{a_m^t=k}] - P(b_t=0)))P(b_t=1) \\
&= \sum_{t=L}^{T} |M_t^H|\mu_{i^*} - \sum_{t=L}^{T}(\sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k (E[1_{a_m^t=k}] - P(b_t=0)) \\
&= \sum_{t=L}^{T} |M_t^H|\mu_{i^*} - \sum_{t=L}^{T} \sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k E[1_{a_m^t=k}] + \sum_{t=L}^{T} \sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k P(b_t=0). \qquad (6)
\end{aligned}$$

As established in Theorem 2 of (Lamport et al., 2019), consensus ($b_t = 1$) is guaranteed so long as the signatures of honest validators cannot be forged. In our model, adversaries are restricted to existential forgery in an adaptive chosen-message attack. By the main result of (Goldwasser et al., 1988), the probability of such an attack succeeding is bounded above by $\frac{1}{Q(l)}$, for any polynomial Q and sufficiently large signature length l.

Formally, the honest participants' signatures are unforgeable with probability at least $1 - \frac{1}{T^2 l^{T-1}}$, ensuring consensus ($b_t = 1$), which is formally expressed as

$$P(b_t = 1) \geq 1 - \tfrac{1}{T^2 l^{T-1}}. \qquad (7)$$

Consequently, we observe the following:

$$\begin{aligned}
(6) &\leq \sum_{t=L}^{T} |M_t^H|\mu_{i^*} - \sum_{t=L}^{T} \sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k E[1_{a_m^t=k}] + \sum_{t=L}^{T} \sum_{m\in M_t^H} \sum_{k=1}^{K} \mu_k (\tfrac{1}{T^2 l^{T-1}}) \\
&\leq \sum_{m\in M_t^H} \sum_{t=L}^{T} (\mu_{i^*} - \sum_{k=1}^{K} \mu_k E[1_{a_m^t=k}]) + \sum_{t=L}^{T} \sum_{m\in M_t^H} K \tfrac{1}{T^2} l^{T-1} \\
&= \sum_{m\in M_H} \sum_{k=1}^{K} \Delta_k E[n_{m,k}(t)] + \sum_{t=L}^{T} \sum_{m\in M_t^H} K \tfrac{1}{T^2} l^{T-1} \\
&\doteq T_{21} + T_{22}
\end{aligned}$$

Meanwhile, we derive that with probability $1 - \frac{1}{T^2}$

$$|(M_t^H - M_0) - \lambda_A^H \cdot t| \leq \tfrac{1}{2},$$

Subsequently, we derive that the second term is upper bounded by the following

$$T_{22} \leq \sum_{t=L}^{T} \sum_{m\in M_t^H} K \tfrac{1}{T^2} l^{T-1} \leq (M_0) + \lambda_A^H + 1$$

And for each honest participant, they are using the estimators based on the validated estimators, as long as the block is approved. Consider the following event, $A = \{\forall 1 \leq t \leq T, b_t = 1\}$. Based on (7) and the Bonferroni's inequality, we obtained that

$$\begin{aligned}
P(A) &= P(\forall 1 \leq t \leq T, b_t = 1) \\
&= 1 - P(\exists 1 \leq t \leq T, b_t = 0) \\
&\geq 1 - \sum_{t=1}^{T} P(b_t = 0) \\
&\geq 1 - \tfrac{1}{T l^{T-1}} - \tfrac{1}{T(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2}.
\end{aligned}$$

On event A, the blockchain always gets approved, and then all the honest participants follow the validated estimators from the validators. By (5) and Lemma 2 in (Zhu et al., 2023), we have that the validated estimator $\tilde{\mu}_i(t)$ can be expressed as

$$\hat{\mu}_i(t) = \sum_{j \in A_t \cap M_t^H} w_{j,i}(t) \bar{\mu}_i^j(t)$$

where the weight $w_{j,i}(t)$ meets the condition

$$\sum_{j \in A_t \cap M_t^H} w_{j,i}(t) = 1,$$

which immediately implies that

$$E[\hat{\mu}_i(t)] = \mu_i.$$

We note that the variance of $\hat{\mu}_i(t)$, $var(\hat{\mu}_i(t))$, satisfies that,

$$\begin{aligned}
var(\hat{\mu}_i(t)) &= var(\sum_{j \in A_t \cap M_t^H} w_{j,i}(t) \bar{\mu}_i^j(t)) \\
&\leq |A_t \cap M_t^H| \sum_{j \in A_t \cap M_t^H} w_{j,i}(t)^2 var(\bar{\mu}_i^j(t))) \\
&\leq |A_t \cap M_t^H| \sum_{j \in A_t \cap M_t^H} w_{j,i}^2(t) \sigma^2 \tfrac{1}{N_{j,i}(t)} \\
&\leq |A_t \cap M_t^H| \sum_{j \in A_t \cap M_t^H} w_{j,i}^2(t) \sigma^2 \tfrac{k_i}{N_{m,i}(t)} \\
&= |M_t^H| \tfrac{k_i}{N_{m,i}(t)} \sum_{j \in M_t^H} w_{j,i}^2(t) \sigma^2 \\
&\leq |M_t^H| \tfrac{k_i \sigma^2}{N_{m,i}(t)}
\end{aligned}$$

