Variance-Aware Bandit Framework for Dynamic Probabilistic Maximum Coverage Problem with Triggered or Self-Reliant Arms

Xiangxiang Dai, *Student Member, IEEE*, Xutong Liu, *Member, IEEE*, Jinhang Zuo, *Member, IEEE*, Hong Xie, *Member, IEEE*, Carlee Joe-Wong, *Senior Member, IEEE*, John C.S. Lui, *Fellow, IEEE, ACM*

***Abstract*—The Probabilistic Maximum Coverage (PMC) problem plays a pivotal role in modeling various network applications, such as mobile crowdsensing, which involves selecting nodes within a graph that probabilistically cover other nodes. Our study focuses on PMC within the framework of online learning, termed the PMC bandit, where the network parameters are initially unknown. In this scenario, the decision-maker is tasked with learning these parameters to maximize the cumulative rewards from covered nodes. Despite prior research on the PMC bandit, we propose a novel variant, dynamic PMC-G bandit, which extends the semi-bandit feedback model to represent applications more accurately. To tackle the complexities of the time-varying combinatorial arm set rather than traditional static, we enhance the Combinatorial Upper Confidence Bound (CUCB) algorithms by developing two innovative variance-aware strategies: the Variance-Adaptive Combinatorial Upper Confidence Bound (VACUCB) for probabilistically triggered arms, and the Action-Based Combinatorial Upper Confidence Bound (ABCUCB) for self-reliant arms, i.e., independent arms with probabilistically triggered outcomes. Based on variance-aware properties, our contributions notably reduce the dependence on the number of nodes K selected per round, demonstrating that: (i) VACUCB effectively minimizes the regret associated with the triggered arms, enhancing the CUCB by a factor of $\tilde{O}(K)$; (ii) ABCUCB further diminishes the dependence on K in the leading term. Empirical results from synthetic and real-world datasets confirm that our proposed algorithms outperform current benchmarks in three network applications.**

***Index Terms*—Probabilistic maximum coverage, online learning, combinatorial bandit, network application optimization**

I. Introduction

The probabilistic maximum coverage (PMC) problem [1] is an effective model widely used in various network applications, including network content delivery [2], mobile crowdsensing [3], and channel allocation [4]. This problem is modeled using a bipartite graph $G = (L, V, E)$, where L represents the nodes to be selected, V the nodes to be covered, and E the edges connecting these nodes. Each edge $(u, v) \in E$ carries an associated probability $p(u, v)$, indicating the likelihood that node $u \in L$ can cover node $v \in V$. Additionally, each node $v \in V$ is assigned a weight $w(v)$, representing the reward for covering that node. The objective for the decision-maker is to select no more than K nodes from L to maximize the sum of the weights of the covered nodes in V.

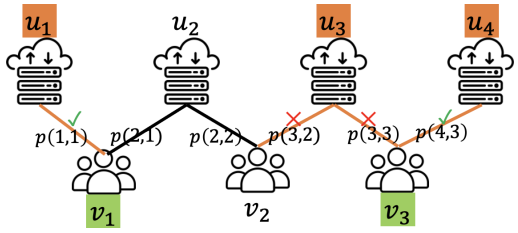


Fig. 1: An example of PMC for content delivery: the decision-maker chooses (orange) servers $\{u_1, u_3, u_4\}$ which cover users $\{v_1, v_3\}$ via successful (check mark) edges $\{(1, 1), (4, 3)\}$.

In a content delivery network (CDN), contents such as images and videos are stored across various mirror servers to ensure rapid accessibility for end users through the nearest server [2]. The challenge of strategically selecting a subset of servers (K in number) to enhance user experience can be effectively represented by the Probabilistic Maximum Coverage (PMC) model (see Fig. 1). Here, L denotes the set of candidate mirror servers responsible for content distribution, while V encompasses the users accessing the content. Each edge $(u, v) \in E$ in this model signifies the likelihood ($p(u, v)$) that server u can deliver content on time to user v (thus, u covers v), and $w(v)$ represents the probability that user v will ultimately access the content. The primary objective of PMC in this scenario is to optimize user satisfaction by maximizing the total number of users who successfully access the content.

In the PMC framework, accurately setting parameters such as edge probabilities is crucial for making optimal decisions. Previous studies have assumed these parameters are known beforehand [5]. However, in real-world network applications, these parameters are typically unknown and subject to change dynamically. For example, in network content delivery scenarios [6], user demands and preferences for content, denoted

Manuscript received May 22, 2024; revised September 22, 2024; accepted December 04, 2024. The work of Xutong Liu was supported in part by a fellowship award from the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK PDFS2324-4S04). The work of Hong Xie was supported in part by the National Natural Science Foundation of China under Grant 62476261. The work of Carlee Joe-Wong was supported in part by NSF grant CNS-2103024. The work of John C.S. Lui was supported in part by the RGC GRF-14202923. (*Corresponding author: Xutong Liu.*)

Xiangxiang Dai and John C.S. Lui are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. (Email: {xxdai23, cslui}@cse.cuhk.edu.hk.)

Xutong Liu and Carlee Joe-Wong are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. (Email: {xutongl, cjoewong}@andrew.cmu.edu.)

Jinhang Zuo is with the Department of Computer Science, City University of Hong Kong, Hong Kong. (Email: jinhang.zuo@cityu.edu.hk.)

Hong Xie is with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China. (Email: xiehong2018@foxmail.com.)

by $w(v)$, can vary unpredictably. Similarly, the delivery probability $p(u, v)$, which reflects the service quality of mirror servers, is influenced by factors like varying distances and potential network congestion and thus remains uncertain. These parameters must be dynamically estimated by network operators as they are not known in advance. Additionally, decision-making in network optimization often involves combinatorial choices, where multiple servers, participants, or items are selected simultaneously. For example, a mobile crowdsensing organizer may need to select the top participants.

A. Dynamic PMC with General Feedback and Targeted Arms

To address the challenge of unknown parameters, the PMC problem can be explored within an online learning framework, termed the PMC bandit [7]. In this model, each edge (u, v) (or node v) is treated as an arm, with its probability $p(u, v)$ (or $w(v)$) unknown and to be learned over T consecutive decision rounds. During each round t, the decision-maker, functioning as the learning agent, is required to choose a set of arms, referred to as *actions*, and observes the results of these actions as feedback. This feedback is then used to estimate the unknown probabilities and refine future decision-making strategies. This type of feedback is categorized as *semi-bandit* feedback [1], [7], [8]. The primary objective of the agent is to maximize the expected rewards over T rounds or, equivalently, to minimize the expected *regret*. Regret is defined as the difference in expected rewards between consistently selecting the optimal actions and following the agent's actual policy.

For PMC bandit, a good learning algorithm must carefully handle the exploration-exploitation trade-off: whether the agent should explore arms in search of a better action, or should the agent stick to the best action observed so far to gain rewards. To deal with this trade-off, combinatorial upper confidence bound algorithms (CUCB) are proposed [7]. Specifically, CUCB uses the empirical mean as the unbiased estimator for each arm and constructs a Chernoff-type confidence interval. Such an interval serves as the exploration bonus to handle the parameter uncertainty and helps to achieve sublinear regret bounds [7].

Although the PMC bandit has been extensively studied, the bandit model and its CUCB algorithm have notable drawbacks that can be substantially improved. Firstly, regarding feedback handling, the traditional semi-bandit feedback, which focuses solely on direct outcomes from deterministic arm selections, inadequately addresses feedback contingent on stochastically varying outcomes. For example, in CDN scenarios, this feedback model fails to accurately represent the uncertain user consumption probability $w(v)$, which is only observable following successful content delivery. Secondly, most existing CUCB studies assume static combinations of arms, i.e., fixed action sets [1], [9], [10], an assumption misaligned with the dynamic nature of real-world networks where action sets evolve continually. To accommodate these dynamic network applications, it is essential to further develop the PMC bandit framework with a general feedback model that supports time-dependent combinations of arms. This is exemplified in mobile crowdsensing (see Fig. 2), where the available participants fluctuate due to factors like changing engagement levels.

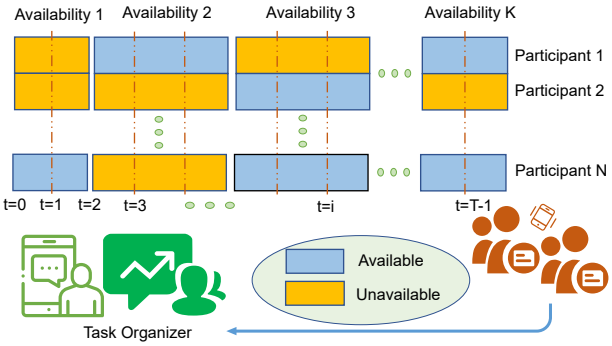


Fig. 2: Illustrating the dynamic action set on participant availability for task organizer in mobile crowdsensing.

Moreover, the current CUCB algorithm primarily relies on a generalized empirical mean of the arms, lacking a tailored design that accounts for the specific characteristics of different arm types (e.g., the independence and observability of the arms) and adaptability to variance changes. This results in excessively wide confidence intervals for exploration. More importantly, when selecting K arms in each round, these enlarged confidence intervals significantly affect the selection, especially in the boundary regions of the unknown parameters. Consequently, this introduces an additional factor of K in the regret calculation, where K can range from hundreds to thousands, depending on the specific application.

B. Our Contributions

Based on the drawbacks and findings aforementioned, this article makes four contributions as follows.

(1) Model Formulation: We introduce the PMC-G model under the dynamic combinatorial arm set, an innovative PMC bandit framework designed to process general feedback through the incorporation of an arm observation probability, also covering the previous volatile combinatorial multi-armed bandit settings (an arbitrary subset of arms is unavailable at any given time instant) [11]. This model categorizes arms into two types: "probabilistically triggered arms" (hereafter referred to as "triggered arms") and "self-reliant arms," which operate independently under the triggering mechanism. Our model demonstrates versatility by accommodating three distinct network applications: mobile crowdsensing, online content delivery, and dynamic channel allocation. These applications are characterized by their unique feedback mechanisms: probabilistic, semi-bandit, and cascading, respectively.

(2) Algorithm Design: We propose two novel variance-aware bandit algorithms tailored for the PMC-G model. The first, a Variance-Adaptive Combinatorial Upper Confidence Bound algorithm (VACUCB), is designed specifically for triggered arms. It utilizes the empirical variance to construct a Bernstein-type confidence interval, which adaptively narrows the Chernoff-type confidence interval used by the CUCB when the arm exhibits low empirical variance. This adjustment significantly reduces unnecessary exploration, leading to tighter regret bounds. The second algorithm, the Action-Based Combinatorial Upper Confidence Bound algorithm (ABCUCB), is optimized for self-reliant arms and leverages the variance-aware reward

feature. Unlike traditional approaches, ABCUCB maintains an upper confidence bound for actions rather than arms and reduces its dependency on the size of the arm selection.

(3) Theoretical Analysis: For VACUCB, we establish that it achieves a regret bound of $O(\sum_{i\in[m]} \frac{|V|\log K\log T}{\Delta^i_{\min}})$. It significantly improves the regret bound of CUCB by a factor of $\tilde{O}(K)$ (where $\tilde{O}$ hides logarithmic factors of K), and matches the lower bound by logarithmic factors. Compared to [12], we further remove $\log K$ dependency in the regret bound. For ABCUCB, we achieve a regret bound of $O\left(\sum_{i\in[m]} \frac{|V|\log T}{\Delta_i^{\min}}\right)$, where the leading term totally removes the $O(\log K)$ dependency. We overcome several technical challenges to prove the improved regret bounds for PMC-G, such as dealing with the non-deterministic observation, associating the triggering probabilities with the expected random triggering event, and bounding the arm over-estimation. One of our key strategies is to use a variance-aware reward sensitivity and smoothness lemma to distribute the total regret across inaccurate estimations.

(4) Experimental Evaluation: In applying the two proposed algorithms to the aforementioned network applications (mobile crowdsensing, online content delivery, and dynamic channel allocation), which are based on specific arm types relevant to PMC-G network problems, we conduct a comprehensive series of experiments. These experiments cover three network applications, using both synthetic and real-world datasets to validate our theoretical findings. The empirical outcomes are compelling, demonstrating that our proposed algorithms consistently achieve over 15% lower regret compared to benchmark algorithms across these varied contexts on both static and dynamic network environments.

II. System Model

The system model of the dynamic PMC bandit with general feedback (or PMC-G in short) can be described by a tuple $(G, [m], \mathcal{S}, \mathcal{D}, D_{\text{obs}}, R)$ as follows: $G = (L, V, E)$ is the underlying bipartite graph, where L is the set of candidate nodes, V is the target nodes to be covered by the L, and E is the set of edges connecting L and V; $[m] = \{1, 2, ..., m\}$ is the set of base arms and each base arm is associated with an unknown parameter to be learned. Depending on different application scenarios in Section V, the base arms for PMC-G could refer to the edge set E, or the edge and target node sets $E \cup V$, therefore we use $[m]$ to cover both cases. Note that the base arms are "*volatile*", meaning that the available arms can vary over time [13]. Based on these volatile arms, $\mathcal{S}$ is defined as the overall set of all combinatorial arms, i.e., all actions, and $\mathcal{S}_t \subseteq \mathcal{S}$ is the set of dynamic sets of eligible actions at round $t \in \mathcal{T} = \{1, 2, ..., T\}$, where $S \in \mathcal{S}_t$ is an individual action. Similar to $[m]$, $\mathcal{S}_t$ varies with the application, is time-dependent, and can be either a collection of subsets of $[m]$, or subsets of L. A specific example of the PMC-G model is as follows: In the context of a CDN, consider a PMC-G model where the graph $G = (L, V, E)$ represents CDN servers (L) and users (V), connected by potential delivery paths (E). The base arms $[m]$ are these edges, each associated with an uncertain success probability. Actions from $\mathcal{S}$ involve selecting subsets of servers to serve subsets of users, with dynamic action sets $\mathcal{S}_t$ reflecting the current network conditions. $\mathcal{D}$ is the set of possible *Bernoulli* distributions over the outcomes of base arms with support $\{0, 1\}^m$; D_{obs} is the observation function to model the general feedback and R is the reward function, the definitions of which will be introduced shortly.

In PMC-G, the learning agent interacts with the unknown environment sequentially as follows. First, the environment chooses a distribution $D \in \mathcal{D}$ unknown to the agent. At round t, the environment reveals the time-dependent action set $\mathcal{S}_t \in \mathcal{S}$, the agent selects an action $S_t \in \mathcal{S}_t$, and the environment draws from the unknown distribution D a Bernoulli outcome $\boldsymbol{X}_t = (X_{t,1}, ...X_{t,m}) \in \{0, 1\}^m$. Intuitively, for $e = (u, v) \in E$, $X_{t,e} = 1$ means the target node $v \in V$ is covered when $u \in L$ is selected and for $v \in V$, $X_{t,v} = 1$ means the target node yields one unit of reward when v is covered. Similar to [7], we assume that the outcome $X_{t,e}$ on edge $e \in E$ is independent with any other outcomes $X_{t,i}$, $i \in [m], i \neq e$, yet the outcomes X_v and $X_{v'}$ of nodes $v', v \in V$ could be dependent or independent, relying on the subsequent arm types.

When the action S_t is played, the agent will receive a non-negative reward $R(S_t, \boldsymbol{X}_t)$. For PMC-G, the reward at round t is the total rewards received from the covered nodes,

$$R(S_t, \boldsymbol{X}_t) = \textstyle\sum_{v\in V} \mathbb{I}\{\exists\, u \in S_t \text{ s.t. } X_{t,(u,v)} = 1\} X_{t,v}. \quad (1)$$

Let $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)$ denote the mean vector of base arms' outcomes, which are unknown initially. Given the independence assumption, the expected reward $r(S; \boldsymbol{\mu}) \triangleq \mathbb{E}[R(S, \boldsymbol{X}_t)]$ is

$$r(S; \boldsymbol{\mu}) = \sum_{v\in V} \mu_v \left(1 - \prod_{u\in S} (1 - \mu_{(u,v)})\right). \quad (2)$$

Note that this expected reward function is highly non-linear and finding the optimal solution S_t^* is NP-hard in general [5], [7]. Fortunately, using the submodular set function maximization technique, one can achieve $(1-1/e)$-approximate solutions [5].

