Thompson Sampling for Combinatorial Network Optimization in Unknown Environments

Alihan Hüyük and Cem Tekin, *Senior Member, IEEE*

***Abstract*— Influence maximization, adaptive routing, and dynamic spectrum allocation all require choosing the right action from a large set of alternatives. Thanks to the advances in combinatorial optimization, these and many similar problems can be efficiently solved given an environment with known stochasticity. In this paper, we take this one step further and focus on combinatorial optimization in unknown environments. We consider a very general learning framework called combinatorial multi-armed bandit with probabilistically triggered arms and a very powerful Bayesian algorithm called Combinatorial Thompson Sampling (CTS). Under the semi-bandit feedback model and assuming access to an oracle without knowing the expected base arm outcomes beforehand, we show that when the expected reward is Lipschitz continuous in the expected base arm outcomes CTS achieves $O(\sum_{i=1}^{m} \log T/(p_i \Delta_i))$ regret and $O(\max\{\mathbb{E}[m\sqrt{T \log T/p^*}], \mathbb{E}[m^2/p^*]\})$ Bayesian regret, where m denotes the number of base arms, p_i and Δ_i denote the minimum non-zero triggering probability and the minimum suboptimality gap of base arm i respectively, T denotes the time horizon, and p^* denotes the overall minimum non-zero triggering probability. We also show that when the expected reward satisfies the triggering probability modulated Lipschitz continuity, CTS achieves $O(\max\{m\sqrt{T \log T}, m^2\})$ Bayesian regret, and when triggering probabilities are non-zero for all base arms, CTS achieves $O(1/p^* \log(1/p^*))$ regret independent of the time horizon. Finally, we numerically compare CTS with algorithms based on upper confidence bounds in several networking problems and show that CTS outperforms these algorithms by at least an order of magnitude in majority of the cases.**

***Index Terms*— Combinatorial network optimization, multi-armed bandits, Thompson sampling, regret bounds, online learning.**

I. Introduction

HOW should an advertiser promote its products in a social network to reach to a large set of users with a limited budget [2], [3]? How should a search engine suggest a ranked list of items to its users to maximize the click-through rate [4]? How should a base station allocate its users to channels to maximize the system throughput [5]? How should a mobile crowdsourcing platform dynamically assign available tasks to its workers to maximize the performance [6]? How can we identify the most reliable paths from source to destination under probabilistic link failures [7]? All of these problems require optimizing decisions among a vast set of alternatives. When the probabilistic description of the environment is fully specified, these problems—and many others—are solved using computationally efficient exact or approximation algorithms. In this paper, we focus on a much more difficult and realistic problem: How should we learn the optimal decisions in these complex problems via repeated interaction with the environment when the probabilistic description of the environment is unknown or only partially known?

It is natural to assume that the environment is unknown in many real-world applications. For instance, the advertiser may not know with what probability user i will influence its neighbor j in a social network or the search engine may not know with what probability user i will click the item shown on position j beforehand. Moreover, decisions are need to be made sequentially over time. For instance, the recommender system should show a new list of items to each arriving user and the base station should reallocate network resources when the channel conditions change or the users leave/enter the system. Obviously, future decisions of the learner must be guided based on what it has observed thus far, i.e., the trajectory of actions, observations and rewards generated by the learner's past decisions. Importantly, both the cumulative reward of the learner and what it has learned so far also depend on this trajectory. Therefore, the learner needs to balance how much it earns (by exploiting the actions it believes to be the best) and how much it learns (by exploring actions it does not know much about) in order to maximize its long-term performance. In this paper, we solve the formidable task of combinatorial optimization in unknown environments by modeling it as a combinatorial multi-armed bandit (MAB).

MAB problems have a long history as they exhibit the prime example of the tradeoff between exploration and exploitation [5], [8]. In the classical MAB, at each round the learner selects an arm (action) which yields a random reward that comes from an unknown distribution. The goal of the learner is to maximize its expected cumulative reward over all rounds by learning to select arms that yield high rewards. The learner's performance is measured by its regret with respect to an oracle that always selects the arm with the highest expected reward.