where the inequality holds by the Cauchy-Schwarz inequality, the second inequality holds by the definition of sub-Gaussian distributions, the third inequality results from the construction of A_t, and the last inequality is as a result of $(a+b)^2 \geq a^2 + b^2$.

Next, we show by induction that $var(\tilde{\mu}_i(t)) \leq 3|M_t^H| \frac{k_i \sigma^2}{N_{m,i}(t)}$ for $t \geq 3K$.

At time step $3K$, we have that $var(\tilde{\mu}_i(t)) \leq 1$ since $E[\tilde{\mu}_i(t)] = \mu_i \leq 1$. In the meantime,

$$\begin{aligned}
&3|M_t^H| \tfrac{k_i \sigma^2}{N_{m,i}(t-1)} \\
&\geq 3|M_t^H| \tfrac{k_i \sigma^2}{3} \\
&= |M_t^H| k_i \sigma^2 \geq 1
\end{aligned}$$

since we have $k_i \geq 1$ and $\sigma^2 \geq \frac{1}{M_0^H} \geq \frac{1}{M_t^H}$ with probability $1 - \frac{1}{T^2}$.

First, assume that for $t-1$, we have $var(\tilde{\mu}_i(t-1)) \leq 3|M_t^H| \frac{k_i \sigma^2}{N_{m,i}(t-1)}$.

Meanwhile, by the update rule such that $\tilde{\mu}_i(t) = (1 - P_t)\hat{\mu}_i(t) + P_t \tilde{\mu}_i(\tau)$ where $\tau = \max_{s<t}\{b_s = 1\}$.

Note that with probability at least $P(A) = 1 - \frac{1}{Tl^{T-1}}$, $b_s = 1$ for all $s < t$. This implies that on event A, $\tau = t-1$. Therefore, by the cauchy-schwartz inequality, we obtain that

$$\begin{aligned}
var(\tilde{\mu}_i(t)) &\leq 2(1 - P_t)^2 (var(\hat{\mu}_i(t))) + 2P_t^2 var(\tilde{\mu}_i(t-1)) \\
&\leq \tfrac{1}{2}|M_t^H| \tfrac{k_i \sigma^2}{N_{m,i}(t)} + \tfrac{1}{2} 3|M_t^H| \tfrac{k_i \sigma^2}{N_{m,i}(t-1)} \\
&\leq 3|M_t^H| \tfrac{k_i \sigma^2}{N_{m,i}(t)}
\end{aligned}$$

where the last inequality holds by the fact that $N_{m,i}(t-1) \geq N_{m,i}(t) - 1 \geq \frac{2}{3} N_{m,i}(t)$ when $t > 3 \cdot K$.

Subsequently, we have that

$$\begin{aligned}
&P(\tilde{\mu}_i^m(t) - \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}} > \mu_i, N_{m,i}(t-1) \geq l) \\
&\leq \exp\{-\tfrac{(\sqrt{\frac{C_1 \log t}{N_{m,i}(t)}})^2}{2var(\tilde{\mu}_i^m)}\} \\
&\leq \exp\{-\tfrac{(\sqrt{\frac{C_1 \log t}{N_{m,i}(t)}})^2}{6|M_t^H|\frac{k_i \sigma^2}{N_{m,i}(t)}}\} \\
&= \exp\{-\tfrac{C_1 \log t}{6|M_t^H|k_i\sigma^2}\} \leq \tfrac{1}{t^2}
\end{aligned} \tag{8}$$

where the first inequality holds by Chernoff bound, the second inequality is derived by plugging in the above upper bound on $var(\tilde{\mu}_i^m(t))$, and the last inequality results from then choice of C_1 that satisfies $\frac{C_1}{6|M_t^H|k_i\sigma^2} \geq 1$.

Likewise, by symmetry, we have

$$P(\tilde{\mu}_i^m(t) + \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}} < \mu_i, N_{m,i}(t-1) \geq l) \leq \tfrac{1}{t^2}. \tag{9}$$

Meanwhile, we have that

$$\sum_{t=L+1}^{T} P(\mu_i + 2\sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l) = 0 \tag{10}$$

if the choice of l satisfies $l \geq [\frac{4C_1 \log T}{\Delta_i^2}]$ with $\Delta_i = \mu_{i^*} - \mu_i$.