At the end of round t, the agent has the opportunity to observe the outcomes of certain arms as feedback, which is critical to improving future decisions. Let us denote the outcome distribution by D for action S. In this context, we now introduce two different types of arms respectively.

Triggered Arms: Within a randomly selected set $\tau_t \sim D_{\text{obs}}(S_t, \boldsymbol{X}_t)$, the outcomes of arms, denoted as $(X_t)_{t\in\tau_t}$, are disclosed to the agent. This process, governed by the function D_{trig}, models general feedback and is thus termed the *general feedback function*. Typically, the selection of τ_t is influenced by S_t and X_t, although it may also incorporate additional randomness. For ease of reference, we introduce the term *observation probability* $p_i^{D,D_{\text{obs}},S}$, which represents the probability of observing base arm i given action S, under outcome distribution D and feedback function D_{obs}. Given that D_{obs} remains constant within a particular application, we simplify our notation to $p_i^{D,S}$ moving forward. Define $\tilde{S}$ as the set of arms that can be triggered by the action S, i.e., $\tilde{S} := \{i \in [m] : p_i^{D,S} > 0\}$ This set represents the target nodes that can be covered by the selected source node, along with the edges connecting these target nodes under the PMC-G model.

Self-Reliant Arm: We define base arms as self-reliant if, for any distribution $D \in \mathcal{D}$, the Bernoulli outcome vectors $\boldsymbol{X}_t \sim$

D are independent across base arms at round t, with outcomes that can also be probabilistically triggered. Specifically, this means that D can be expressed as a tensor product $\otimes_{i\in[m]} D_i$, where each D_i has an expected mean $\mu_i = \mathbb{E}_{X_i\sim D_i}[X_i]$. In this context, the time-varying set of eligible subsets $\mathcal{S}t \in \mathcal{S}$ still represents the collections available of subsets of $[m]$ at round t. Similarly, only arms triggered by the set $\tau_t \sim D_{\text{obs}}(S_t, \boldsymbol{X}_t)$, with observation probability $p_i^{D,S}$, are disclosed as feedback. Additionally, the outcome distribution D_i for each arm i, with mean μ_i, is considered to be $C_1\mu_i(1-\mu_i)$-sub-Gaussian, where C_1 is a coefficient indicating the variability of each arm's outcomes [14] (which will be explained in Section V).

The former arm type can be observed in CDN (see Fig. 1), where each caching server (i.e., an arm) chooses to distribute content but may fail to deliver it to the user with a certain probability. The latter type is typical in crowdsourcing (see Fig. 2), where participants, who do not know each other, independently engage in tasks across the crowdsourcing initiative.

Note that the PMC-G model significantly generalizes the modeling power of previous PMC bandit [7] as it not only models semi-bandit feedback that is deterministic but can also model the probabilistic feedback when τ_t is randomly determined or even the partial feedback that depends on certain stopping criteria under the dynamic environment, which will be discussed in details in Section V. The goal of PMC-G is to accumulate as much reward as possible over T rounds, by learning the Bernoulli distribution D, or equivalently the unknown mean vector $\boldsymbol{\mu}$. The performance of an online learning algorithm A is measured by its *regret*, defined as the difference of the expected cumulative reward between always playing the best action $S_t^* \triangleq \text{argmax}_{S\in\mathcal{S}_t} r(S;\boldsymbol{\mu})$ and playing actions chosen by algorithm A at each round t. As mentioned before, it could be NP-hard to compute the exact S_t^* even when $\boldsymbol{\mu}$ is known, so similar to [1], [7], [15], we assume that the algorithm A has access to an offline (α,β)-approximation oracle. This oracle, given the mean vector $\boldsymbol{\mu}$, outputs an action S such that $\Pr[r(S;\boldsymbol{\mu}) \geq \alpha \cdot r(S_t^*;\boldsymbol{\mu})] \geq \beta$. For PMC-G applications with the monotone submodular reward in Eq. (2), the offline (α,β)-approximation oracle is typically an $(1-1/e,1)$-approximation greedy oracle, as described in [5], [16]. Consequently, the T-round $(1-1/e,1)$-approximate regret is defined as

$$Reg(T;\boldsymbol{\mu}) = \mathbb{E}\left[\sum_{t=1}^{T}(1-1/e)\cdot r(S_t^*;\boldsymbol{\mu}) - r(S_t;\boldsymbol{\mu})\right], \quad (3)$$

where the expectation is taken over the randomness of outcomes $\boldsymbol{X}_1,...,\boldsymbol{X}_T$, the observation sets $\tau_1,...,\tau_T$, and the inherent randomness of the algorithm A under the dynamic action set.

III. Algorithm Design

In this section, we design two different algorithms under two types of arms for the dynamic PMC-G problem. At a high-level design perspective, VACUCB (Variance-Adaptive Combinatorial Upper Confidence Bound) and ABCUCB (Action-Based Combinatorial Upper Confidence Bound) algorithms share the common goal of optimizing selections in a variance-aware combinatorial multi-armed bandit framework. On the other hand, VACUCB focuses on individual arm variance and updates based on triggered arms, while ABCUCB computes confidence intervals for entire actions with self-reliant arms.

Algorithm 1 VACUCB: Variance-Adaptive Combinatorial Upper Confidence Bound with Triggered Arms

1: **Input:** Base arms $[m]$, offline ORACLE.
2: **Initialize:** For each arm i, $T_{0,i} \leftarrow 0, \hat{\mu}_{0,i} = 0, \hat{V}_{0,i} = 0$.
3: **for** $t = 1, ..., T$ **do**
4: For arm i, compute $\rho_{t,i}$ according to Eq. (4) and set UCB value $\bar{\mu}_{t,i} = \min\{\hat{\mu}_{t-1,i} + \rho_{t,i}, 1\}$.
5: Select action S_t = ORACLE$(\bar{\mu}_{t,1}, ..., \bar{\mu}_{t,m})$ from dynamic time-dependent action set $\mathcal{S}_t \subseteq \mathcal{S}$.
6: Play S_t and observe arms τ_t with outcome $X_{t,i}$ from ORACLE, where $i \in \tau_t$ are the available base arms.
7: For every $i \in \tau_t$, update $T_{t,i} = T_{t-1,i} + 1$, $\hat{\mu}_{t,i} = \hat{\mu}_{t-1,i} + (X_{t,i} - \hat{\mu}_{t-1,i})/T_{t,i}$, $\hat{V}_{t,i} = \frac{T_{t-1,i}}{T_{t,i}}\left(\hat{V}_{t-1,i} + \frac{1}{T_{t,i}}(\hat{\mu}_{t-1,i} - X_{t,i})^2\right)$.
8: **end for**

A. Triggered Arm Algorithm (VACUCB) for PMC-G

Algorithm 1 maintains the empirical estimate $\hat{\mu}_{t,i}$ and $\hat{V}_{t,i}$ for the true mean and the true variance of the base arm outcomes, respectively. As discussed earlier, we follow the principle of Optimism in the Face of Uncertainty (OFU), which guides decision-making by favoring actions with the most optimistic potential outcomes under uncertainty. Specifically, Algorithm 1 computes the upper confidence bound (UCB) value $\bar{\mu}_i = \hat{\mu}_{t,i} + \rho_{t,i}$ as an optimistic estimate of μ_i. Intuitively, confidence interval $\rho_{t,i}$ serves as a bonus term to explore the unknown mean μ_i: when arm i is not observed often (i.e., $T_{t,i}$ is small), $\rho_{t,i}$ will be large and encourages the algorithm to select arm i.

Compared with the CUCB algorithm [7] which uses confidence interval $\rho_{t,i} = \sqrt{\frac{3\log t}{2T_{t-1,i}}}$ based on Chernoff-type concentration bound [17] for the PMC problem, the key difference is that we leverage on the stronger Bernstein-type concentration bound and use empirical variance $\hat{V}_{t-1,i}$ to construct the following "variance-adaptive" confidence interval:

$$\rho_{t,i} = \sqrt{\frac{6\hat{V}_{t-1,i}\log t}{T_{t-1,i}}} + \frac{9\log t}{T_{t-1,i}} \quad (4)$$

As we will show in Section IV, adopting a variance-adaptive interval is crucial for attaining tighter regret bounds. This is particularly relevant because the expected reward in Eq. (2) is highly sensitive to arms with means near 0 or 1. Overestimations ($\rho_{t,i}$) for these arms lead to substantial regret. Interestingly, these arms typically exhibit lower variance, which implies that a variance-adaptive approach effectively minimizes overestimations, thereby reducing the overall regret.

To select the action S_t from the set of available actions $\mathcal{S}_t$ at round t, the next step is to insert the UCB values into the offline $(1-1/e,1)$-approximation oracle. After playing action S_t, the agent will observe a set of arms τ_t as feedback and update the estimation accordingly. Regarding the time complexity, the offline process of finding the best action corresponds to a typical submodular maximization problem [1], [16]. The entire online and offline processes can be analyzed based on

the procedure for selecting actions over T rounds. Each round involves a greedy oracle for selecting K arms from a set of $|L|$ candidate arms. For each arm added to the arm set S, the potential reward $r(S;\boldsymbol{\mu})$ must be computed, which requires evaluating combinations from the bipartite graph $G=(L,V,E)$ and considering all $|L|\times|V|$ potential interactions. The overall time complexity is given by $O(TK|L|^2|V|)$.

B. Self-Reliant Arm Algorithm (ABCUCB) for PMC-G

For the PMC-G problem with self-reliant arms, the Action-Based Combinatorial Upper Confidence Bound (ABCUCB) algorithm is presented in Algorithm 2. Unlike works that maintain a single UCB for each base arm i, ABCUCB maintains a UCB for a super arm S_t (i.e., action) at round t. To circumvent the inefficiencies of a brute-force search through all possible actions [18], we propose a method that incrementally builds the arm set. At each round t, we initialize a temporary arm set S'_t and sequentially add one available base arm at a time until S'_t comprises K arms, forming the action S_t. The core challenge is devising a method that can efficiently handle the selection process over a non-linear set function.

For any set S, the function $r(S;\boldsymbol{\mu})$ is observed to be monotone and submodular. We leverage this property by designing a confidence interval $\rho_t(S)$ such that the optimistic reward $\bar{r}_t(S) := r(S;\hat{\boldsymbol{\mu}})+\rho_t(S)$ retains these properties, where $\hat{\boldsymbol{\mu}}=\{\hat{\mu}_i\}$ for any $i\in S$. This allows the use of a greedy $(1-1/e,1)$-approximation oracle, based on the fact that the sum of two submodular functions remains submodular.

Specifically, with $\tilde{S}$ denoting the set of arms that can be triggered by action S, let $T^{\min}_{t-1,\tilde{S}}$ represent the minimum count of triggers for the arm set $\tilde{S}$ by S, i.e., $T^{\min}_{t-1,\tilde{S}}=\min_{i\in\tilde{S}} T_{t-1,i}$. We define σ_{t-1} as the maximum of the following: $\sigma_{t-1}=\max\left\{\sqrt{\sum_{i\in S}\frac{\log(2|\mathcal{S}_t|T)}{T^2_{t-1,i}}},\frac{\log(2|\mathcal{S}_t|T)}{T^{\min}_{t-1,\tilde{S}}}\right\}$. Subsequently, the confidence interval in Line 5 is defined as:

$$\rho_t(S)=\sqrt{\sum_{i\in\tilde{S}}\frac{C_1|V|}{T_{t-1,i}}+8\sigma_{t-1}C_1|V|}\,, \tag{5}$$

where C_1 is a sub-Gaussian coefficient. Eq. (5) complies with concentration bounds for sub-exponential random variables [19], [20], assuming that the estimation error ζ_i for each base arm $i\in S$ behaves as independent sub-Gaussian random variables, influenced by the variance-aware reward smoothness (see Property 2 in Section IV). This aggregated approach results in a sub-Exponential distribution that is more concentrated, thus enhancing the estimation's accuracy over potentially dependent variables. Importantly, ABCUCB employs the minimum counter $T^{\min}_{t-1,\tilde{S}}$ for constructing the second segment of the interval, instead of aggregating all counters $T_{t-1,i}$ for the arm set S. Then, the algorithm selects the base arm i^* that offers the highest incremental reward (Line 6).

Once S'_t expands to include K arms, Algorithm 2 selects $S_t=S'_t$ and consults the oracle to observe the outcomes for round t. It then updates the relevant estimates, similar to Algorithm 1. The design of $\rho_t(S)$ ensures that the optimistic reward $\bar{r}_t(S)$ retains its monotone submodular property, which facilitates efficient optimization through a greedy $(1-1/e,1)$-approximation oracle. Based on this, the time complexity of ABCUCB mirrors that of VACUCB, which is also $O(TK|L|^2|V|)$. This approach avoids the enumeration method with $O(|\mathcal{S}|T)$, where $|\mathcal{S}|$ can be exponentially large.

Algorithm 2 ABCUCB: Action-Based Combinatorial Upper Confidence Bound with Self-Reliant Arms

1: **Input:** Base arms $[m]$, coefficient C_1, offline ORACLE.
2: **Initialize:** For each arm i, $T_{0,i}=0,\hat{\mu}_{0,i}=0$.
3: **for** $t=1,...,T$ **do**
4: **for** $|S'_t|\le K$ **do**
5: For each $i\in([m]\setminus S'_t)\cap\mathcal{S}_t$, set UCB value $\bar{r}_t(S'_t\cap\{i\})=r(S'_t\cap\{i\};\hat{\boldsymbol{\mu}}_{t-1})+\rho_t(S'_t\cap\{i\})$, where confidence interval $\rho_t(S'_t\cap\{i\})$ is computed according to Eq. (5).
6: Select $i^*=\arg\max_{i\in([m]\setminus S'_t)\cap\mathcal{S}_t}\bar{r}_t(S'_t\cap\{i\})-\bar{r}_t(S'_t)$, and set $S'_t=S'_t\cap\{i^*\}$.
7: **end for**
8: Set $S_t=S'_t$, play S_t, and observe arms τ_t with outcome $X_{t,i}$ from ORACLE, where $i\in\tau_t$ are the available arms.
9: For every $i\in\tau_t$, update $T_{t,i}=T_{t-1,i}+1$, $\hat{\mu}_{t,i}=\hat{\mu}_{t-1,i}+(X_{t,i}-\hat{\mu}_{t-1,i})/T_{t,i}$.
10: **end for**

IV. Performance Analysis

In this section, we present our main theoretical results, the related analysis, and some discussions for two algorithms. For ease of exposition, we put the proofs in the Appendix.

A. Analysis Preliminaries

Initially, we present some definitions in the dynamic action set. In the following subsections, we primarily present a distribution-dependent regret analysis based on the suboptimality gap defined below. For distribution-independent regret analysis, please refer to Appendix, available online.

Definition 1 (Suboptimality Gap)**.** *Fix a distribution $D\in\mathcal{D}$ and its mean vector $\boldsymbol{\mu}$, for each action $S\in\mathcal{S}$, we define the (approximation) gap as $\Delta_{t,S}=\max\{0,\alpha r(S^*_t;\boldsymbol{\mu})-r(S;\boldsymbol{\mu})\}$, where $S^*_t=\operatorname{argmax}_{S\in\mathcal{S}_t} r(S;\boldsymbol{\mu})$. For each arm i, we define $\Delta^{\min}_i=\inf_{S\in\mathcal{S}_t,\mathcal{S}_t\in\mathcal{S},t\in\mathcal{T}:p^{D,S}_i>0,\,\Delta_{t,S}>0}\Delta_{t,S}$, $\Delta^{\max}_i=\sup_{S\in\mathcal{S}_t,\mathcal{S}_t\in\mathcal{S},t\in\mathcal{T}:p^{D,S}_i>0,\,\Delta_{t,S}>0}\Delta_{t,S}$. As a convention, if there is no action $S\in\mathcal{S}$ such that $p^{D,S}_i>0$ and $\Delta_{t,S}>0$, then $\Delta^{\min}_i=+\infty,\Delta^{\max}_i=0$. We define $\Delta_{\min}=\min_{i\in[m]}\Delta^{\min}_i$ and $\Delta_{\max}=\max_{i\in[m]}\Delta^{\min}_i$.*

Our analysis uses several events to filter the total regret and then bound these event-filtered regrets accordingly. Below we define the event-filtered regret.