Manuscript received July 7, 2019; revised June 29, 2020; accepted September 10, 2020; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor L. Huang. Date of publication October 2, 2020; date of current version December 16, 2020. This work was supported in part by the Scientific and Technological Research Council of Turkey under Grant 215E342. A preliminary version of this work was presented in AISTATS 2019. *(Corresponding author: Cem Tekin.)*

Alihan Hüyük was with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey. He is now with the Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, U.K. (e-mail: ah2075@cam.ac.uk).

Cem Tekin is with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: cemtekin@ee.bilkent.edu.tr).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNET.2020.3025904

It is shown that when the arms' rewards are independent, any uniformly good policy will incur at least logarithmic in time regret [9].

Several classes of policies are proposed for the learner to minimize its regret. One example is Thompson sampling [10]–[12], which is a Bayesian method. In this method, the learner keeps a posterior distribution over the expected arm rewards, and at each round takes a sample from each arm's posterior, and then, plays the arm with the largest sample. Reward observed from the played arm is then used to update its posterior. This sampling strategy allows the learner to frequently select the arms whose probabilities of being optimal are the highest based on their posteriors and to occasionally explore inferior arms to refine their posteriors. Policies in the other end of the spectrum use the principle of optimism under the face of uncertainty. Notable examples include policies based on upper confidence bound (UCB) indices [9], [13], [14], which are usually composed of sample mean reward of an arm plus an exploration bonus that accounts for the uncertainty in the arm's reward estimates. The strategy is to play the arm with the highest UCB index to trade-off exploration and exploitation. Unlike Thompson sampling, performance of this type of policies heavily rely on the confidence sets used to compute the exploration bonus [12]. This together with the superior performance of Thompson sampling documented in numerous applications [15], [16] motivate us to consider a Thompson sampling based approach for our problem.

Our main focus in this paper, i.e., combinatorial MAB (CMAB) [5], [17]–[19], is an extension of MAB where the learner selects a *super arm* at each round, which is defined to be a subset of the *base arms*. Then, the learner observes and collects the reward associated with the selected super arm, and also observes the outcomes of the base arms that are in the selected super arm. This type of feedback is also called *semi-bandit* feedback. For instance, when allocating users to orthogonal channels, each user-channel pair represents a base arm, the super arm is the set of user-channel pairs in the selected allocation, outcomes of base arms are indicators of successful packet transmissions and the reward is the number of packets successfully transmitted, i.e., sum of the indicators. While CMAB is general enough to model the aforementioned resource allocation problem, it does not fully capture the probabilistic structure of influence maximization, item list recommendation and reliable packet routing applications discussed in the preceding paragraphs. Therefore, we consider a generalized version of CMAB, called CMAB with probabilistically triggered arms (CMAB-PTA) [20], where the selected super arm probabilistically triggers a set of base arms, and the expected reward obtained in a round is a function of the set of triggered base arms and their expected outcomes. For instance, in influence maximization, each edge of the graph represents a base arm, the super arm is the selected seed set of nodes, outcomes of base arms are indicators of influence propagation on the corresponding edge (see, e.g., the independent cascade model [21]) and the reward is the number of influenced nodes, i.e., the set of nodes reachable from the seed set of nodes after the outcomes of base arms are realized. Triggered base arms in this case correspond to the set of edges that originate from all influenced nodes (including the seed set).

The regret for CMAB-PTA is defined as the difference between the expected cumulative reward of an oracle that always selects the super arm with the highest expected reward and that of the learner given a particular environment. Then, the Bayesian regret is the expected regret over all possible environments. Our goal is to design an algorithm that achieves the smallest rate of growth of the (Bayesian) regret over time, as this will ensure that the average reward of the learner will converge to the highest possible expected reward. To this end, we propose a Bayesian algorithm called combinatorial Thompson sampling (CTS) and analyze its regret assuming that the learner does not know the expected base arm outcomes beforehand but has access to an exact optimization oracle. Essentially, this oracle outputs an estimated optimal super arm given estimates of expected base arm outcomes as inputs. When the expected reward is Lipschitz continuous in the expected base arm outcomes, we show that CTS achieves $O(\sum_{i=1}^{m} \log T/(p_i \Delta_i))$ regret and $O(\max\{\mathbb{E}[m\sqrt{T \log T/p^*}], \mathbb{E}[m^2/p^*]\})$ Bayesian regret, where m denotes the number of base arms, p_i denotes the minimum non-zero triggering probability of base arm i, Δ_i denotes the minimum suboptimality gap of base arm i, T denotes the time horizon, and p^* denotes the overall minimum non-zero triggering probability. We also show that when the expected reward satisfies the triggering probability modulated (TPM) Lipschitz continuity in [22], which is a stronger assumption than the regular Lipschitz continuity yet still satisfied by the network optimization problems that we consider, CTS achieves $O(\max\{m\sqrt{T \log T}, m^2\})$ Bayesian regret independent of the triggering probabilities.