Based on the decision rule, we have the following hold for $N_{m,i}(T)$ with $l \geq [\frac{4C_1 \log T}{\Delta_i^2}]$,

$$\begin{aligned}
N_{m,i}(T) &\leq l + \sum_{t=L+1}^{T} 1_{\{a_t^m = i, N_{m,i}(t) > l\}} \\
&\leq l + \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_i^m - \sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_i, N_{m,i}(t-1) \geq l\}} \\
&\quad + \sum_{t=L+1}^{T} 1_{\{\tilde{\mu}_{i^*}^m + \sqrt{\frac{C_1 \log t}{N_{m,i^*}(t-1)}} < \mu_{i^*}, N_{m,i}(t-1) \geq l\}} \\
&\quad + \sum_{t=L+1}^{T} 1_{\{\mu_i + 2\sqrt{\frac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l\}}.
\end{aligned}$$

By taking the expectation over $n_{m,i}(t)$, we obtain

$$\begin{aligned}
E[N_{m,i}(t)] &\leq l + \sum_{t=L+1}^{T} P(\tilde{\mu}_i^m(t) - \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}} > \mu_i, N_{m,i}(t-1) \geq l) \\
&\quad + \sum_{t=L+1}^{T} P(\tilde{\mu}_i^m(t) + \sqrt{\tfrac{C_1 \log t}{N_{m,i}(t)}} < \mu_i, N_{m,i}(t-1) \geq l) \\
&\quad + \sum_{t=L+1}^{T} P(\mu_i + 2\sqrt{\tfrac{C_1 \log t}{N_{m,i}(t-1)}} > \mu_{i^*}, N_{m,i}(t-1) \geq l) \\
&\leq l + \sum_{t=L+1}^{T} \tfrac{1}{t^2} + \sum_{t=L+1}^{T} \tfrac{1}{t^2} + 0 \\
&\leq l + \tfrac{\pi^2}{3} = [\tfrac{4C_1 \log T}{\Delta_i^2}] + \tfrac{\pi^2}{3}
\end{aligned} \tag{11}$$

where the second inequality holds by using (8), (9), and (10).

Then by the definition of T_{21}, we derive

$$\begin{aligned} E[T_{21}|A] &= \sum_{k=1}^{K} \sum_{m \in M_t^H} \Delta_k E[N_{m,k}(t)] \\ &= \sum_{i \neq i*} N_{m,i}(T)(\mu_{i*}^m - \mu_i^m) \\ &\leq K([\tfrac{4C_1 \log T}{\min_i \Delta_i}] + \tfrac{\pi^2}{3} C \max_i \Delta_i) \end{aligned}$$

where the inequality results from (11).

Consequently, we obtain

$$\begin{aligned} (6) &\leq E[T_{21}|A] + E[T_{22}] \\ &\leq K([\tfrac{4C_1 \log T}{\min_i \Delta_i}] + \tfrac{\pi^2}{3} C \max_i \Delta_i) + (M_0) + \lambda_A^H + 1. \end{aligned} \tag{12}$$

Furthermore, we have probability $1 - \frac{1}{T(\sqrt{(\lambda_A^H)} - \sqrt{(2\lambda_A^A)})^2)} - \frac{1}{Tl^{T-1}}$,

$$\begin{aligned} (4) &\leq (c+1) \cdot L + E[T_1|A] + E[T_2|A] \\ &\leq (c+1) \cdot L + K([\tfrac{4C_1 \log T}{\min_i \Delta_i}] + \tfrac{\pi^2}{3} C \max_i \Delta_i) + (M_0^H) + \lambda_A^H + 1 + 0 \qquad (13) \\ &= O(M_0^H + \log T + \lambda_A^H + \lambda_A^H \cdot \tfrac{\log T^2}{(M_0^H)}) \qquad (14) \end{aligned}$$

which completes the proof.

□

F.3.4 Proof of Theorem 4

Proof. Here again, let us denote the length of the burn-in period is $L = \frac{\log T}{M_0}$. Based on Lemma 2, we again derive that with probability $1 - \frac{1}{T^2}$

$$|(M_L - M_0) - (\lambda_A^H - \lambda_D^H) \cdot \tfrac{\log T}{M_0}| \leq \tfrac{1}{2},$$

In a like manner, we obtain that for any $1 \leq t \leq T$, the following relation holds between the numbers of honest and malicious agents. In particular, by Lemma 4, the probability that $M_t^H > 2M_t^A$ for all $L \leq t \leq T$ is at least

$$\begin{aligned} &1 - \sum_{t=L}^{T} e^{-(\sqrt{(\lambda_A^H + \lambda_D^A)} - \sqrt{(2\lambda_A^A + 2\lambda_D^H)})^2 t} \\ &\geq 1 - \frac{1}{T((\sqrt{(\lambda_A^H + \lambda_D^A)} - \sqrt{(2\lambda_A^A + 2\lambda_D^H)})^2} \end{aligned}$$

The remainder of the proof proceeds analogously to that of Theorem 3, except that λ_A^H is replaced with $\lambda_A^H + \lambda_D^A$ and λ_A^A with $\lambda_A^A + \lambda_D^H$ in the corresponding steps. These substitutions represent the net growth of honest agents per time step. This growth captures both the accumulation of honest information about the unknown bandit problems and the relative proportion of honest versus malicious agents, which together determine the regret upper bound.