Definition 2 (Event-filtered Regret)**.** *For any series of events $(\mathcal{E}_t)_{t\ge0}$ indexed by round number t, we define the $Reg^A_{\alpha,\boldsymbol{\mu}}(T,(\mathcal{E}_t)_{t\ge0})$ as the regret filtered by events $(\mathcal{E}_t)_{t\ge0}$, or the regret is only counted in t if $\mathcal{E}$ happens in t. Formally, $Reg^A_{\alpha,\boldsymbol{\mu}}(T,(\mathcal{E}_t)_{t\ge0})=\mathbb{E}\left[\sum_{t\in[T]}\mathbb{I}(\mathcal{E}_t)(\alpha\cdot r(S^*_t;\boldsymbol{\mu})-r(S_t;\boldsymbol{\mu}))\right]$. For simplicity, we will omit $A,\alpha,\boldsymbol{\mu},T$ and rewrite $Reg^A_{\alpha,\boldsymbol{\mu}}(T,(\mathcal{E}_t)_{t\ge0})$ as $Reg(T,\mathcal{E}_t)$ when contexts are clear.*

B. Performance Guarantee of VACUCB

1) Main Result: We first give our main result of VACUCB, following the given definitions and lemmas.

Theorem 1. *For a PMC-G problem instance $(G, [m], \mathcal{S}, \mathcal{D}, D_{obs}, R)$, the regret of VACUCB (Algorithm 1) is bounded by $O\left(\left(\sum_{i\in[m]} \frac{|V|\log K}{\Delta_i^{\min}} + \log\left(\frac{K}{\Delta_i^{\min}}\right)\right)\log T\right)$.*

2) Proof Analysis: To give concrete events filtering the leading regret, we leverage the following property.

Property 1 (Variance-aware Reward Sensitivity)**.** *For PMC-G with semi-bandit, probabilistic, and cascading feedback model, and any parameter change $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [0,1]^m$ s.t. $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$, the reward sensitivity $r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})$ satisfies*

$$r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu}) \le \sqrt{|V|}\,\|\boldsymbol{x}_v\|_2 + \|\boldsymbol{x}_1\|_1\,, \tag{6}$$

where $\boldsymbol{x}_v \triangleq \left(\frac{p_i^{D,S}\zeta_i}{\sqrt{(1-\mu_i)\mu_i}}\right)_{i\in[m]}$, $\boldsymbol{x}_1 \triangleq \left(p_i^{D,S}\eta_i\right)_{i\in[m]}$.

Intuitively, this property bounds the reward difference with the impact of variance considered, by ℓ_2 and ℓ_1 norm of each arm's over-estimation $\boldsymbol{x}_v$ and $\boldsymbol{x}_1$ given by VACUCB, and $\sqrt{|V|}$ is to bound the non-linearity of $r(S;\boldsymbol{\mu})$. Notice that both $x_{v,i}$ and $x_{1,i}$ are re-weighted by $p_i^{D,S}$ which reduces the regret contribution from unlikely observed arms to handle the general feedback model.

We have the following lemma for the regret decomposition.

Lemma 1 (Regret Decomposition)**.** *We define two error terms*

$$e_{t,1}(S_t) = 4\sqrt{3}\sqrt{|V|}\sqrt{\sum_{i\in\tilde{S}_t}\left(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28}\right)(p_i^{D,S_t})^2} \tag{7}$$

$$e_{t,2}(S_t) = 28\sum_{i\in\tilde{S}_t}\left(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28}\right)(p_i^{D,S_t}) \tag{8}$$

and two events $E_{t,1} = \{\Delta_{S_t} \le 2e_{t,1}(S_t)\}$, $E_{t,2} = \{\Delta_{S_t} \le 2e_{t,2}(S_t)\}$. The regret of Algorithm 1, when used with (α,β) approximation oracle is bounded by

$$Reg(T) \le Reg(T, E_{t,1}) + Reg(T, E_{t,2}) + \frac{2\pi^2}{3} m\Delta_{\max}. \tag{9}$$

Our final step is to bound $Reg(T, E_{t,1})$ and $Reg(T, E_{t,2})$, which corresponds to the first term and the second term in Theorem 1, respectively. By employing a refined reverse amortization technique originated in [15], we allocate the regret $\Delta_{t,S}$ across base arms using carefully designed thresholds. Furthermore, we establish a probability equivalence to associate the triggering probabilities with the expected random triggering event, which is a highly non-trivial endeavor. We defer the detailed proofs of $Reg(T, E_{t,1})$ and $Reg(T, E_{t,2})$ in Appendix.

3) Discussions: The leading term in the regret bound is $O(\sum_{i=1}^m \frac{|V|\log K\log T}{\Delta_i^{\min}})$ when gaps are not too large, i.e., $\Delta_{\min}^i \le |V|^{1-\epsilon}/\log K$, for any $\epsilon > 0$. The dependence over K is $O(\log K)$. For PMC bandit with general feedback, [15] can only give $O(\sum_{i=1}^m \frac{|V|K\log T}{\Delta_i^{\min}})$ for PMC-G. Our result is strictly better than theirs by a factor of $O(K/\log K)$. For the classical PMC bandit with semi-bandit feedback, [21] recently gives a regret lower bound $\Omega(\frac{L|V|^2}{\Delta_{\min}})$, which means our regret bound is near-optimal (by setting $m = L|V|, \Delta_{\min} \le \Delta_i^{\min}$) and matches the lower bound up to $O(\log K)$. Compared to [10], [12], we have eliminated a factor of $O(\log K)$ from the second main term in the regret upper bound of [10] and the overall regret upper bound in [12] by minimizing the impact of observation randomness.

C. Performance Guarantee of ABCUCB

1) Main Results: To state the regret bound, let $p_{\tilde{S}}^{D,S}$ be the probability that the arm set $\tilde{S}$ is triggered when super arm S is selected and $p^* = \min_{S\in\mathcal{S}: p_{\tilde{S}}^{D,S}} p_{\tilde{S}}^{D,S}$. The following theorem summarizes the regret upper bound for ABCUCB.

Theorem 2. *For a PMC-G problem instance $(G, [m], \mathcal{S}, \mathcal{D}, D_{obs}, R)$, the regret of ABCUCB (Algorithm 2) is bounded by $O\left(\sum_{i\in[m]} \frac{|V|}{p^*\Delta_i^{\min}}\log T + \frac{m|V|K}{p^*\Delta_{\min}}\right)$.*

2) Proof Analysis: We use the following lemma to bound the reward difference. This represents the triggering version with independent arms of Property 1.

Property 2 (Variance-aware Reward Smoothness)**.** *For PMC-G with independent instance $([m], \mathcal{S}, \mathcal{D}, R)$, if for any action $S \in \mathcal{S}$, for any parameter change $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^m$ s.t. $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$, the reward smoothness $r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})$ satisfies*

$$|r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})| \le \sqrt{|V|}\,\|\boldsymbol{x}'_v\|_2 + \|\boldsymbol{x}'_1\|_1\,, \tag{10}$$

where $\boldsymbol{x}_v \triangleq \left(\frac{\zeta_i}{\sqrt{(1-\mu_i)\mu_i}}\right)_{i\in\tilde{S}}$, $\boldsymbol{x}_1 \triangleq (\eta_i)_{i\in\tilde{S}}$ for the triggering arm set $\tilde{S}$ of action S.

In contrast to Property 1, Property 2 is unidirectional, allowing both $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ to assume negative values, as empirical mean reward may not always exceed the unknown true mean reward of the base arm.

We first define the error term $e_t(S_t) = 2\rho_t(S_t)$ as in Line 5. We focus on the regret conditioned on the event $\{\Delta_{t,S} \le e_t(S_t)\}$. The central strategy leverages Property 2, establishing a bound on $|r(S;\hat{\boldsymbol{\mu}}) - r(S;\boldsymbol{\mu})|$. Denoting $u_{t,i}$ as a sub-Gaussian random variable and $Y_{t,S} = \sum_{i\in\tilde{S}} u_{t,i}^2$, $Y_{t,S}$ behaves as a sub-Exponential random variable. Applying concentration bounds on $Y_{t,S}$, we derive the specified form of $e_t(S_t)$. Next, we dissect the scenario based on the magnitude of $\sum_{i\in\tilde{S}_t}\frac{1}{T_{t-1,i}}$. In both cases, we employ the reverse amortization technique [15]. Moreover, the definition of $e_t(S_t)$ allows us to limit our consideration of regret to the contributions from the minimum-arm in S_t, leading to eliminating the $O(\log K)$ term in the regret upper bound.

3) Discussions: Examining the regret bound presented above, the leading term eliminates the $O(\log K)$ dependency, in contrast to Theorem 1. [10] considers a simpler scenario involving independent but non-triggering arms, a case that our work can address by setting $p_i^{D,S} = 1$ for arm i and $p_i^{D,S} = 0$ for the others. Compared to the findings in [22], our regret bound shows an improvement by a factor of $O(\log^2 K)$. Regarding the applications of PMC-G, Theorem 2 provides an improvement of $O(\sum_{i\in[m]} \frac{|V|\log T}{\Delta_i^{\min}})$.

V. Applications for PMC-G

We consider three applications with semi-bandit, probabilistic, and cascading feedback to illustrate the utility of our PMC-G framework: mobile crowdsensing, online content delivery, and dynamic wireless channel allocation. We compare the regret of our VACUCB and ABCUCB algorithms to two baselines: CUCB [1], a state-of-the-art combinatorial bandit algorithm that does not use variance-adaptive confidence intervals; and ϵ-greedy, which explores new actions with fixed probability ϵ and otherwise greedily chooses the empirically optimal action. In the context of mobile crowdsensing, we extend our comparison to include ESCB [18], which is notable for its efficient exploitation strategy within the *semi-bandit* feedback. Moreover, risk-averse multi-armed bandit, which also incorporates the variance term and is popular in financial portfolio selection [23]–[25], is compared on reward outcomes in Appendix, given its different regret definition from Eq. (3). In our theoretical analysis, ABCUCB (Algorithm 2) addresses the self-reliant arms, which are independent and can be probabilistically triggered. Under the network applications of mobile crowdsensing, online content delivery, and dynamic wireless channel allocation, where the dependence of arms increases from semi-bandit to probabilistic to cascading feedback, we still explore the performance of ABCUCB.

TABLE I: Summary of feedback and oracles for different PMC network problem applications.

Application	**Feedback**	(α, β)**-Oracle**
Mobile Crowdsensing	Semi-bandit	Greedy, $(1-1/e, 1)$
Online Content Delivery	Probabilistic	Greedy, $(1-1/e, 1)$
Dynamic Channel Allocation	Cascading	Greedy, $(1-1/e, 1)$

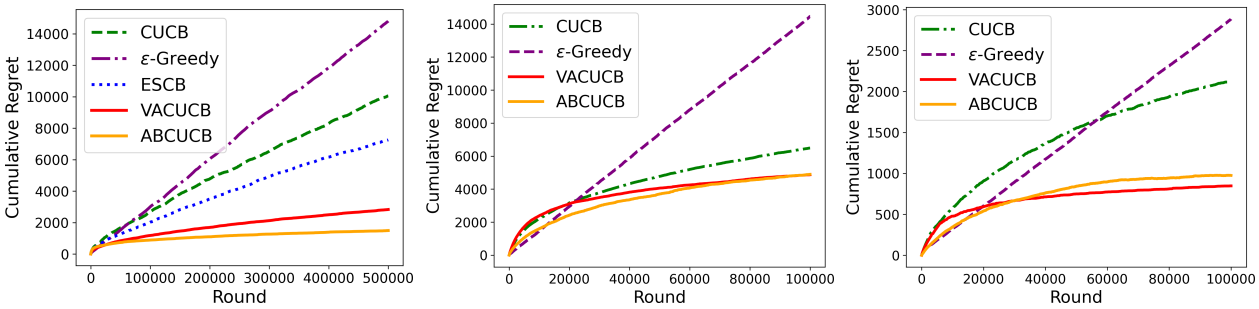
A. Mobile Crowdsensing

1) Problem Description: Mobile devices today, including smartphones, tablets, and wearables, come equipped with advanced sensors like GPS, accelerometers, and gravity sensors. These sensors enable the devices to gather and analyze environmental data based on the users' locations. Mobile crowdsensing leverages this capability, organizing individuals to collect data across various locations using their personal mobile devices as they move around an area. This approach is particularly beneficial for large-scale sensing projects, allowing for the aggregation of diverse data points across a broad geographic area [26]. For example, a task organizer may want to organize a group of participants and use their cameras, gravity sensors, and GPS as sensors to monitor dust levels or a possible earthquake in a large city [27]. However, the quality of data obtained through mobile crowdsensing can be inconsistent, influenced by factors like the different paths participants travel and the variable quality of sensors across devices. Additionally, participant availability can fluctuate due to personal obligations or device constraints [3]. The goal of the mobile crowdsensing task organizer is to select a group of individuals to maximize the amount of high-quality data collected from different locations in the city.

The mobile crowdsensing application can be modeled by our PMC-G problem. Consider a bipartite graph $G(L, V, E)$, where L is the set of candidate participants, V is the set of locations in a city, and E models the data collection process. At each time t, the agent (or the task organizer) wants to choose at most K participants to conduct the sensing task. For example, K may be chosen based on a budget for paying fixed recruitment incentives to each chosen participant. However, note that the availability of participants varies, with not all being accessible for tasks at any given time. Each selected participant $u \in S_t$ independently uploads their sensor data at location $v \in V$, which is modeled as a Bernoulli random variable $X_{t,(u,v)} \in \{0, 1\}$ with probability $\mu_{u,v}$ that the data can be used as valid information to cover location v. In this case, we know the arms are exactly E. The agent can get *semi-bandit feedback*, i.e., observe whether the uploaded data is valid or not for (u, v) s.t. $u \in S_t$. Using the PMC-G formulation, the observation probability $p^{D,S_t}_{(u,v)} = 1$ if $u \in S_t$ or 0 otherwise. The reward is the weighted total number of locations that are covered with valid information: $r(S; \boldsymbol{\mu}) = \sum_{v \in V} \mu_v \left(1 - \prod_{u \in S}(1 - \mu_{(u,v)})\right)$, where the known weight μ_v represents the importance of covering location v to the crowdsensing task. Busy areas, for example, may have higher sensing importance as their environmental conditions affect more people.

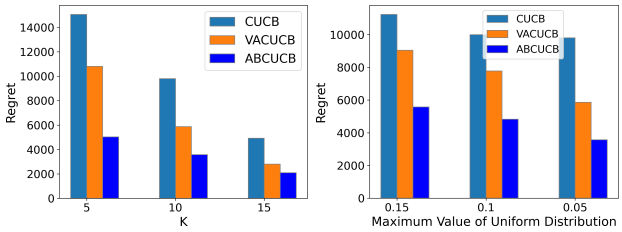
2) Performance Evaluation: **Fixed Action Set.** To simulate the mobile crowdsensing problem, we employ a complete bipartite graph comprising 17 candidate nodes (representing participants) and 30 target nodes (denoting locations). The significance of each location is determined by weights drawn from a uniform distribution $U(0, 0.5)$, with these weights being known to the task organizer. Selecting $K = 15$ participants, we model the success probability $\mu_{u,v}$ for each participant-location pair using a uniform distribution $U(0, 0.15)$. According to [14] where the value of C_1 can be set as $C_1 = \max_{i \in [m]} \frac{1-2\mu_i}{2\ln(\frac{1-\mu_i}{\mu_i})(1-\mu_i)\mu_i}$ for Bernoulli arms with mean μ_i, we utilize a sub-Gaussian parameter of $C_1 = 3$ for the ABCUCB algorithm. Fig. 3a displays the cumulative regret observed across different algorithms throughout $500,000$ rounds (we choose $\epsilon = 0.2$ for the ϵ-greedy algorithm in all experiments). The performance of the VACUCB algorithm is notably superior to other strategies, outstripping the ESCB by 50%, the CUCB by 63%, and the ϵ-greedy by 79%. ABCUCB further showcases remarkable improvement percentages, which eclipses VACUCB by 47%.