In addition to these more general cases, we also prove that when triggering probabilities are non-zero for all base arms, CTS achieves $O(1/p^* \log(1/p^*))$ regret independent of the time horizon. This setting is of particular interest since it can model random behavior of users in a recommender system. For instance, a user may rate an item even when it is not in the list of recommended items as a result of an exogenous event (by rating the item on a partner website or by explicitly navigating to the item to rate it). Moreover, it is also closely linked to related work on online learning with probabilistic graph feedback [23], [24] and MAB with side observations [25]. Specifically, the models in [24] and [25] become special cases of our work when the graph is fully-connected for the one-step case and connected for the cascade case in [24] and when the probability of having an observation from any arm is non-zero in [25].

We complement our theoretical findings via extensive simulations in the following combinatorial network optimization problems: cascading bandits [4], probabilistic maximum coverage bandits [20] and influence maximization bandits [20]. For cascading bandits, we show that CTS, which uses Beta posterior on base arms significantly outperforms all competitor algorithms that use either UCB indices [4] or Thompson sampling with Gaussian posterior [26]. The latter finding emphasizes the importance of working with the correct type of posterior. For probabilistic maximum coverage bandits,

we show that CTS achieves an order of magnitude improvement over combinatorial UCB (CUCB) in [20] when both algorithms use an exact oracle. For influence maximization bandits, we show a similar result even when both algorithms use an approximation oracle instead of an exact oracle.

In summary, the main contribution of this paper is to analyze Thompson sampling for a very general combinatorial online learning framework that is comprehensive enough to model many different sequential decision-making applications defined over networks and show its optimality both theoretically and experimentally. The rest of the paper is organized as follows. Related work is given in Section II followed by problem formulation in Section III. Applications of CMAB-PTA are detailed in Section IV. Description of CTS and regret bounds are given in Section V. Proofs of the main results are explained in Sections VI and VII (some proofs are left to the supplemental document). Numerical results are presented in Section VIII and concluding remarks are given in Section IX.

II. Related Work

CMAB has been studied under various assumptions on the relation between super arms, base arms and rewards [17]. Here, we mainly discuss the related works that assume semi-bandit feedback as we do in our work. A version of CMAB in which the expected reward of a super arm is a linear combination of the expected outcomes of the base arms in that super arm is studied in [5]. For this problem, it is shown in [18] that a combinatorial version of UCB1 in [14] achieves $O(Km \log T/\Delta)$ gap-dependent and $O(\sqrt{KmT \log T})$ gap-free (worst-case) regrets, where m is the number of base arms, K is the maximum number of base arms in a super arm, and Δ is the gap between the expected reward of the optimal super arm and the second best super arm.

Later on, this setting is generalized to allow the expected reward of each super arm to be a more general function of the expected outcomes of the base arms that obeys certain monotonicity and bounded smoothness conditions [19]. The main challenge in the general case is that the optimization problem itself is NP-hard, but an approximately optimal solution can usually be computed efficiently for many special cases [27]. Therefore, it is assumed that the learner has access to an approximation oracle, which can output a super arm that has expected reward that is at least α fraction of the optimal reward with probability at least β when given the expected outcomes of the base arms. Thus, the regret is measured with respect to the $\alpha\beta$ fraction of the optimal reward, and it is proven that a combinatorial variant of UCB1, called CUCB, achieves $O(\sum_{i=1}^{m} \log T/\Delta_i)$ regret when the bounded smoothness function is $f(x) = \gamma x$ for some $\gamma > 0$, where Δ_i is the minimum gap between the expected reward of the optimal super arm and the expected reward of any suboptimal super arm that contains base arm i.