This completes the proof of Theorem 4.

□

F.4 Proof of Regret Lower Bounds in Section 5

Proof of Theorem 5. Consider any policy $\pi \in \Pi_{PB}$ where Π_{PB} represents all possible policies in PB-MA-MAB.

Let us denote the σ-algebra induced by the full information I_j^s of agent j at time step s as $\sigma_{PB}^{t,m} = \sigma(\{\{I_j^s\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$, where I_j^s represents the information of all arms available to agent j at time step s. Similarly, let us denote the σ-algebra induced by the limited information $I_j^s(a_s^j)$ corresponding to the pulled arm as $\sigma_B^{t,m} = \sigma(\{\{I_j^s(a_s^j)\}_{j\in\mathcal{N}_m(s)}\}_{s\le t})$, where $I_j^s(a_s^j)$ represents the information of arm a_s^j available to agent j at time step s. In other words, $\sigma_{PB}^{t,m}$ captures the history of the information corresponding to agent m and its (malicious) neighbors' time-dependent actions up to time t, whereas $\sigma_B^{t,m}$ contains the information corresponding to agent m and its neighbors' time-dependent actions up to time t. Hence, we have

$$\sigma_{PB}^{t,m} \subset \sigma_B^{t,m}.$$

Since the policy only requires the information of agents' actions σ^t, and $\sigma_{PB}^t \subset \sigma_B^t$ (with respect to all agents; honest information may be polluted by malicious agents), it follows that $\pi \in \Pi_B$, where Π_B denotes the set of policies in MA-MAB. Consequently, we obtain $\Pi_{PB} \subset \Pi_B$ by the arbitrary choice of π, which implies

$$\min_{\pi\in\Pi_B} R_T^\pi \;\le\; \min_{\pi\in\Pi_{PB}} R_T^\pi,$$

or equivalently,

$$R_T \le R_T^{PB}.$$

Therefore, the lower bound on R_T also serves as a lower bound on R_T^{PB}, which completes the proof.

□

F.4.1 Proof of Theorems 6

Let us assume that the malicious agents perform existential forgery on the signatures of honest agents with an adaptive chosen message attack. Let us assume that the algorithms we consider are consistent in the sense that

$$E[N_i(T)] \le o((\lambda_A \cdot S_T + T \cdot M_0)^a)$$

for any $a > 0$ where $S_T = \frac{T(T+1)}{2}$. Then we have that

$$E[R_T|A] \ge O(\log T) + O(\log(\lambda_A \cdot T + M_0))$$

Proof. First, we adopt the standard conditions from the seminal work of (Lai and Robbins, 1985), while keeping the notation consistent with theirs.

$$0 < I(\theta,\lambda) < \infty \quad \text{whenever } \mu(\lambda) > \mu(\theta). \tag{15}$$

$$\begin{aligned}&\forall\varepsilon > 0,\ \forall\lambda \text{ with } \mu(\lambda) > \mu(\theta),\ \exists\delta > 0 :\ |I(\theta,\lambda) - I(\theta,\lambda')| < \varepsilon\\ &\text{if } \mu(\lambda) \le \mu(\lambda') \le \mu(\lambda) + \delta.\end{aligned} \tag{16}$$

$$\forall\theta \in \Theta,\ \forall\delta > 0,\ \exists\theta' \in \Theta \text{ such that } \mu(\theta) < \mu(\theta') < \mu(\theta) + \delta. \tag{17}$$

To fix ideas, let $j = 1$, $\theta \in \Theta_1$, and $\theta^* = \theta_2$. Then $\mu(\theta^*) > \mu(\theta_1)$ and $\mu(\theta^*) \ge \mu(\theta_i)$ for $3 \le i \le k$. Fix any $0 < \delta < 1$. By (21), (22), and (23), we can choose $\lambda \in \Theta$ such that

$$\mu(\theta^*) < \mu(\lambda), \quad \text{and} \quad I(\theta_1,\lambda) - I(\theta_1,\theta^*) < \delta I(\theta_1,\theta^*).$$

Define the new parameter vector $\gamma = (\lambda, \theta_2, \ldots, \theta_k)$. Then $\gamma \in \Theta_1^*$, so by the assumption that

$$E[N_i(T)] \leq o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

we derive

$$\sum_{i \neq 1} E_\gamma N_i(T) = o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

for any $a > 0$ where $S_T = \frac{T(T+1)}{2}$.