To verify how K would affect the regret, we then generate each $\mu_{u,v}$ with $U(0, 0.05)$ and show the total regret for different K after $100,000$ rounds with $|L| = 20$ in Fig. 4a. Note that with the change of K, the optimal reward will also change, which explains why the regret of a small K is larger than that of a large K. We find that with the increase of K, VACUCB's improvement over the CUCB baseline also increases (25% for $K = 5$ and 50% for $K = 15$), which is consistent with our theoretical result in Theorem 1. Additionally, we observe a diminishing return in regret improvement for ABCUCB as K escalates to 15, diverging from the trends noted at $K = 5$ and $K = 10$. This trend indicates that ABCUCB's sensitivity to variations in K is less pronounced than that of VACUCB, corroborating with the findings in Theorem 2.



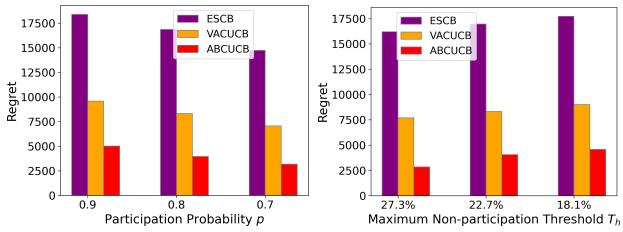
(a) Mobile Crowdsensing (b) Online Content Delivery (c) Dynamic Channel Allocation

Fig. 3: Cumulative regret across three network applications under fixed action set: averaged outcomes from 20 experiments, highlighting our proposed algorithms via the solid line. Only in mobile crowdsensing, each arm operates independently.



(a) Varying K (b) Varying x ($\mu_{u,v} \sim U(0, x)$)

Fig. 4: Total regret after 100000 rounds in static settings.



(a) Varying p (b) Varying T_h

Fig. 5: Total regret after 100000 rounds in dynamic settings.

Fig. 4b compares the total regret of CUCB and VACUCB when varying the value of $\mu_{u,v}$. We set $K = 10$ and generate each $\mu_{u,v}$ with $U(0, x)$, where $x \in \{0.05, 0.1, 0.15\}$. As x enlarges, the regret for ABCUCB, VACUCB, and CUCB escalates correspondingly. This increment is attributed to the widening gap between the rewards from the algorithm-selected participants and those selected based on unknown optimal action as $U(0, x)$'s range broadens. Nevertheless, even at the minimal regret increment, ABCUCB and VACUCB manifest notable performance enhancements of 19% and 50%, respectively, when benchmarked against CUCB.

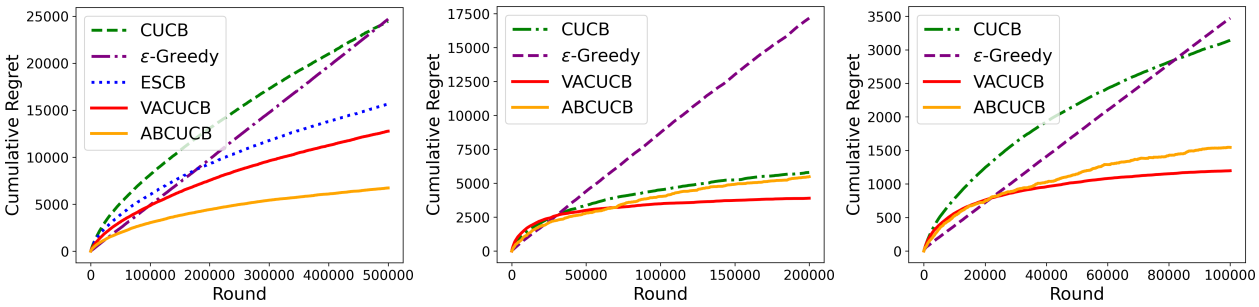
Dynamic Action Set. Furthermore, consistent with the above setting, for the tasks distributed by the task organizer, each crowdsensing participant now has a certain participation probability p (set as 0.8) to decide whether they can engage in the mobile crowdsensing task. Additionally, to facilitate the seamless execution of the crowdsensing tasks, a maximum non-participation threshold is established at $T_h = 22.7\%$, applicable to the overall count of 22 participants. Concisely, as shown in Fig. 6a, the VACUCB algorithm demonstrates superior performance relative to other strategies, achieving improvement percentages of 47% over CUCB and 18% over ESCB. In a similar vein, ABCUCB registers even more pronounced improvements: 72% over CUCB, 56% above ESCB, and outperforming VACUCB by 46%.

We then explore the impact of the probability of participant engagement and the maximum non-participation threshold under the number of candidate participants $L = 23$. As shown in Fig. 5, with a decrease in participation probability p or an increase in the maximum non-participation threshold T_h, the number of available participants decreases. In other words, both the arm and action search space are reduced, leading to a decrease in regret. Moreover, it is observable that under variations in either p or T_h, our proposed algorithms, ABCUCB and VACUCB, outperform the benchmark ESCB, demonstrating the robustness of our algorithms.

B. Online Content Delivery

1) Problem Description: We investigate the challenge of online content delivery within content delivery networks (CDNs), a critical component of web services such as video streaming, web browsing, and software distribution, as documented in existing literature [2], [6], [28]. Unlike traditional approaches that rely on a single central server, CDNs distribute and cache content across multiple mirror servers, enabling end users to retrieve data from the nearest available source. This architecture significantly improves the speed and reliability of content delivery. Our proposed model and algorithm are designed to assist content providers, such as media firms or e-commerce platforms, in selecting an optimal set of mirror servers to maximize user satisfaction. Furthermore, we acknowledge the dynamic nature of CDN environments, where the availability of mirror servers can fluctuate due to maintenance, high traffic, failures, or other network disruptions.

The above application scenario naturally fits into our PMC-G problem with a bipartite graph $G(L, V, E)$, where L models the set of candidate servers, V are the end users, and E models the user-server interactions as follows. At each time slot t, the agent (or the content owner) needs to choose $S_t \subseteq L$ mirror servers that can send contents to users via the CDN network. We assume the number of selected servers at each round is less than K, i.e., $|S_t| \leq K$, since the maintenance of each server usually incurs certain costs, and the content owner has a limited budget. The selected servers $u \in S_t$ then independently



(a) Mobile Crowdsensing (b) Online Content Delivery (c) Dynamic Channel Allocation

Fig. 6: Cumulative regret across three network applications under dynamic action set: averaged outcomes from 20 experiments, highlighting our proposed algorithms via the solid line. Only in mobile crowdsensing, each arm operates independently.

send contents for each user $v \in V$ with unknown success rates $\mu_{(u,v)}$, depending on varying geometric distances and the network congestion [29]. By "success," we mean the content is delivered in time, which can be modeled by a Bernoulli random variable $X_{t,u,v} \in \{0,1\}$ with mean $\mu_{(u,v)}$. We suppose that each user v attempts to preload content from the selected servers to its device, and we use a Bernoulli random variable with unknown mean μ_v to represent whether this preloaded content is ultimately consumed (e.g., video is viewed) by the user [2], [30]. To this end, we can see that arms correspond to the success probability $\mu_{(u,v)}$ for $(u,v) \in E$ and the consuming probability μ_v for users $v \in V$. The question is how to select K mirror servers for content delivery to maximize the total number of users that consume the contents with unknown success rates and consuming probabilities. Every server is subjected to a certain probability of either becoming unavailable or resuming availability. Moreover, server availability is variable, with each facing a distinct likelihood of downtime or reactivation. A strategic server selection policy must, therefore, aim to optimize content delivery to those users with higher propensities for content utilization, accounting for uncertain success rates and varying server availability.

Regarding feedback, the agent can observe whether the content is successfully delivered from the selected servers, i.e., $X_{t,(u,v)}$ for $u \in S_t, v \in V$. We know that the observation probability $p^{D,S_t}_{(u,v)}$ equals 1 if $u \in S_t$ and 0 otherwise, which is known as semi-bandit feedback. If user v successfully receives the content, the agent (i.e., the content owner or CDN provider) can observe whether the user consumes the content, i.e., $X_{t,v}$ is observed when $\exists v$ s.t. $X_{u,v} = 1$. This feedback is called *probabilistic feedback* since it depends on other random outcomes and the probability of observation $p^{D,S_t}_v = 1 - \prod_{u \in S_t}(1 - \mu_{u,v})$. In Table I, the result of $\alpha = 1 - 1/e$ for the dynamic channel allocation application represents the worst-case scenario for algorithms based on base arm selection and action selection, where the arm-based algorithm can achieve an optimal solution with $\alpha = 1$. The expected reward is essentially Eq. (2) and the agent's goal is to minimize the total regret in Eq. (3).

2) Performance Evaluation: **Fixed Action Set.** For the online content delivery experiments, we consider 10 mirror servers located at some of the point-of-presence (POP) locations of Microsoft Azure CDN in North America[1]. We assume the users are distributed in 20 POP locations (including the servers' locations). We extract the average latency data between these locations[2], and assume the realized latency at each round is the average latency plus a random delay ranging from 0ms to 30ms, which is 76% of the average observed delay. We simulate the random delivery deadlines of the contents with the range from 10ms to 20ms. The users will successfully receive the content if their latencies to the mirror servers are less than the delivery deadline. The probability that user v will consume content, μ_v, is sampled from $U(0, 0.5)$ and is unknown to the server selector. Fig. 3b shows the cumulative regret of different algorithms for $100,000$ rounds when all servers are operational. Notably, VACUCB demonstrates a 32% and 65% reduction in regret compared to CUCB and ϵ-greedy, respectively. Moreover, ABCUCB and VACUCB exhibit similar overall performance.

Dynamic Action Set. In the setting of the dynamic combinatorial arm set, the availability of the server depends on its load levels. Monitoring these conditions allows the identification of servers under heavy loads, which are more likely to be unavailable. Accordingly, each round is designed to account for up to two servers becoming unavailable, while a selection of no more than six servers is made from the pool for content delivery purposes. As shown in Fig. 6b, the VACUCB algorithm outperforms comparative strategies, registering a 33% improvement over CUCB and a 77% enhancement relative to the ϵ-greedy approach. As the frequency of observations decreases and the dependency among arms increases relative to the mobile crowdsensing context, the estimation errors for each base arm deviate from behaving as independent sub-Gaussian random variables. Under these conditions, VACUCB exhibits a 27% performance improvement over ABCUCB.

C. Dynamic Channel Allocation

1) Problem Description: We consider a centralized dynamic channel allocation problem where a central controller chooses K channels from the candidate channel set L and allocates them to a group of users V. Each channel $i \in L$ can be viewed as a base arm with unknown Bernoulli availability. Similarly to the centralized online channel allocation setting

[1]https://docs.microsoft.com/en-us/azure/cdn/cdn-pop-locations
[2]https://wondernetwork.com/pings

in [4], [31], where users are assigned predetermined disjoint channel lists to avoid collisions, the controller allocates a specific sublist of these channels to each user. A user will receive a reward only if at least one of the allocated channels is available in a given round. The expected total reward of all users is then $\sum_{j\in V}\left(1-\prod_{i\in S_{j,t}}(1-\mu_i)\right)$, where $S_{j,t}$ is the set of channels assigned to the user j in round t and μ_i is the expected availability of the channel i. Unlike the offline NP-hard problem in [4], the offline optimization problem with such a reward function can be exactly solved by a greedy algorithm that sequentially allocates channels with the maximum marginal returns to users. As shown in Fig. 7, each user j will have an ordered list of allocated channels $o_j^t=(o_{j1}^t,o_{j2}^t,\cdots)$ with length $|S_{j,t}|$. We consider cascading feedback in this application, as each user will sequentially check the availability of allocated channels and stop when finding the first available one to send data. More specifically, only the outcomes of o_{jl}^t for all $l\le L_t$ are observed, where L_t is the index of the first channel available in the list ($L_t=|S_{j,t}|$ if all channels in the list are unavailable).

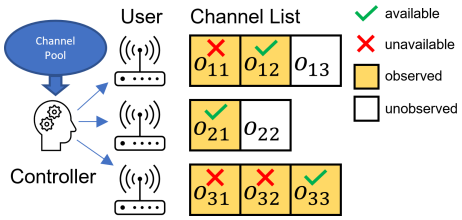


Fig. 7: Illustration of centralized dynamic channel allocation.

In addition, certain channel allocation scenarios, such as those in cognitive radio networks, differentiate between primary and secondary users. Primary users, having licensed access to specific spectrum bands, are prioritized in channel allocation and remain unaffected by secondary user activities [32], [33]. The necessity to avoid interference with high-priority primary users introduces the challenge of opportunistic spectrum access for secondary users [34]. This challenge entails secondary users accessing the available spectrum solely when it is not in use by primary users, necessitating consideration of channel unavailability in the allocation process for secondary users.

2) Performance Evaluation: **Fixed Action Set.** Following in [4], we utilize a real wireless data trace [35] that contains the availability of 16 channels. We choose the most competitive 4 channels among them with average available probabilities μ_i less than 0.1, and consider 4 copies of each to build the candidate channel set with $|L|=16$. Also, there is no real optimal online policy, so we adjust our regret to compare with the optimal policy when assuming the channel availability is uniformly sampled from the whole data trace, i.e., the expected availability is equal to the average availability. We consider a central controller that chooses $K=8$ channels from the 16 candidate channels and allocates them to $|V|=4$ users. Fig. 3c shows the cumulative regret of different algorithms, where VACUCB achieves 13%, 57%, and 68% less regret than the ABCUCB, CUCB, and ϵ-greedy algorithms, respectively.

Dynamic Action Set. Expanding on the above dynamic channel allocation problem, we establish a more complex setting in a cognitive radio network scenario, informed by the same wireless data trace [35], but with a configuration of 20 channels. While maintaining the selection of 8 channels for 4 secondary users, the scenario now includes four primary users who favor channels with higher availability according to the data traces. Should any primary user choose a channel, the availability probability for that channel drops to 0 for the secondary users. Fig. 6c shows that the VACUCB algorithm achieves a 62% improvement over CUCB and a 66% improvement over ϵ-greedy, respectively. Despite the strong dependency between arms due to cascading feedback impacting the observability of outcomes, ABCUCB manages to outperform CUCB by 50% and the ϵ-greedy strategy by 55%.

VI. Related Work

There has been vast literature focusing on online learning problems under the multi-armed bandit (MAB) model, which was first studied by [36] and then extended by many other works (cf. [37]–[40]). The principle of Optimism in the Face of Uncertainty (OFU) [41] is one of the most fundamental concepts in MAB, and has been widely used in MAB algorithms [39]. While most algorithms rely on Hoeffding-type concentration bounds to build the upper confidence bound (UCB) of an arm, a few works [22], [42], [43], apply Bernstein-type bounds and successfully show superior performance, both in theory and in experiments.

Probabilistic maximum coverage (PMC) problem [1] is a widely studied topic with many applications in computer science, especially in the area of network optimization. Besides the three applications mentioned in this article, PMC also covers many other applications, including wireless sensor placement [44] and social network advertising [43], [45]. The online learning version of the PMC problem (or PMC bandit) is first proposed by [7], and then followed by [1], [22]. Different from these works that only consider the semi-bandit feedback, we propose a new PMC-G model that generalizes the semi-bandit feedback and can model broader applications with the general probabilistic feedback and the cascading feedback.