Also, notably, we have with probability $1 - \frac{1}{T^2}$

$$|(M_t - M_0) - (\lambda_A) \cdot t| \leq \tfrac{1}{2},$$

Therefore, for some $0 < \alpha < \delta$,

$$\begin{aligned}
&(M_0 + \lambda_A \cdot \tfrac{T+1}{2})T - N_1(T) \\
&= \sum_{i \neq 1} N_i(T) \\
&= o(((\lambda_A \cdot S_T + T \cdot M_0^H))^a).
\end{aligned}$$

This implies

$$\begin{aligned}
&P_\gamma\{N_1(T) < (1-\alpha)(M_0 + \lambda_A \cdot T)T\} \\
&\leq \tfrac{E_\gamma[(M_0 + \lambda_A \cdot T)T - N_1(T)]}{\alpha(M_0 + \lambda_A \cdot T)T} \\
&= o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}).
\end{aligned}$$

Let $Y_1, Y_2, \ldots$ be successive observations from Π_1, and define

$$L_m = \sum_{i=1}^{m} \log \tfrac{f(Y_i;\theta_1)}{f(Y_i;\lambda)}.$$

Then,

$$P_\gamma(C_n) = o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}),$$

where

$$C_n = \{N_1(T) < (1-\delta)(\log(\lambda_A \cdot S_T + T \cdot M_0^H))/I(\theta_1, \lambda),\ L_{N_1(T)} \leq (1-\alpha)\log T\}.$$

Note that

$$\begin{aligned}
&P_\gamma\{N_1(T) = T_1, \ldots, N_K(T) = T_K, L_{N_1(T)} \leq (1-\alpha)\log(\lambda_A \cdot S_T + T \cdot M_0)\} \\
&= \int \prod_{i=1}^{T_1} f(y_i;\lambda) \prod_{i=2}^{K} \prod_{j=1}^{T_i} f(y_j;\theta_i) dy \\
&\leq \exp(-(1-\alpha)\log(\lambda_A \cdot S_T + T \cdot M_0))\, P_\theta\{N_1(T) = T_1, \ldots, N_K(T) = T_K, L_{N_1(T)} \\
&\qquad \leq (1-\alpha)\log(\lambda_A \cdot S_T + T \cdot M_0)\}.
\end{aligned}$$

Summing over disjoint events gives

$$P_\gamma(C_n) \leq (\lambda_A \cdot S_T + T \cdot M_0)^{a-1} P_\theta(C_n).$$

By the strong law of large numbers, $L_m/m \to I(\theta_1, \lambda)$ a.s. under P_θ. Since $I(\theta_1, \lambda) > 0$, it follows that

$$\begin{aligned}
&P_\theta\{L_m > (1-\alpha)\log(\lambda_A \cdot S_T + T \cdot M_0) \\
&\qquad \text{for some } m < (1-\delta)(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda)\} \to 0.
\end{aligned}$$

Combining these bounds shows that

$$\lim_{T\to\infty} P_\theta\{N_1(T) < (1-\delta)(\log(\lambda_A \cdot S_T + T \cdot M_0^H))/I(\theta_1,\theta^*)\} = 0,$$

which yields the conclusion in the regret statement.

□

F.4.2 Proof of Theorem 7

Let us assume that the malicious agents perform existential forgery on the signatures of honest agents with an adaptive chosen message attack. Let us assume that the algorithms we consider are consistent in the sense that

$$E[n_i(T)] \leq (\lambda_A \cdot S_T + T \cdot M_0)^a$$

for any $a > 0$ where $S_T = \frac{T(T+1)}{2}$. Then we have that

$$E[R_T|A] \geq O(M_0)$$

Proof. Notably, before any arms are pulled, the agents have no prior knowledge of the bandit problems.

If we assume that the initial set of agents must explore once, then the lower bound holds. This completes the first part of the proof.

If a linear proportion of the initial set of agents pull sub-optimal arms at least once, then the lower bound also holds, which also completes the proof.

Otherwise, without this assumption, we consider the following. The equivalent statement of this assumption is that: only sublinear subset of agents $o(M_0)$ pull sub-optimal arms at least once, i.e. linear set of agents (among all agents) $\sum_{t=1}^{T} M_t - o(M_0)$, only pull the same optimal arms (unique), namely arm i^*, from the beginning until the end, which means that they do not explore at all.

If we switch the optimal arm and sub-optimal arm (where the sub-optimality gap meets the condition in (Bubeck et al., 2013)), then the algorithm would incur linear regret (the two setting makes no difference for the initial set of agents at the beginning), as they have no prior information. In this case, there would exist instances in which the algorithm incurs linear regret, contradicting the assumption of consistency.

Consequently, in any case, the regret is at least $O(M_0)$ as the result of exploration of the initial set of agents, which completes the proof.

□

F.4.3 Proof of Theorem 8

Proof. First, we adopt the standard conditions from the seminal work of (Lai and Robbins, 1985), while keeping the notation consistent with theirs.

$$0 < I(\theta,\lambda) < \infty \quad \text{whenever } \mu(\lambda) > \mu(\theta). \tag{18}$$

$$\begin{aligned}&\forall \varepsilon > 0,\ \forall \lambda \text{ with } \mu(\lambda) > \mu(\theta),\ \exists \delta > 0 :\ |I(\theta,\lambda) - I(\theta,\lambda')| < \varepsilon \\ &\text{if } \mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta.\end{aligned} \tag{19}$$

$$\forall \theta \in \Theta,\ \forall \delta > 0,\ \exists \theta' \in \Theta \text{ such that } \mu(\theta) < \mu(\theta') < \mu(\theta) + \delta. \tag{20}$$

To fix ideas, let $j = 1$, $\theta \in \Theta_1$, and $\theta^* = \theta_2$. Then $\mu(\theta^*) > \mu(\theta_1)$ and $\mu(\theta^*) \geq \mu(\theta_i)$ for $3 \leq i \leq k$. Fix any $0 < \delta < 1$. By (21), (22), and (23), we can choose $\lambda \in \Theta$ such that

$$\mu(\theta^*) < \mu(\lambda), \quad \text{and} \quad I(\theta_1,\lambda) - I(\theta_1,\theta^*) < \delta I(\theta_1,\theta^*).$$

Define the new parameter vector $\gamma = (\lambda, \theta_2, \ldots, \theta_k)$. Then $\gamma \in \Theta_1^*$, so by the assumption that

$$E[N_i(T)] \leq o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

we derive

$$\sum_{i \neq 1} E_\gamma N_i(T) = o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

for any $a > \frac{M_0-2}{M_0-1}$ where $S_T = \frac{T(T+1)}{2}$.

Also, notably, we have with probability $1 - \frac{1}{T^2}$

$$|(M_t - M_0) - (\lambda_A) \cdot t| \leq \tfrac{1}{2},$$

Therefore, for some $0 < \alpha < \delta$,

$$\begin{aligned}
&(M_0 + \lambda_A \cdot \tfrac{T+1}{2})T - N_1(T) \\
&= \sum_{i \neq 1} N_i(T) \\
&= o(((\lambda_A \cdot S_T + T \cdot M_0))^a).
\end{aligned}$$

This implies

$$\begin{aligned}
&P_\gamma\{N_1(T) < (1-\delta)M_0^H(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda)\} \\
&\leq \tfrac{E_\gamma[(M_0+\lambda_A \cdot T)T - N_1(T)]}{(\alpha)(M_0+\lambda_A \cdot T)T - O(M_0 \log T)} \\
&= o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}).
\end{aligned}$$

Let $Y_1, Y_2, \ldots$ be successive observations from Π_1, and define

$$L_m = \sum_{i=1}^{m} \log \tfrac{f(Y_i;\theta_1)}{f(Y_i;\lambda)}.$$

Then,

$$P_\gamma(C_n) = o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}),$$

where

$$\begin{aligned}
C_n = \{&N_1(T) < (1-\delta)M_0(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda), \\
&L_{N_1(T)} \leq (1-\alpha)M_0 \log(\lambda_A \cdot S_T + T \cdot M_0)\}.
\end{aligned}$$

Note that

$$\begin{aligned}
&P_\gamma\{N_1(T) = T_1, \ldots, N_K(T) = T_K, L_{N_1(T)} \leq (1-\alpha)M_0 \log(\lambda_A \cdot S_T + T \cdot M_0)\} \\
&= \int \prod_{i=1}^{T_1} f(y_i;\lambda) \prod_{i=2}^{K} \prod_{j=1}^{T_i} f(y_j;\theta_i) dy \\
&\leq \exp(-(1-\alpha)M_0 \log(\lambda_A \cdot S_T + T \cdot M_0))\, P_\theta\{N_1(T) = T_1, \ldots, N_K(T) = T_K, L_{N_1(T)} \\
&\qquad \leq (1-\alpha)M_0 \log(\lambda_A \cdot S_T + T \cdot M_0)\}.
\end{aligned}$$

Summing over disjoint events gives when $a > \frac{M_0-2}{M_0-1}$

$$P_\gamma(C_n) \leq (\lambda_A \cdot S_T + T \cdot M_0)^{a-1} P_\theta(C_n).$$

By the strong law of large numbers, $L_m/m \to I(\theta_1, \lambda)$ a.s. under P_θ. Since $I(\theta_1, \lambda) > 0$, it follows that

$$P_\theta\{L_m > (1-\alpha)M_0 \log(\lambda_A \cdot S_T + T \cdot M_0) \\ \text{for some } m < (1-\delta)M_0(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda)\} \to 0.$$