The classic variance-aware algorithms can be traced back to [42]. In contrast to reinforcement learning (RL) works with variance considered [46], [47], our study examines a different setting, as we do not account for state transitions. Regarding variance-aware bandits, [48] concentrates on distribution-independent regret bounds for cascading bandits. [10] explores the smoothness condition that incorporates variance information. Our research focuses on the context of specific network applications within the novel dynamic PMC-G model, successfully reducing the dependence on K on regret upper bounds.

The stochastic combinatorial MAB (CMAB) has received much attention recently [1], [7], [15], [21], [22], [49], [50], and PMC bandit fits into the CMAB framework. For CMAB with semi-bandit feedback, [49] is the first study on stochastic CMAB, and its regret bound has been improved by [8], [18]. Later, [1], [15] considered probabilistic feedback to generalize the semi-bandit feedback model. However, all CMAB frameworks above suffer an additional $O(K)$ factor in their regret bound and the best of them only achieve

$O(\sum_{i\in[m]}(K\log T)/\Delta^i_{\min})$, since they use combinatorial upper confidence bound (CUCB) algorithms that ignore the variance of the arm. [12] presents initial results for the CMAB model in the static combinatorial arm set scenario. [11] examines the CMAB with volatile arms and submodular rewards. Our article provides a more comprehensive analysis including both triggered and self-reliant arms.

VII. Conclusion

In this article, we propose a general variance-aware PMC bandit model, which is equipped with a general feedback mechanism designed to cater to a wide array of network applications under both static and dynamic action set settings. We develop variance-aware online learning algorithms specifically developed for two distinct arm types: triggered arms and self-reliant arms. We establish that each algorithm consistently outperforms in terms of regret minimization. To corroborate our theoretical claims in network problems, we embark on empirical studies across three applications: mobile crowdsensing, content delivery, and channel allocation. The results from these experiments demonstrate our model's superior efficacy. Exploring the removal of dependency p^* for independently probabilistically triggered arms can be an interesting direction for our future research.

References

[1] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1746–1778, 2016.

[2] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 915–929, 2018.

[3] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao, "Automatically characterizing places with opportunistic crowdsensing using smartphones," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 481–490.

[4] J. Zuo, X. Zhang, and C. Joe-Wong, "Observe before play: Multi-armed bandit with pre-observations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 7023–7030.

[5] A. Krause and C. Guestrin, "Near-optimal observation selection using submodular functions," in *AAAI*, vol. 7, 2007, pp. 1650–1654.

[6] L. Chen, J. Xu, S. Ren, and P. Zhou, "Spatio–temporal edge service placement: A bandit learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8388–8401, 2018.

[7] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*. PMLR, 2013, pp. 151–159.

[8] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits." in *AISTATS*, 2015.

[9] S. Li, B. Wang, S. Zhang, and W. Chen, "Contextual combinatorial cascading bandits," in *International conference on machine learning*. PMLR, 2016, pp. 1245–1253.

[10] X. Liu, J. Zuo, S. Wang, C. Joe-Wong, J. Lui, and W. Chen, "Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 904–14 916, 2022.

[11] L. Chen, J. Xu, and Z. Lu, "Contextual combinatorial multi-armed bandits with volatile arms and submodular reward," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] X. Liu, J. Zuo, H. Xie, C. Joe-Wong, and J. C. Lui, "Variance-adaptive algorithm for probabilistic maximum coverage bandits with general feedback," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[13] Z. Bnaya, R. Puzis, R. Stern, and A. Felner, "Volatile multi-armed bandits for guaranteed targeted social crawling," in *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[14] O. Marchal and J. Arbel, "On the sub-gaussianity of the beta and dirichlet distributions," *Electronic Communications in Probability*, vol. 22, pp. 1–14, 2017.

[15] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," in *Advances in Neural Information Processing Systems*, 2017, pp. 1161–1171.

[16] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[17] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[18] R. Combes, M. S. Talebi Mazraeh Shahi, A. Proutiere *et al.*, "Combinatorial bandits revisited," *Advances in neural information processing systems*, vol. 28, 2015.

[19] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[20] J. Honorio and T. Jaakkola, "Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees," in *Artificial Intelligence and Statistics*. PMLR, 2014, pp. 384–392.

[21] N. Merlis and S. Mannor, "Tight lower bounds for combinatorial multi-armed bandits," in *Conference on Learning Theory*. PMLR, 2020, pp. 2830–2857.

[22] ——, "Batch-size independent regret bounds for the combinatorial multi-armed bandit problem," in *Conference on Learning Theory*. PMLR, 2019, pp. 2465–2489.

[23] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," *Advances in neural information processing systems*, vol. 25, 2012.

[24] S. Vakili and Q. Zhao, "Risk-averse multi-armed bandit problems under mean-variance measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 6, pp. 1093–1111, 2016.

[25] Q. Shao, J. Ye, and J. C. Lui, "Risk-aware multi-agent multi-armed bandits," in *Proceedings of the Twenty-fifth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2024, pp. 61–70.

[26] P. Yang, N. Zhang, S. Zhang, K. Yang, L. Yu, and X. Shen, "Identifying the most valuable workers in fog-assisted spatial crowdsourcing," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1193–1203, 2017.

[27] X. Wang, J. Ye, and J. C. Lui, "Online learning aided decentralized multi-user task offloading for mobile edge computing," *IEEE Transactions on Mobile Computing*, vol. 23, no. 4, pp. 3328–3342, 2023.

[28] X. Dai, Z. Zhang, P. Yang, Y. Xu, X. Liu, and J. C. Lui, "Axiomvision: Accuracy-guaranteed adaptive visual model selection for perspective-aware video analytics," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7229–7238.

[29] S. A. Bitaghsir, A. Dadlani, M. Borhani, and A. Khonsari, "Multi-armed bandit learning for cache content placement in vehicular social networks," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2321–2324, 2019.

[30] X. Dai, Z. Wang, J. Xie, X. Liu, and J. C. Lui, "Conversational recommendation with online learning and clustering on misspecified users," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 7825–7838, 2024.

[31] K. Cai, X. Liu, Y.-Z. J. Chen, and J. C. Lui, "Learning with guarantee via constrained multi-armed bandit: Theory and network applications," *IEEE Transactions on Mobile Computing*, vol. 22, no. 9, pp. 5346–5358, 2022.

[32] O. Avner and S. Mannor, "Concurrent bandits and cognitive radio networks," in *Proc. of Springer ECML-PKDD*, 2014, pp. 66–81.

[33] X. Dai, Z. Wang, J. Ye, and J. C. Lui, "Quantifying the merits of network-assist online learning in optimizing network protocols," in *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*. IEEE, 2024, pp. 1–10.

[34] K. Cai, X. Liu, Y.-Z. J. Chen, and J. C. Lui, "An online learning approach to network application optimization with guarantee," in *Proc. of IEEE INFOCOM*, 2018, pp. 2006–2014.

[35] S. Wang, "Multichannel dqn channel model," https://github.com/ANRGUSC/MultichannelDQN-channelModel, 2018.

[36] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[37] S. Bubeck, N. Cesa-Bianchi *et al.*, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[38] A. Slivkins *et al.*, "Introduction to multi-armed bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[39] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[40] X. Dai, Z. Wang, J. Xie, T. Yu, and J. C. Lui, "Online learning and detecting corrupted users for conversational recommendation systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 8939–8953, 2024.

[41] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[42] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[43] X. Liu, J. Zuo, X. Chen, W. Chen, and J. C. Lui, "Multi-layered network exploration via random walks: From offline optimization to online learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7057–7066.

[44] M. Hefeeda and H. Ahmadi, "A probabilistic coverage protocol for wireless sensor networks," in *2007 IEEE International Conference on Network Protocols*. IEEE, 2007, pp. 41–50.

[45] J. Zuo, X. Liu, C. Joe-Wong, J. C. Lui, and W. Chen, "Online competitive influence maximization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 11 472–11 502.

[46] D. Zhou, Q. Gu, and C. Szepesvari, "Nearly minimax optimal reinforcement learning for linear mixture markov decision processes," in *Conference on Learning Theory*. PMLR, 2021, pp. 4532–4576.

[47] Z. Zhang, J. Yang, X. Ji, and S. S. Du, "Improved variance-aware confidence sets for linear bandits and linear mixture mdp," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4342–4355, 2021.

[48] D. Vial, S. Sanghavi, S. Shakkottai, and R. Srikant, "Minimax regret for cascading bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29 126–29 138, 2022.

[49] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 5, pp. 1466–1478, 2012.

[50] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvári, "Combinatorial cascading bandits," in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 1450–1458.

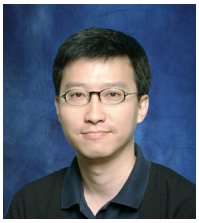
Xiangxiang Dai (Student Member, IEEE) is a PhD student at the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2023, advised by Prof. John C. S. Lui. He received his B.E. degree from the School of Electronic Information and Communications, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2023. His research interests include online learning theory and its algorithm design for various applications, such as web recommendation systems, multimedia platforms, and computer networks.

Xutong Liu (Member, IEEE) received the bachelor's degree from the School of Computer Science and Technology, University of Science and Technology of China, in 2017, and the PhD degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2022. He is now a postdoctoral researcher in the Department of Electrical and Computer Engineering at Carnegie Mellon University. Previously, he was a visiting postdoc at University of Massachusetts Amherst, and a postdoctoral fellow at the Chinese University of Hong Kong. His research interests include reinforcement learning, online learning, network science, combinatorial optimization, and stochastic modeling.

Jinhang Zuo (Member, IEEE) is an assistant professor in the Department of Computer Science at City University of Hong Kong. Prior to this, he was a joint postdoc at California Institute of Technology and University of Massachusetts Amherst. He received his Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University in 2022. Before that, he completed his B.E. in Communication Engineering at Nanjing University of Posts and Telecommunications in 2017. His main research interests include online learning, networked systems, and resource allocation. He was a recipient of the Center for Data Science (CDS) Postdoctoral Fellowship from UMass Amherst, ACM SIGMETRICS 2022 Best Poster Award, and Carnegie Institute of Technology Dean's Fellowship.

Hong Xie (Member, IEEE) is currently a research professor at School of Computer Science and Technology, University of Science and Technology of China (USTC). He received Ph.D. degree in the Department of Computer Science and Engineering at The Chinese University of Hong Kong (CUHK) in 2015. He received his B.Eng. degree from the School of Computer Science and Technology at USTC in 2010. Hong Xie was a postdoctoral research fellow at the Department of Computing Science and Engineering, CUHK hosted by Prof. John C.S. Lui, and a postdoctoral research fellow at the School of Computing, National University of Singapore. He was also a faculty member at Chongqing University. His research interests lie in bandits and reinforcement learning, as well as LLMs, with a focus on in-context learning and reasoning. He is a member of CCF, a member of IEEE, and a member of ACM.

Carlee Joe-Wong (Senior Member, IEEE) received the A.B. degree (magna cum laude) in Mathematics, and M.A. and Ph.D. degrees in Applied and Computational Mathematics, from Princeton University in 2011, 2013, and 2016, respectively. From 2013 to 2014, she was the Director of Advanced Research at DataMi, a startup she co-founded from her research on mobile data pricing. She is the Robert E. Doherty Associate Professor of Electrical and Computer Engineering at Carnegie Mellon University. Her work has received best paper and poster awards at several conferences, including IEEE INFOCOM, ACM/IEEE IPSN, ACM SIGMETRICS, and IEEE ICDCS. She received the NSF CAREER award in 2018, the Army Young Investigator award in 2019, and the Department of Energy Early Career Research Program Award in 2024. Her research interests are optimizing networked systems, including applications of machine learning and pricing to cloud computing, and mobile/wireless networks.



John C.S. Lui (Fellow, IEEE/ACM) is currently the Choh-Ming Li Chair Professor in the Department of Computer Science & Engineering (CSE) at The Chinese University of Hong Kong (CUHK). He received his Ph.D. in Computer Science from UCLA. After his graduation, he joined the IBM Laboratory and participated in research and development projects on file systems and parallel I/O architectures. He later joined the CSE Department at CUHK. His current research interests are in online learning algorithms and applications (e.g., multi-armed bandits, reinforcement learning), machine learning on network sciences and networking systems, large-scale data analytics, network/system security, network economics, large-scale storage systems, and performance evaluation theory. He has served at the IEEE Fellow Review Committees. He has served at the IEEE Fellow Review Committees. He is an elected member of the IFIP WG 7.3, Fellow of ACM, Fellow of IEEE, Senior Research Fellow of the Croucher Foundation, Fellow of the Hong Kong Academy of Engineering Sciences (HKAES).

APPENDIX

A. Fact and Definition

We introduce the following tail bounds for our analysis.

Lemma 2 (Empirical Bernstein Inequality [42]). *Let $(X_i)_{i\in[n]}$ be n i.i.d random variables with bounded support $[0,1]$ and mean $\mathbb{E}[X_i]=\mu$. Let $\hat{X}_n$ and $\hat{V}_n$ be the empirical mean and empirical variance of $(X_i)_{i\in[n]}$. Then for any $n\in\mathbb{N}$ and $y>0$, it holds that* $\Pr\left[|\hat{X}_n-\mu|\geq\sqrt{\frac{2\hat{V}_n y}{n}}+\frac{3y}{n}\right]\leq 3e^{-y}$.

Next, we use the following Bernstein Inequality to bound the difference between the empirical variance and the true variance under our variance-aware framework.

Lemma 3 (Bernstein Inequality [17]). *Let $(X_i)_{i\in[n]}$ be n independent random variables in $[0,1]$ with mean $\mathbb{E}[X_i]=\mu$ and variance $Var[X_i]=V$. Then with probability $1-\delta$: $\frac{1}{n}\sum_{i\in[n]}X_i\leq\mu+\frac{2\log 1/\delta}{3n}+\sqrt{\frac{2V\log 1/\delta}{n}}$.*

We define the following event for arm-level concentration.

Definition 3 (Nice Sampling). *We say that the sampling is nice at the beginning of round t if: (1) for every base arm $i\in[m]$, $|\hat{\mu}_{t-1,i}-\mu_i|\leq\rho_{t,i}$, where $\rho_{t,i}=\sqrt{\frac{6\hat{V}_{t-1,i}\log t}{T_{t-1,i}}}+\frac{9\log t}{T_{t-1,i}}$; (2) for every base arm $i\in[m]$, $\hat{V}_{t-1,i}\leq 2\mu_i(1-\mu_i)+\frac{3.5\log t}{T_{t-1,i}}$. We denote such event as $\mathcal{N}_t^s$.*

The following lemma bounds the probability that $\neg\mathcal{N}_t^s$.

Lemma 4. *For each round t, $\Pr[\neg\mathcal{N}_t^s]\leq 4mt^{-2}$.*

Proof. Let $\mathcal{N}_t^{s,1},\mathcal{N}_t^{s,2}$ be the event (1) and event (2) in Definition 3. For $\Pr[\neg\mathcal{N}_t^{s,1}]$, we can bound it using Lemma 2 by setting $y=3\log t$ We then bound the probability that the second event $\neg\mathcal{N}_t^{s,2}$ using the similar proof of Eq. (7) in [22]. Fix $T_{t-1,i}=\tau$ and consider $(Y_i^1,...,Y_i^\tau)$, where $Y_i^k=(X_i^k-\mu_i)^2\in[0,1]$ and X_i^k is the random outcome of the k-th i.i.d trial. In this case, one can verify that $\hat{V}_{t-1,i}\leq\frac{1}{\tau}\sum_{k=1}^{\tau}Y_i^k$; $\mathbb{E}[Y_i^k]\leq(1-\mu_i)\mu_i$; and $\text{Var}[Y_i]\leq\mathbb{E}[Y_i^k]\leq(1-\mu_i)\mu_i$. By Lemma 3, it holds with probability at least $1-t^{-3}$ that $\hat{V}_{t-1,i}\leq\mu_i(1-\mu_i)+\frac{2\log t}{\tau}+\sqrt{\frac{6(1-\mu_i)\mu_i\log t}{\tau}}\leq\mu_i(1-\mu_i)+\frac{2\log t}{\tau}+\mu_i(1-\mu_i)+\frac{3\log t}{2\tau}=2\mu_i(1-\mu_i)+\frac{3.5\log t}{\tau}$ Now by applying union bound over $i\in[m]$ and $\tau\in[t]$, $\mathcal{N}_t^{s,1}$ and $\mathcal{N}_t^{s,2}$, we have $\Pr[\neg\mathcal{N}_t^s]\leq 4mt^{-2}$. ■

The next lemma bounds each arm's actual over-estimation, and the $\sqrt{(1-\mu_i)\mu_i}$ is the key term to cancel the denominator in $x_{v,i}$ to give improved bounds.[3]

Lemma 5 (Arm-level Over-estimation). *For every base arm $i\in[m]$ and every time $t\in[T]$, it holds with probability at least $1-4mt^{-3}$ that $\mu_i\leq\bar{\mu}_{t,i}\leq\min\left\{\mu_i+4\sqrt{3}\sqrt{\frac{\mu_i(1-\mu_i)\log t}{T_{t-1,i}}}+\frac{28\log t}{T_{t-1,i}},1\right\}$.*

Proof. Under event $N_t^{s,1}$, we have $|\mu_i-\hat{\mu}_{t,i}|\leq\rho_{t,i}$ by Lemma 4, hence the first and the second inequality in Lemma 5 holds. For the last inequality, it holds under event $N_t^{s,2}$ to replace $\hat{V}_{t-1,i}$ with $2\mu_i(1-\mu_i)+\frac{3.5\log t}{T_{t-1,i}}$). Since $\mathcal{N}_t^s=\mathcal{N}_t^{s,1}\bigcap\mathcal{N}_t^{s,2}$ and by Lemma 4, Lemma 5 holds with probability at least $1-4mt^{-2}$. ■

[3]For a bounded random variable $X\in[0,1]$ with mean μ_i and variance $V_i\leq(1-\mu_i)\mu_i$, where maximum variance occurs if X is Bernoulli.

B. Proof of Property 1

For cascading feedback, without loss of generality, let the mean of action in the group $j\in V$ be $\{\mu_{1,j},...,\mu_{K,j}\}$, then the reward function is $r(S;\boldsymbol{\mu})=\sum_{j\in V}1-\prod_{i=1}^{K}(1-\mu_{i,j})$ and the observation probability is $p_{i,j}^{D,S}=\prod_{\ell=1}^{i-1}(1-\mu_{\ell,j})$. To simplify the notation, let $[K]$ represent the set of all nodes that can be maximally covered, with a slight abuse of the notation. Let $\bar{\boldsymbol{\mu}}=(\bar{\mu}_{i,j})_{i\in[K],j\in V}$ and $\boldsymbol{\mu}=(\mu_{i,j})_{i\in[K],j\in V}$, where $\bar{\boldsymbol{\mu}}=\boldsymbol{\mu}+\boldsymbol{\zeta}+\boldsymbol{\eta}$ with $\bar{\boldsymbol{\mu}},\boldsymbol{\mu}\in(0,1)^{[K]\times V},\boldsymbol{\zeta},\boldsymbol{\eta}\in[0,1]^{[K]\times V}$. Now we can derive $r(S;\bar{\boldsymbol{\mu}})-r(S;\boldsymbol{\mu})$ equals

$$\sum_{j\in V,i\in[K]}(\bar{\mu}_{i,j}-\mu_{i,j})\prod_{\ell=1}^{i-1}(1-\mu_{\ell,j})\prod_{\ell=i+1}^{K}(1-\bar{\mu}_{\ell,j}) \tag{11}$$

$$\leq\sum_{j\in V}\sum_{i\in[K]}(\zeta_{i,j})(\prod_{\ell=1}^{i-1}(1-\mu_{\ell,j})\prod_{\ell=i+1}^{K}(1-\mu_{\ell,j}))$$
$$+\sum_{j\in V}\sum_{i\in[K]}(\eta_{i,j})\prod_{\ell=1}^{i-1}(1-\mu_{\ell,j}) \tag{12}$$

$$\leq\sqrt{\sum_{j\in V,i\in[K]}\frac{\zeta_{i,j}^2(p_{i,j}^{D,S})^2}{(1-\mu_{i,j})\mu_{i,j}}}\cdot\sqrt{\sum_{j\in V,i\in[K]}\prod_{\ell=i+1}^{K}(1-\mu_{\ell,j})\mu_{i,j}}$$
$$+\sum_{j\in V}\sum_{i\in[K]}\eta_{i,j}p_{i,j}^{D,S}, \tag{13}$$

where the first inequality is by definition of $\zeta_{i,j},\eta_{i,j}$ and $\bar{\mu}_{i,j}\geq\mu_{i,j}$, the second inequality is by Cauchy-Schwarz inequality and definition of $p_{i,j}^{D,S}$, concluding the lemma by $\sqrt{\sum_{j\in V}(1-\prod_{\ell=1}^{K}(1-\mu_{\ell,j}))}\leq\sqrt{|V|}$.

For probabilistic feedback, let the effective base arms $\boldsymbol{\mu}=(\boldsymbol{x},\boldsymbol{y})\in(0,1)^{(K|V|+|V|)},\bar{\boldsymbol{\mu}}=(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})\in(0,1)^{(K|V|+|V|)}$, where $\bar{\boldsymbol{x}}=\boldsymbol{\zeta}_x+\boldsymbol{\eta}_x+\boldsymbol{x},\bar{\boldsymbol{y}}=\boldsymbol{\zeta}_y+\boldsymbol{\eta}_y+\boldsymbol{y}$, for $\boldsymbol{\zeta},\boldsymbol{\eta}\in[0,1]^{(K|V|+|V|)}$. For the target node $j\in V$, the per-target reward function $r_j(S;\boldsymbol{x},\boldsymbol{y})=y_j(1-\prod_{i\in[K]}(1-x_{i,j}))$. Denote $\bar{p}_j^{D,S}=1-\prod_{i\in[K]}(1-\bar{x}_{i,j})$. Now we can derive $r(S;\bar{\boldsymbol{\mu}})-r(S;\boldsymbol{\mu})=\sum_{j\in V}r_j(S;\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})-r_j(S;\boldsymbol{x},\boldsymbol{y})=\sum_{j\in V}\bar{y}_j\left(\prod_{i\in[K]}(1-x_{i,j})-\prod_{i\in[K]}(1-\bar{x}_{i,j}))\right)+\sum_{j\in V}(\bar{y}_j-y_j)p_j^{D,S}$. For the first summation, we follow exactly the derivation of the cascading feedback, we have RHS $\leq\sqrt{\sum_{j\in V,i\in[K]}(\frac{\zeta_{x,i,j}^2}{(1-x_{i,j})x_{i,j}})+\sum_{j\in V}\frac{\zeta_{y,j}^2(p_j^{D,S})^2}{(1-y_j)y_j}}\cdot\sqrt{\sum_{j\in V}\bar{y}_j^2+(1-y_j)y_j}+(\sum_{j\in V,i\in[K]}|\eta_{x,i,j}|+\sum_{j\in V}|\eta_{y,j}|\,p_j^{D,S})$ and replacing $\sqrt{\sum_{j\in V}\bar{y}_j^2+(1-y_j)y_j}\leq\sqrt{1.25|V|}$ concludes the proof.

For semi-bandit feedback, it is easy to follow the probabilistic feedback but set $p_{i,j}^{D,S}=1$ if $i\in S$ and 0, otherwise.

C. Proof of Lemma 1

Proof. Under event $\mathcal{N}_t^s$, by Lemma 5, it is easy to check that $\bar{\mu}_{t,i}\leq\min\{\mu_{t-1,i}+4\sqrt{3}\sqrt{\frac{\mu_i(1-\mu_i)\log t}{T_{t-1,i}}}+\frac{28\log t}{T_{t-1,i}},1\}\leq$

$\mu_{t-1,i} + 4\sqrt{3}\sqrt{\mu_i(1-\mu_i)(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})} + 28(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})$. Therefore, it holds that

$$\alpha r(S_t^*; \boldsymbol{\mu}) \leq \alpha r(S_t^*; \bar{\boldsymbol{\mu}}_t) \leq r(S_t; \bar{\boldsymbol{\mu}}_t) \tag{14}$$

$$\leq r(S_t; \boldsymbol{\mu}) + e_{t,1}(S_t) + e_{t,2}(S_t), \tag{15}$$

where the first inequality is because the reward function is monotone and second inequality is due to the computation oracle, the third inequality is because of the inequality above and Property 1 by plugging in $\zeta_i = 4\sqrt{3}\sqrt{\mu_i(1-\mu_i)(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})}$ and $\eta_i = 28(\frac{\log t}{T_{t-1,i}} \wedge \frac{1}{28})$. So $Reg(T, \mathcal{N}_t^s) \leq Reg(T, E_{t,1}) + Reg(T, E_{t,2})$. Now for $Reg(T, \neg\mathcal{N}_t^s)$, by Lemma 4 it holds that $Reg(T, \neg\mathcal{N}_t^s) \leq \sum_{t=1}^{T} \Pr[\neg\mathcal{N}_t^s] \leq \sum_{t=1}^{T} 4mt^{-2} \leq \frac{2\pi^2}{3} m\Delta_{\max}$, which concludes the lemma. ■

D. Proof of Property 2

Unlike Property 1, Property 2 is unidirectional, allowing both $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ to take negative values. This is because, in Algorithm 2, we do not need to construct the UCB for each base arm but rather base it directly on the action, i.e., the super arm. In this scenario, the optimistic reward we designed does not require monotonicity, thereby permitting $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ to be negative.

For cascading feedback, use the same notations as in Appendix B, but with $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^{[K]\times V}$. We can derive $|r(S; \bar{\boldsymbol{\mu}}) - r(S; \boldsymbol{\mu})|$ equals

$$\sum_{j\in V, i\in[K]} |(\bar{\mu}_{i,j} - \mu_{i,j})|((1-\mu_{1,j})...(1-\mu_{i-1,j})\cdot (1-\bar{\mu}_{i+1,j})...(1-\bar{\mu}_{K,j}))$$

$$\leq \sum_{j\in V, i\in[K]} |\zeta_{i,j}|((1-\mu_{1,j})...(1-\mu_{i-1,j}) \quad (1-\mu_{i+1,j})...\cdot (1-\mu_{K,j})) + \sum_{j\in V, i\in[K]} |\eta_{i,j}|((1-\mu_{1,j})...(1-\mu_{i-1,j})) \tag{16, 17}$$

$$\leq \sum_{j\in V, i\in[K]} \frac{|\zeta_{i,j}| p_{i,j}^{D,S}}{\sqrt{(1-\mu_{i,j})\mu_{i,j}}} \sqrt{(1-\mu_{i+1})...(1-\mu_K)\mu_{i,j}} + \sum_{j\in V}\sum_{i\in[K]} |\eta_{i,j}| p_{i,j}^{D,S} \tag{18}$$

$$\leq \sqrt{\sum_{j\in V, i\in[K]} \frac{\zeta_{i,j}^2 p_{i,j}^{D,S}}{(1-\mu_{i,j})\mu_{i,j}}} \sqrt{\sum_{j\in V} 1-(1-\mu_{1,j})...(1-\mu_{K,j})} + \sum_{j\in V, i\in[K]} |\eta_{i,j}| p_{i,j}^{D,S} \tag{19}$$

$$\leq \sqrt{|V| \sum_{j\in V, i\in[K]} \frac{\zeta_{i,j}^2}{(1-\mu_{i,j})\mu_{i,j}}} + \sum_{j\in V, i\in[K]} |\eta_{i,j}|, \tag{20}$$

where Eq. (17) is due to $\bar{\mu}_{i,j}, \mu_{i,j} \in [0,1]$ for any $i \in [K], j \in V$, and the definition of $\boldsymbol{\zeta}, \boldsymbol{\eta}$, Eq. (18) is due to the definition of $p_{i,j}^{D,S}$ and we multiply the first term by $\sqrt{\mu_{i,j}}$ but divide it by $\sqrt{(1-\mu_{i,j})\mu_{i,j}}$, Eq. (19) is due to the Cauchy–Schwarz inequality on the first term with some math calculation, and Eq. (20) is due to $\sqrt{\sum_{j\in V} 1-(1-\mu_{1,j})...(1-\mu_{K,j})} \leq \sqrt{|V|}$. In the cascading feedback, the trigger arm set $\tilde{S}$ of action S equals $[K] \times V$, which concludes the proof.

For probabilistic feedback, similarly, let effective base arms $\boldsymbol{\mu} = (\boldsymbol{x}, \boldsymbol{y}) \in (0,1)^{(K|V|+|V|)}, \bar{\boldsymbol{\mu}} = (\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \in (0,1)^{(K|V|+|V|)}$, where $\bar{\boldsymbol{x}} = \boldsymbol{\zeta}_x + \boldsymbol{\eta}_x + \boldsymbol{x}, \bar{\boldsymbol{y}} = \boldsymbol{\zeta}_y + \boldsymbol{\eta}_y + \boldsymbol{y}$, for $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^{(K|V|+|V|)}$. Define $\tilde{V}$ as the set of nodes that can be triggered by V, the set of nodes to be covered under the PMC model, to concretize the trigger action $\tilde{S}$. For the target node $j \in V$, the per-target reward function $r_j(S; \boldsymbol{x}, \boldsymbol{y}) = y_j(1 - \prod_{i\in[K]}(1 - x_{i,j}))$. Denote $\bar{p}_j^{D,S} = 1 - \prod_{i\in[K]}(1-\bar{x}_{i,j})$. We can derive $|r(S; \bar{\boldsymbol{\mu}}) - r(S; \boldsymbol{\mu})| = \left|\sum_{j\in V} r_j(S; \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) - r_j(S; \boldsymbol{x}, \boldsymbol{y})\right| = \left|\sum_{j\in V} \bar{y}_j \bar{p}_j^{D,S} - \bar{y}_j p_j^{D,S} + \bar{y}_j p_j^{D,S} - y_j p_j^{D,S}\right|$. Since $\bar{x}_{i,j} \in [0,1]$ for any i, j, we have:

$$|r(S; \bar{\boldsymbol{\mu}}) - r(S; \boldsymbol{\mu})| \leq \sum_{j\in V}\sum_{i\in[K]} (|\zeta_{x,i,j}| + |\eta_{x,i,j}|)\left((1-x_{1,j})...(1-x_{i-1,j})\right)\bar{y}_j + \sum_{j\in V}(|\zeta_{y,j}| + |\eta_{y,j}|)p_j^{D,S}, \tag{21}$$

Multiplying the first term by $\sqrt{x_{i,j}}$ but divide it by $\sqrt{x_{i,j}(1-x_{i,j})}$ and multiplying the second term by $\sqrt{y_j}$ but divide it by $\sqrt{y_j(1-y_j)}$, we have $Eq.\ (21) \leq \sum_{j\in V, i\in[K]} (\frac{|\zeta_{x,i,j}|}{\sqrt{(1-x_{i,j})x_{i,j}}}) \sqrt{\left((1-x_{1,j})...(1-x_{i-1,j})\right) x_{i,j}} \bar{y}_j + \sum_{j\in V} \left(\frac{|\zeta_{y,j}| p_j^{D,S}}{\sqrt{(1-y_j)y_j}}\right) \sqrt{(1-y_j)y_j} + \sum_{j\in V}\sum_{i\in[K]} |\eta_{x,i,j}| + \sum_{j\in V} |\eta_{y,j}| p_j^{D,S}$. Applying the Cauchy-Schwarz inequality with $x_{i,j} \in [0,1]$ for any i, j and $y_j, \bar{y}_j \in [0,1]$ for any j, we have $RHS \leq \sqrt{\sum_{j\in V}\sum_{i\in[K]} (\frac{\zeta_{x,i,j}^2}{(1-x_{i,j})x_{i,j}}) + \sum_{j\in V} \frac{\zeta_{y,j}^2 (p_j^{D,S})^2}{(1-y_j)y_j}} \cdot \sqrt{1.25|V|} + \left(\sum_{j\in V}\sum_{i\in[K]} |\eta_{x,i,j}| + \sum_{j\in V} |\eta_{y,j}| p_j^{D,S}\right)$. Finally, we obtain $RHS \leq \left(\sum_{j\in V}\sum_{i\in[K]} |\eta_{x,i,j}| + \sum_{j\in\tilde{V}} |\eta_{y,j}|\right) + \sqrt{1.25|V|}\sqrt{\sum_{j\in V}\sum_{i\in[K]} (\frac{\zeta_{x,i,j}^2}{(1-x_{i,j})x_{i,j}}) + \sum_{j\in\tilde{V}} \frac{\zeta_{y,j}^2}{(1-y_j)y_j}}$. Under probabilistic feedback, the trigger arm set $\tilde{S}$ for action S is given by $\tilde{S} = ([K] \times V) \cup \tilde{V}$, thus concluding the proof.