Combining these bounds shows that

$$\lim_{T\to\infty} P_\theta\{N_1(T) < (1-\delta)M_0(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \theta^*)\} = 0,$$
$$\lim_{T\to\infty} P_\theta\{N_1(T) < (1-\delta)(M_0 + \log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \theta^*)\}$$

which yields the conclusion in the regret statement, by noting that

$$M_0 \cdot \log(\lambda_A \cdot S_T + T \cdot M_0) \\ \geq M_0 + \log(\lambda_A \cdot S_T + T \cdot M_0)$$

holds for $M_0 > 1 + \frac{1}{\log T}$.

In other words, the regret lower bound matches the upper bound asymptotically conditional on (1) the new stricter family of consistent algorithms and (2) $M_0 > 1 + \frac{1}{\log T}$. This completes the proof step, and thus concludes the proof of Theorem 8.

□

F.4.4 Proof of Theorem 9

Proof. First, we adopt the standard conditions from the seminal work of (Lai and Robbins, 1985), while keeping the notation consistent with theirs.

$$0 < I(\theta, \lambda) < \infty \quad \text{whenever } \mu(\lambda) > \mu(\theta). \tag{21}$$

$$\forall \varepsilon > 0,\ \forall \lambda \text{ with } \mu(\lambda) > \mu(\theta),\ \exists \delta > 0 :\ |I(\theta, \lambda) - I(\theta, \lambda')| < \varepsilon \\ \text{if } \mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta. \tag{22}$$

$$\forall \theta \in \Theta,\ \forall \delta > 0,\ \exists \theta' \in \Theta \text{ such that } \mu(\theta) < \mu(\theta') < \mu(\theta) + \delta. \tag{23}$$

To fix ideas, let $j = 1$, $\theta \in \Theta_1$, and $\theta^* = \theta_2$. Then $\mu(\theta^*) > \mu(\theta_1)$ and $\mu(\theta^*) \geq \mu(\theta_i)$ for $3 \leq i \leq k$. Fix any $0 < \delta < 1$. By (21), (22), and (23), we can choose $\lambda \in \Theta$ such that

$$\mu(\theta^*) < \mu(\lambda), \quad \text{and} \quad I(\theta_1, \lambda) - I(\theta_1, \theta^*) < \delta I(\theta_1, \theta^*).$$

Define the new parameter vector $\gamma = (\lambda, \theta_2, \ldots, \theta_k)$. Then $\gamma \in \Theta_1^*$, so by the assumption that

$$E[N_i(T)] \leq o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

we derive

$$\sum_{i \neq 1} E_\gamma N_i(T) = o((\lambda_A \cdot S_T + T \cdot M_0)^a),$$

for any $a > \frac{M_0 - 2}{M_0 - 1}$ where $S_T = \frac{T(T+1)}{2}$.

Also, notably, we have with probability $1 - \frac{1}{T^2}$

$$|(M_t - M_0) - (\lambda_A) \cdot t| \leq \tfrac{1}{2},$$

Therefore, for some $0 < \alpha < \delta$,

$$\begin{aligned}&(M_0 + \lambda_A \cdot \tfrac{T+1}{2})T - N_1(T)\\ &= \sum_{i\neq 1} N_i(T)\\ &= o(((\lambda_A \cdot S_T + T \cdot M_0))^a).\end{aligned}$$

This implies

$$\begin{aligned}&P_\gamma\{N_1(T) < (1-\delta)\log T(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda)\}\\ &\leq \tfrac{E_\gamma[(M_0+\lambda_A \cdot T)T - N_1(T)]}{(\alpha)(M_0+\lambda_A \cdot T)T - O(\log T^2)}\\ &= o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}).\end{aligned}$$

Let $Y_1, Y_2, \ldots$ be successive observations from Π_1, and define

$$L_m = \sum_{i=1}^{m} \log \tfrac{f(Y_i;\theta_1)}{f(Y_i;\lambda)}.$$

Then,

$$P_\gamma(C_n) = o((\lambda_A \cdot S_T + T \cdot M_0)^{a-1}),$$

where

$$\begin{aligned}C_n = \{&N_1(T) < (1-\delta)\log T(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda),\\ &L_{N_1(T)} \leq (1-\alpha)\log T \log(\lambda_A \cdot S_T + T \cdot M_0)\}.\end{aligned}$$