For semi-bandit feedback, it is easy to follow the probabilistic feedback but set $p_{i,j}^{D,S} = 1$ if $i \in S$ and 0, otherwise.

E. Upper bound of $Reg(T, E_{t,1})$

Denote $\tilde{\Delta}_{t,S_t} = \frac{\Delta_{t,S_t}}{p_i^{D,S_t}}$ for action S_t at $t \in \mathcal{T}$, and $\tilde{\Delta}_i^{\min} = \min_{S\in\mathcal{S}_t, \mathcal{S}_t\in\mathcal{S}, t\in\mathcal{T}: p_i^{D,S}>0, \Delta_{t,S}>0} \Delta_{t,S}/p_i^{D,S}$. Regarding Δ_{t,S_t} on the selected action S_t at round t, we have:

$$\Delta_{t,S_t} \leq \sum_{i\in[m]} \frac{4(4\sqrt{3})^2 |V| (p_i^{D,S_t})^2 \frac{\log t}{T_{t-1,i}}}{\Delta_{t,S_t}} \tag{22}$$

$$= -\Delta_{t,S_t} + 2\sum_{i\in[m]} \frac{192|V| (p_i^{D,S_t})^2 \frac{\log t}{T_{t-1,i}}}{\Delta_{t,S_t}} \tag{23}$$

$$\leq \sum_{i\in[m]} p_i^{D,S_t} \left(\frac{384|V|\frac{\log t}{T_{t-1,i}}}{\tilde{\Delta}_{t,S_t}} - \frac{\tilde{\Delta}_{t,S_t}}{K}\right). \tag{24}$$

Defining a regret allocation function

$$\kappa_{i,T}(\ell)=\begin{cases}\frac{48|V|}{\tilde{\Delta}_i^{\min}}, & \text{if } \ell=0,\\ 2\sqrt{\frac{192|V|\log T}{\ell}}, & \text{if } 1\le\ell\le\frac{192|V|\log T}{(\tilde{\Delta}_i^{\min})^2},\\ \frac{384|V|\log T}{\tilde{\Delta}_i^{\min}}\frac{1}{\ell}, & \text{if } \frac{192|V|\log T}{(\tilde{\Delta}_i^{\min})^2}<\ell\le\frac{384|V|K\log T}{(\tilde{\Delta}_i^{\min})^2},\\ 0, & \text{if } \ell>\frac{384|V|K\log T}{(\tilde{\Delta}_i^{\min})^2},\end{cases} \tag{25}$$

via $\Delta_{t,S_t}=\mathbb{E}_t[\tilde{\Delta}_{t,S_t}]$, we then have:

$$\Delta_{t,S_t}\le\mathbb{E}_t\left[\sum_{i\in[m]}p_i^{D,S_t}\left(\frac{384|V|\frac{\log t}{T_{t-1,i}}}{\tilde{\Delta}_{t,S_t}}-\frac{\tilde{\Delta}_{t,S_t}}{K}\right)\right] \tag{26}$$

$$=\mathbb{E}_t\left[\sum_{i\in\tau_t}\left(\frac{384|V|\frac{\log t}{T_{t-1,i}}}{\tilde{\Delta}_{t,S_t}}-\frac{\tilde{\Delta}_{t,S_t}}{K}\right)\right] \tag{27}$$

$$\le\mathbb{E}_t\left[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right]. \tag{28}$$

Now, we will demonstrate the validity of Eq. (28).

1) **When** $T_{t-1,i}>\frac{384|V|K\log T}{(\tilde{\Delta}_i^{\min})^2}$**,**

we can deduce that $(Eq.\ (28),i)\le\frac{384|V|\log T}{T_{t-1,i}\cdot\tilde{\Delta}_{t,S_t}}-\frac{\tilde{\Delta}_{t,S_t}}{K}<\frac{\tilde{\Delta}_i^{\min})^2}{K\tilde{\Delta}_{t,S_t}}-\frac{\tilde{\Delta}_{t,S_t}}{K}\le 0=\kappa_{i,T}(T_{t-1,i})$.

2) **When** $\frac{192|V|\log T}{(\tilde{\Delta}_i^{\min})^2}<T_{t-1,i}\le\frac{384|V|K\log T}{(\tilde{\Delta}_i^{\min})^2}$**,**

it follows that $(Eq.\ (28),i)\le\frac{384|V|\log T}{T_{t-1,i}\cdot\tilde{\Delta}_{t,S_t}}-\frac{\tilde{\Delta}_{t,S_t}}{K}<\frac{384|V|\log T}{T_{t-1,i}\cdot\tilde{\Delta}_{t,S_t}}\le\frac{384|V|\log T}{T_{t-1,i}\cdot\tilde{\Delta}_{\min}^{i,\lambda}}=\kappa_{i,T}(T_{t-1,i})$.

3) **When** $T_{t-1,i}\le\frac{192|V|\log T}{(\tilde{\Delta}_i^{\min})^2}$,

we analyze two scenarios as follows:

$$\begin{cases}1) & T_{t-1,i}\le\frac{192|V|\log T}{\tilde{\Delta}_{t,S_t}^2},\\ 2) & \frac{192|V|\log T}{\tilde{\Delta}_{t,S_t}^2}<T_{t-1,i}\le\frac{192|V|\log T}{(\tilde{\Delta}_{i,\lambda}^{\min})^2}.\end{cases}$$

In the first scenario, if for any $i\in\tau_t$, it is true that $T_{t-1,i}\le\frac{192|V|\log T}{\tilde{\Delta}_{t,S_t}^2}$, then it is established that $\sum_{q\in\tilde{S}_t}\kappa_{q,T}(T_{t-1,q})\ge\kappa_{i,T}(T_{t-1,i})=2\sqrt{\frac{192|V|\log T}{T_{t-1,i}}}\ge 2\sqrt{\tilde{\Delta}_{t,S_t}^2}\ge\Delta_{t,S_t}$, affirming the truth of Eq. (28) under any circumstance. Hence, this case does not require further consideration.

In the second scenario, when $\frac{192|V|\log T}{\tilde{\Delta}_{t,S_t}^2}<T_{t-1,i}$, it is deduced that

$$\begin{aligned}(Eq.\ (28),i)&\le\frac{384|V|\log T}{\tilde{\Delta}_{t,S_t}}\frac{1}{T_{t-1,i}}\\&=2\sqrt{\frac{192|V|\log T}{(\tilde{\Delta}_{t,S_t})^2}\frac{1}{T_{t-1,i}}}\sqrt{\frac{192|V|\log T}{T_{t-1,i}}}\\&\le 2\sqrt{\frac{\tilde{\Delta}_{t,S_t}}{\tilde{\Delta}_{t,S_t}}}\sqrt{\frac{192|V|\log T}{T_{t-1,i}}}\\&\le 2\sqrt{\frac{192|V|\log T}{T_{t-1,i}}}\\&=\kappa_{i,T}(T_{t-1,i}).\end{aligned}$$

4) **When** $\ell=0$,

we have $(Eq.\ (28),i)\le\frac{384|V|}{\tilde{\Delta}_{t,S_t}}\cdot\frac{1}{28}-\frac{\tilde{\Delta}_{t,S_t}}{K}\le\frac{48|V|}{\tilde{\Delta}_{t,S_t}}\le\frac{48|V|}{\tilde{\Delta}_{i,\lambda}^{\min}}=\kappa_{i,T}(T_{t-1,i})$.

Combining all above cases, we have $\Delta_{t,S_t}\le\mathbb{E}[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})]$, i.e., Eq. (28) holds.

Next, according to the tower rule and the fact that $T_{t-1,i}$ is increased by 1 if and only if $i\in\tau_t$, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t\in[T]}\mathbb{E}_t\left[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right]\right]\\&=\mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right]\\&=\mathbb{E}\left[\sum_{i\in[m]}\sum_{s=0}^{T_{T-1,i}}\kappa_{i,T}(s)\right].\end{aligned} \tag{29}$$

Applying Eq. (28) and the definition of regret allocation function $\kappa_{i,T}(\ell)$ in Eq. (25), the expected regret $Reg(T,E_{t,1})=\mathbb{E}\left[\sum_{t=1}^T\Delta_{t,S_t}\right]$ under event $E_{t,1}$ is bounded as:

$$\begin{aligned}Reg(T,E_{t,1})\le&\sum_{i\in[m]}\frac{48|V|}{\tilde{\Delta}_i^{\min}}+\sum_{i\in[m]}\frac{384|V|\log T}{\tilde{\Delta}_i^{\min}}(1+\log K)\\&+\sum_{i\in[m]}\frac{768|V|\log T}{\tilde{\Delta}_i^{\min}}.\end{aligned} \tag{30}$$

F. Upper bound of $Reg(T,E_{t,2})$

In the analysis of $Reg(T,E_{t,2})$, similar to that of $Reg(T,E_{t,1})$, under event $E_{t,2}$, we derive:

$$\begin{aligned}\Delta_{t,S_t}&\le\sum_{i\in\tilde{S}_t}56p_i^{D,S_t}\min\left\{1/28,\frac{\log T}{T_{t-1,i}}\right\}\\&\le-\Delta_{t,S_t}+2\sum_{i\in\tilde{S}_t}56p_i^{D,S_t}\min\left\{1/28,\frac{\log T}{T_{t-1,i}}\right\}\\&\le\sum_{i\in[m]}p_i^{D,S_t}\left(-\frac{\Delta_{t,S_t}}{K}+112\min\left\{1/28,\frac{\log T}{T_{t-1,i}}\right\}\right).\end{aligned}$$

By establishing the probability equivalence

$$\begin{aligned}&\mathbb{E}_t\left[\sum_{i\in[m]}p_i^{D,S_t}\left(-\frac{\Delta_{t,S_t}}{K}+112\min\left\{1/28,\frac{\log T}{T_{t-1,i}}\right\}\right)\right]\\&=\mathbb{E}_t\left[\sum_{i\in\tau_t}\left(-\frac{\Delta_{t,S_t}}{K}+112\min\left\{1/28,\frac{\log T}{T_{t-1,i}}\right\}\right)\right],\end{aligned}$$

it follows

$$\Delta_{t,S_t}=\mathbb{E}_t[\Delta_{t,S_t}]\le\mathbb{E}_t\left[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right], \tag{31}$$

where the regret allocation function $\kappa_{i,T}(\ell)$ is adjusted accordingly:

$$\kappa_{i,T}(\ell)=\begin{cases}\Delta_i^{\max}, & \text{if } 0\le\ell\le\frac{112\log T}{\Delta_i^{\max}},\\ \frac{112\log T}{\ell}, & \text{if } \frac{112\log T}{\Delta_i^{\max}}<\ell\le\frac{112K\log T}{\Delta_i^{\min}},\\ 0, & \text{if } \ell>\frac{112K\log T}{\Delta_i^{\min}}+1.\end{cases} \tag{32}$$

Now, we prove the reason why Eq. (31) holds.

1) **When** $T_{t-1,i} > \frac{112K\log T}{\Delta_i^{\min}}$**,**

it follows that $(Eq.\ (31), i) \leq 112\frac{\log T}{T_{t-1,i}} - \frac{\Delta_{t,S_t}}{K} < \frac{\Delta_i^{\min}}{K} - \frac{\Delta_{t,S_t}}{K} \leq 0 = \kappa_{i,T}(T_{t-1,i})$.

2) **When** $T_{t-1,i} \leq \frac{112K\log T}{\Delta_i^{\min}}$**,**

we deduce that $(Eq.\ (31), i) \leq 112\frac{\log T}{T_{t-1,i}} - \frac{\Delta_{t,S_t}}{K} < \frac{112\log T}{T_{t-1,i}} = \kappa_{i,j_i^{S_t},T}(N_{t-1,i,j_i^{S_t}})$.

3) **When** $T_{t-1,i} \leq \frac{192|V|\log T}{(\Delta_i^{\min})^2}$**,**

if for any $i \in \tilde{S}_t$ holds true, then it is established that $T_{t-1,i} \leq \frac{192|V|\log T}{(\Delta_i^{\min})^2}$, then we know $\sum_{q\in\tilde{S}_t}\kappa_{i,T}(T_{t-1,q}) \geq \kappa_{i,T}(T_{t-1,i}) = \Delta_i^{\max} \geq \Delta_{t,S_t}$, which makes Eq. (31) holds under all circumstances. This scenario thus does not require further consideration.

Combining all above cases, we have $\Delta_{t,S_t} \leq \mathbb{E}_t[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})]$.

Applying Eq. (31) and Eq. (32), we can derive the expected regret $Reg(T, E_{t,2})$ under event $E_{t,2}$ equals to:

$$
\begin{aligned}
Reg(T, E_{t,2}) &= \mathbb{E}\left[\sum_{t=1}^{T}\Delta_{t,S_t}\right] \qquad (33)\\
&\leq \mathbb{E}\left[\sum_{t\in[T]}\mathbb{E}_t\left[\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right]\right]\\
&= \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in\tau_t}\kappa_{i,T}(T_{t-1,i})\right]\\
&= \mathbb{E}\left[\sum_{i\in[m]}\sum_{s=0}^{T_{T-1,i}}\kappa_{i,T}(s)\right]\\
&\leq \sum_{i\in[m]}112\left(1+\log\left(\frac{K\Delta_i^{\max}}{\Delta_i^{\min}}\right)\right)\log T\\
&\quad + m\Delta_{\max}. \qquad (34)
\end{aligned}
$$

G. Proof of Theorem 2

For self-reliant arms, we also explored the feasibility of establishing a direct connection between the triggering probabilities and the expected random triggering events to reduce the upper bound of regret. However, this approach proved ineffective because the randomness introduced by triggering is coupled with the outcomes. Additionally, the confidence interval involves $T_{t-1,\tilde{S}}^{\min}$, which must be considered across all arms $\tilde{S}$ with triggers rather than S. Consequently, we adopted an alternative proof method, distinct from the analysis of Theorem 1, which is outlined as follows.

Let $\tilde{S}$ be the set arm arms that can be triggered by super arm S, i.e., $\tilde{S} := \{i \in [m] : p_i^{D,S} > 0\}$. A random variable with mean $\mu = \mathbb{E}[X]$ is sub-Exponential with parameter (ν^2, b) if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\nu^2}{2}}$ for any $|\lambda| < \frac{1}{b}$ [19]. We denote this by $X \in SE(\nu^2, b)$. Next, we give the following properties for sub-Exponential random variables.

Lemma 6 (Independent Sub-Exponential Random Variables Composition [19])**.** *Consider independent sub-Exponential random variables $Y_1, \ldots, Y_n$, where each Y_i is in the class $SE(\nu_i^2, b_i)$ with expected value $\mathbb{E}[Y_i] = \mu_i$. Then, the sum of their centered versions, $\sum_{i=1}^n(Y_i - \mu_i)$, belongs to the sub-Exponential class SE $\left(\sum_{i=1}^n\nu_i^2, \max_i b_i\right)$.*

Let $\delta_{t,i} = \hat{\mu}_{t-1,i} - \mu_i$ and $\delta_{t,i}$ is a sub-Gaussian random variable SE($\frac{C_1(1-\mu_i)\mu_i}{T_{t-1,i}}$) with mean 0 [19]. Let $u_{t,i} = \frac{\delta_{t,i}}{\sqrt{(1-\mu_i)\mu_i}}$, then $u_{t,i}$ is also sub-Gaussian SE($\frac{C_1}{T_{t-1,i}}$). By Property 2, we have $|r(S;\hat{\boldsymbol{\mu}}_{t-1}) - r(S;\boldsymbol{\mu})| \leq \sqrt{|V|\sum_{i\in\tilde{S}}u_{t,i}^2}$.

Fix a super arm S, we will focus on random variable $Y_{t,S} \triangleq \sum_{i\in\tilde{S}}u_{t,i}^2$. Since the square of sub-Gaussian random variable is sub-Exponential [20], we have $Y_{t,S} \in$ SE($48\sum_{i\in\tilde{S}}\frac{32}{T_{t-1,i}^2}, \frac{4C_1}{T_{t-1,\tilde{S}}^{\min}}$) from Lemma 6. For the mean of $Y_{t,S}$, we can also show that $\mathbb{E}[Y_{t,S}] \leq \sum_{i\in\tilde{S}}\frac{C_1}{T_{t-1,i}}$, where the last inequality is because the variance of any sub-Gaussian random variable $X \in$ SE(σ^2) is smaller than σ^2 [19]. According to the tail bounds for sub-Exponential random variables in [19], we can give the confidence interval:

$$
\begin{aligned}
Y_{t,S} \leq & C_1\sum_{i\in\tilde{S}}\frac{1}{T_{t-1,i}} + \\
& \max\{\sqrt{6448\log(\frac{2}{\delta})\sum_{i\in\tilde{S}}\frac{1}{T_{t-1,i}^2}}, 8C_1\log(\frac{2}{\delta})\frac{1}{T_{t-1,\tilde{S}}^{\min}}\}. \qquad (35)
\end{aligned}
$$

Then, with probability at least $1-\delta$, it holds that $|r(S;\hat{\boldsymbol{\mu}}_{t-1}) - r(S;\boldsymbol{\mu})| \leq \rho_t(S)$, where $\rho_t(S)$ is defined in Eq. (4). If S_t is selected as the action in any round t, then

$$
\begin{aligned}
\Delta_{t,S_t} &= r(S_t^*;\boldsymbol{\mu}) - r(S_t;\boldsymbol{\mu})\\
&\leq r(S_t^*;\hat{\boldsymbol{\mu}}_{t-1}) + \rho_t(S_t^*) - r(S_t;\boldsymbol{\mu})\\
&\leq r(S_t;\hat{\boldsymbol{\mu}}_{t-1}) + \rho_t(S_t) - r(S_t;\boldsymbol{\mu})\\
&\leq 2\rho_t(S_t),
\end{aligned}
$$

where we can find that S_t can only be selected when $\rho_t(S_t) > \Delta_{t,S_t}/2$ is satisfied.

Utilizing this, we consider two conditions based on the value of $\sum_{i\in\tilde{S}_t}\frac{1}{T_{t-1,i}}$ to use the reverse amortization trick. This analysis diverges from that in Section IV-B by introducing $e_t(S_t)$, which focuses our attention solely on the regret contributions from the min-arm (the least played arm in S_t). This refinement effectively eliminates the $O(\log K)$ term from the leading component of our analysis.

We now consider the **first condition**: $\sum_{i\in\tilde{S}_t}\frac{1}{T_{t-1,i}} < \frac{\Delta_{t,S_t}^2}{36C_1|V|}$. We first show by contraction that the confidence interval lies in the second part of Eq. (35), i.e. $\rho_t(S_t) = \sqrt{C_1|V|(\sum_{i\in\tilde{S}_t}\frac{1}{T_{t-1,i}} + 8\log(\frac{2}{\delta})\frac{1}{T_{t-1,\tilde{S}_t}^{\min}})}$. Based on tail bounds for sub-Exponential random variable [19], if

the confidence interval lies in the first part, then

$$\begin{aligned}\tau \le \frac{\nu^2}{b} &= \sum_{i\in\tilde{S}_t} 8C_1 \frac{T^{\min}_{t-1,\tilde{S}_t}}{T^2_{t-1,i}} \\ &\le \sum_{i\in\tilde{S}_t} 8C_1 \frac{1}{T_{t-1,i}} \\ &\le \frac{2\Delta^2_{t,S_t}}{9|V|}.\end{aligned} \tag{36}$$

This indicates that

$$\begin{aligned}\rho_t(S_t) &= \sqrt{|V|\mathbb{E}[Y_{t,S_t}] + |V|\tau} \\ &\le \sqrt{|V|\mathbb{E}[Y_{t,S_t}] + |V|\frac{2\Delta^2_{t,S_t}}{9|V|}} \\ &\le \sqrt{C_1 \sum_{i\in\tilde{S}_t} \frac{|V|}{T_{t-1,i}} + \frac{2|V|\Delta^2_{t,S_t}}{9|V|}} \\ &\le \sqrt{\frac{4\Delta^2_{t,S_t}}{9} + \frac{2|V|\Delta^2_{t,S_t}}{9}} \\ &= \Delta_{t,S_t}/2,\end{aligned} \tag{37}$$

which contradicts the requirement $\rho_t(S_t) > \Delta_{t,S_t}/2$.

Therefore, from contradiction, assuming the confidence interval falls in the first part defined in $\max$ of Eq. (5) leads to $\tau \le \frac{2\Delta^2_{t,S_t}}{9|V|}$ [19], which contradicts $\rho_t(S_t) > \Delta_{t,S_t}/2$. Therefore, $\rho_t(S_t)$ lies in the second part, with $\tau = \frac{8C_1 \log\left(\frac{2}{\delta}\right)}{T^{\min}_{t-1,\tilde{S}_t}}$. The condition for Δ_{t,S_t} becomes $\Delta_{t,S_t} \le \sqrt{\frac{288C_1|V|\log\left(\frac{2}{\delta}\right)}{T^{\min}_{t-1,\tilde{S}_t}}}$.

Introducing a regret allocation function for the reverse amortization: $\kappa_{i,\delta}(S,\ell) = 2\sqrt{\frac{288C_1|V|\log\left(\frac{2}{\delta}\right)}{\ell}}$ for $1 \le \ell \le \frac{288C_1|V|\log\left(\frac{2}{\delta}\right)}{(\Delta^{\min}_i)^2}$ and $i = \operatorname{argmin}_{j\in S} T_{t-1,j}$; otherwise, $\kappa_{i,\delta}(S,\ell) = 0$. It can be easily shown that $\Delta_{t,S_t} \le \sum_{i\in\tilde{S}_t} \kappa_{i,\delta}(S_t, T_{t-1,i})$. Let $p^{D,S}_{\tilde{S}}$ be the probability that the arm set $\tilde{S}$ is triggered when super arm S is selected and $p^* = \min_{S\in\mathcal{S}: p^{D,S}_{\tilde{S}}} p^{D,S}_{\tilde{S}}$, the overall regret for the first condition is bounded by

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{t,\tilde{S}_t}\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\Delta_{t,\tilde{S}_t} \frac{1}{p^{D,S}_{\tilde{S}_t}} \mathbb{I}[\tau_t = \tilde{S}_t]\right]\right] \\ &\le \frac{1}{p^*}\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_t\left[\Delta_{t,\tilde{S}_t} \mathbb{I}[\tau_t = \tilde{S}_t]\right]\right] \\ &= \frac{1}{p^*}\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{t,\tau_t}\right] \\ &\le \frac{1}{p^*}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in\tau_t} \kappa_{i,\delta}(\tau_t, T_{t-1,i})\right] \\ &\le \frac{1}{p^*}\sum_{i\in[m]} \frac{288C_1|V|\log(\frac{2}{\delta})}{\Delta^{\min}_i}.\end{aligned} \tag{38}$$

For the **second condition**: $\sum_{i\in\tilde{S}_t} \frac{1}{T_{t-1,i}} \ge \frac{\Delta^2_{t,S_t}}{36C_1|V|}$, we have $\frac{K}{T^{\min}_{t-1,\tilde{S}_t}} \ge \sum_{i\in\tilde{S}_t} \frac{1}{T_{t-1,i}} \ge \frac{\Delta^2_{t,S_t}}{36C_1|V|}$. Rewriting the inequality, we have $\Delta_{t,S_t} \le \sqrt{\frac{36C_1K|V|}{T^{\min}_{t-1,\tilde{S}_t}}}$. Applying a similar regret allocation method and argument as in the first condition (where we require $\Delta_{t,S_t} \le \sqrt{\frac{36C_1|V|\log(\frac{2}{\delta})}{T^{\min}_{t-1,\tilde{S}_t}}}$), the second condition's contribution to the total regret is at most $\frac{288mC_1K|V|}{p^*\Delta_{\min}}$, which does not depend on the time horizon T.

Considering the first m rounds to ensure each arm is observed at least once, and using a union bound to cover all potential bad events where the confidence interval does not hold, the extra regret is bounded by $\frac{(m+1)\Delta_{\max}}{p^*}$. Finally, the total regret is bounded by:

$$\begin{aligned}Reg(T) \le &\sum_{i\in[m]} \frac{288C_1|V|\log(2|\mathcal{S}|T)}{p^*\Delta^{\min}_i} \\ &+ \frac{288mC_1K|V|}{p^*\Delta_{\min}} + \frac{(m+1)\Delta_{\max}}{p^*},\end{aligned} \tag{39}$$

encapsulating the regret in terms of the minimal gap $\Delta^{\min}_i$, the set size m, and the logarithm of the time horizon T.

H. Distribution-Independent Regret Analysis

The minimum gap $\Delta_{t,S}$ may appear to be negligible, affecting only a single round among many. To provide a comprehensive analysis that is independent of specific distributions, we derive regret bounds for Algorithm 1 and Algorithm 2 that do not depend directly on $\Delta_{t,S}$. Instead, we introduce a fixed gap Δ, the value of which will be specified later. We analyze the regret under two scenarios: $\{\Delta_{t,S_t} \le \Delta\}$ and $\{\Delta_{t,S_t} > \Delta\}$.

Regarding Algorithm 1, for cases where $\Delta_{t,S_t} \le \Delta$, the regret is trivially upper bounded as: $Reg(T, \{\Delta_{t,S_t} \le \Delta\}) \le T\Delta$. For cases where $\Delta_{t,S_t} > \Delta$, we simplify the regret calculation by approximating all minimal gaps $\Delta^{\min}_i$ with Δ, yielding the bound:

$$\begin{aligned}&Reg(T, \{\Delta_{t,S_t} > \Delta\}) \\ \le &O\left(\frac{m|V|\log K\log T}{\Delta} + m\log\left(\frac{K}{\Delta}\right)\log T\right).\end{aligned} \tag{40}$$

By selecting Δ as:

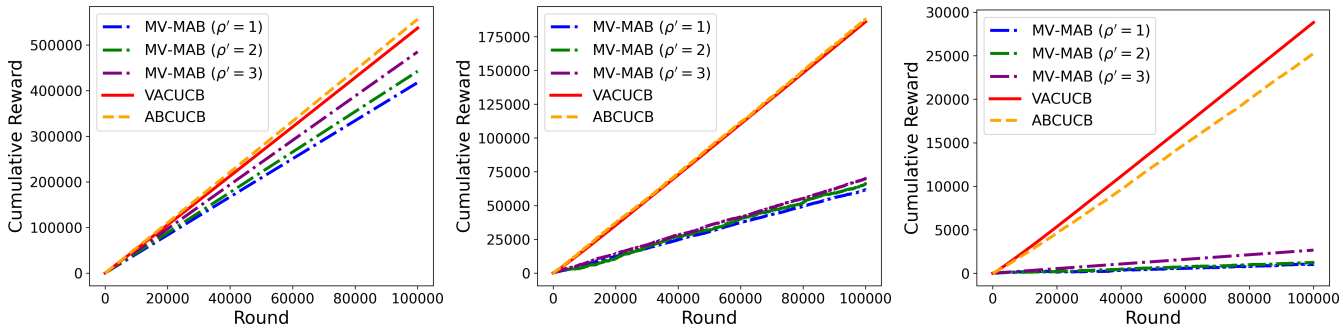
$$\Delta = \Theta\left(\sqrt{\frac{m|V|\log T\log K}{T}} + \frac{m\log K\log T}{T}\right),$$

we establish a new regret bound for Algorithm 1 that is independent of the variations in gaps across different rounds:

$$Reg(T) \le O\left(\sqrt{m|V|(\log K)T\log T} + m\log(KT)\log T\right). \tag{41}$$

This formulation of the regret bound does not depend on specific gap values and provides an $O(\sqrt{\log T})$ improvement over the previously established distribution-independent bound of $O\left(\sqrt{m|V|(\log K)T}\log(KT) + m\log^2(KT)\log T\right)$ in [10].

For Algorithm 2, which operates under the assumption of independent arms with probabilistically determined outcomes, we apply a similar methodology by setting $\Delta = \Theta\left(\sqrt{\frac{m|V|\log T}{T}}\right)$. The resultant distribution-independent regret for this setting is: $O\left(\sqrt{m|V|T\log T}\right)$.



(a) Mobile Crowdsensing (b) Online Content Delivery (c) Dynamic Channel Allocation

Fig. 8: Cumulative reward across three network applications. Only in mobile crowdsensing, each arm operates independently.

I. Extend Experiments

We evaluate our method with the risk-averse multi-armed bandit under the mean-variance measure to assess their performance across different network applications.

In the context of risk-averse strategies, the mean-variance multi-armed bandit (MV-MAB) model, which is often employed in financial portfolio selection, utilizes the mean-variance measure as a crucial factor in selecting arms [23], [24]. Specifically, with $\hat{\mu}_{i,t} = \frac{1}{t}\sum_{s=1}^{t} X_{i,s}$ and $\hat{\sigma}^2_{i,t} = \frac{1}{t}\sum_{s=1}^{t}(X_{i,s} - \hat{\mu}_{i,t})^2$, the empirical mean-variance $\widehat{MV}_i$ of an arm i with t samples is defined as:

$$\widehat{MV}_{i,t} = \hat{\sigma}^2_{i,t} - \rho' \hat{\mu}_{i,t}. \tag{42}$$

This method emphasizes a balance between risk (as variance) and reward (as mean), which is regulated by the risk tolerance coefficient ρ'. As $\rho' \to \infty$, the model reverts to a traditional multi-armed bandit framework. Conversely, as $\rho' \to 0$, the focus shifts toward minimizing variance.

The selection process utilizes the empirical mean-variance instead of the UCB rule at each round t. The concept of risk-averse regret, which integrates expected rewards with risk measures, differs significantly from $(1 - 1/e, 1)$-approximate regret in Eq. (3), making a direct comparison of regrets impractical. Consequently, we evaluate our algorithm against the MV-MAB model by focusing on the reward outcomes of the three network applications.

The comparative results, shown in Fig. 8, illustrate the performance on cumulative rewards of different algorithms. Specifically, in scenarios with independent arms as shown in Fig. 8a, our ABCUCB algorithm outperforms other multi-armed bandit strategies, registering an average improvement of 2.53% over the VACUCB algorithm. When benchmarked against the MV-MAB model at varying levels of risk tolerance (ρ'), the reward gains are as follows: 30.54% over MV-MAB ($\rho' = 1$), 23.28% over MV-MAB ($\rho' = 2$), and 12.48% over MV-MAB ($\rho' = 3$). Similar patterns are observed in the online content delivery and dynamic channel allocation applications (Fig. 8b and Fig. 8c), further evidencing the robustness and effectiveness of our strategies in different network application settings compared to MV-MAB.