Note that

$$\begin{aligned}&P_\gamma\{N_1(T) = T_1, \ldots, N_K(T) = T_K,\\ &\qquad L_{N_1(T)} \leq (1-\alpha)\log T \log(\lambda_A \cdot S_T + T \cdot M_0)\}\\ &= \int \prod_{i=1}^{T_1} f(y_i;\lambda) \prod_{i=2}^{K}\prod_{j=1}^{T_i} f(y_j;\theta_i)dy\\ &\leq \exp(-(1-\alpha)\log T \log(\lambda_A \cdot S_T + T \cdot M_0))\, P_\theta\{N_1(T) = T_1, \ldots, N_K(T) = T_K, L_{N_1(T)}\\ &\qquad \leq (1-\alpha)\log T \log(\lambda_A \cdot S_T + T \cdot M_0)\}.\end{aligned}$$

Summing over disjoint events gives when $a > \frac{\log T - 2}{\log T - 1}$

$$P_\gamma(C_n) \leq (\lambda_A \cdot S_T + T \cdot M_0)^{a-1} P_\theta(C_n).$$

By the strong law of large numbers, $L_m/m \to I(\theta_1, \lambda)$ a.s. under P_θ. Since $I(\theta_1, \lambda) > 0$, it follows that

$$\begin{aligned}P_\theta\{&L_m > (1-\alpha)\log T \log(\lambda_A \cdot S_T + T \cdot M_0)\\ &\text{for some } m < (1-\delta)\log T(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \lambda)\} \to 0.\end{aligned}$$

Combining these bounds shows that

$$\lim_{T\to\infty} P_\theta\{N_1(T) < (1-\delta)\log T(\log(\lambda_A \cdot S_T + T \cdot M_0))/I(\theta_1, \theta^*)\} = 0$$

which yields the conclusion in the regret statement, and thus completes the proof steps.

In other words, the regret lower bound matches the upper bound asymptotically conditional on the new stricter family of consistent algorithms. This completes the proof step, and thus concludes the proof of Theorem 9.

□

F.4.5 Proof of Theorems 10

Proof. Again, we leverage the aforementioned fact that with probability $1 - \frac{1}{T^2}$

$$|(M_t - M_0) - \lambda_A \cdot t| \leq \frac{1}{2},$$

The remainder of the proof follows the same argument as Theorem 6, with λ_A replaced by $\lambda_A - \lambda_D$ in the corresponding steps. The term

$$\lambda_A - \lambda_D$$

represents the net increase in the number of agents per time step, which governs the accumulation of information about the unknown bandit problems and thus determines the corresponding regret upper bound.

□

F.4.6 Proof of Theorem 11

Proof. Similar to the proof of Theorem 7, before any arms are pulled, the initial set of agents still lack prior knowledge of the bandit problems, irrespective of the presence of a departure process.

If we assume that the initial set of agents must explore once, then the lower bound holds. This completes the first part of the proof.

If a linear proportion of the initial set of agents pull sub-optimal arms at least once, then the lower bound also holds, which also completes the proof.

Otherwise, without this assumption, we consider the following. The equivalent statement of this assumption is that: only sublinear subset of agents $o(M_0)$ pull sub-optimal arms at least once, i.e. linear set of agents (among all agents) $\sum_{t=1}^{T} M_t - o(M_0)$, only pull the same optimal arms (unique), namely arm i^*, from the beginning until the end, which means that they do not explore at all.

If we switch the optimal arm and sub-optimal arm (where the sub-optimality gap meets the condition in (Bubeck et al., 2013)), then the algorithm would incur linear regret (the two setting makes no difference for the initial set of agents at the beginning), as they have no prior information. In this case, there would exist instances in which the algorithm incurs linear regret, contradicting the assumption of consistency.

Consequently, in any case, the regret is at least $O(M_0)$ as the result of exploration of the initial set of agents, which completes the proof.

□

F.4.7 Proof of Theorem 12

Proof. We again use the fact that, with probability $1 - \frac{1}{T^2}$,

$$|(M_t - M_0) - \lambda_A \cdot t| \leq \frac{1}{2},$$

The remainder of the proof parallels the argument of Theorem 8, with λ_A replaced by $\lambda_A - \lambda_D$ in the corresponding steps. Here, $\lambda_A - \lambda_D$ denotes the net growth rate of agents per time step, which governs the accumulation of information about the bandit problems and, in turn, determines the regret upper bound.

□

F.4.8 Proof of Theorem 13

Proof. Based on the result in Lemma 1, we have that with probability $1 - \frac{1}{T^2}$

$$|(M_t - M_0) - \lambda_A \cdot t| \leq \frac{1}{2},$$

The rest of the proof follows the argument of Theorem 9, except that λ_A is replaced by $\lambda_A - \lambda_D$. The quantity $\lambda_A - \lambda_D$ captures the net change in the number of agents per time step, which dictates the rate of information accumulation about the bandit problems and thereby determines the regret upper bound.

□