Recently, it is shown in [28] that Thompson sampling can achieve $O(\sum_{i=1}^{m} \log T/\Delta_i)$ regret for the general CMAB under a Lipschitz continuity assumption on the expected reward, given that the learner has access to an exact computation oracle, which outputs an optimal super arm when given the set of expected base arm outcomes. Moreover, it is also shown that in general the learner cannot guarantee sublinear regret when it only has access to an approximation oracle. Since the setting studied in this paper is a special case of ours, for our theoretical analysis we also assume that the learner uses an exact computation oracle. Nevertheless, we show in Section VIII that in practice CTS works well even when used with an approximation oracle. Another related work on CMAB [29] considers a new smoothness condition termed the Gini-weighted smoothness on the expected reward. For some problem types, this leads to regret bounds with better dependency on the sizes of super arms when compared with the common linear dependency of the existing algorithms.

Different from CMAB, papers on CMAB-PTA assume that the expected reward is a function of the expected outcomes of the triggered base arms, which is a random superset of base arms in the selected super arm. For this problem, it is shown in [20] that logarithmic regret is achievable when the expected reward function has the ℓ_∞ bounded smoothness property. However, this bound depends on $1/p^*$, where p^* is the minimum non-zero triggering probability. Later, it is shown in [22] that under a stricter smoothness assumption on the expected reward function, called triggering probability modulated (TPM) bounded smoothness, it is possible to achieve regret that does not depend on $1/p^*$. It is also shown in this work that the dependence on $1/p^*$ is unavoidable for the general case. In another work [30], CMAB-PTA is considered for the case when the arm triggering probabilities are all positive, and it is shown that both CUCB and CTS achieve bounded regret. However, their $O((1/p^*)^4)$ bound has a much worse dependence on p^* than our $O(1/p^* \log(1/p^*))$ bound.

Apart from the works mentioned above, numerous other works also tackle related online learning problems. For instance, [31] considers matroid bandits, which is a special case of CMAB where the super arms are given as independent sets of a matroid with base arms being the elements of the ground set, and the expected reward of a super arm is the sum of the expected outcomes of the base arms in the super arm. Another example is cascading bandits [4], which is a special case of CMAB-PTA, where each super arm corresponds to a ranked list of items and base arms are triggered according to a user click model. A plethora of papers exist on UCB based policies for variants of these two models (see e.g., [32] for a variant of matroid bandits and [33] and [34] for variants of cascading bandits.) Apart from these, [26] considers Thompson sampling with Gaussian posterior for cascading bandits and proves that the worst-case regret is $\tilde{O}(\sqrt{KmT})$. We show in Section VIII that CTS significantly outperforms their algorithm for cascading bandits. We think that this is the case in practice because Beta posterior is more suitable in modeling click probabilities compared to Gaussian posterior.

Several other works focus on contextual CMAB [34]–[36], CMAB with adversarial rewards [37], [38] and CMAB with knapsacks [39]. Most recently there has been a surge of interest in analyzing CMAB under the full-bandit feedback setting, where the learner only observes the reward of the selected super arm but not the outcomes of the base arms [40], [41]. For instance, [41] uses a sampling method based

TABLE I
SUMMARY OF THE RELATED WORK IN COMPARISON WITH OUR WORK

Publ.	Algorithm	Oracle	PTAs	Regret Bound
[19]	CUCB	Approx.	No	$O(\sum_i \log T/\Delta_i)$
[20]	CUCB	Approx.	Yes	$O(\sum_i \log T/(p_i\Delta_i))$
[22]	CUCB	Approx.	Yes	$O(\sum_i \log T/\Delta_i)^\dagger$
[28]	CTS	Exact	No	$O(\sum_i \log T/\Delta_i)$
[30]	CUCB & CTS	Approx.	Yes*	$O((1/p^*)^4)$
Ours	CTS	Exact	Yes	$O(\sum_i \log T/(p_i\Delta_i))$
			Yes	$O(\max\{\mathbb{E}[m\sqrt{T\log T/p^*}], \mathbb{E}[m^2/p^*]\})^\ddagger$
			Yes	$O(\max\{m\sqrt{T\log T}, m^2\})^{\dagger\ddagger}$
			Yes*	$O(1/p^*\log(1/p^*))$

*The case when the arm triggering probabilities are all positive.
†Under the TPM bounded smoothness assumption.
‡Bound for the Bayesian regret.

on Hadamard matrices to estimate base arm rewards from full-bandit feedback. On the other hand, [42] considers a more general feedback model where the learner observes a linear combination of base arm's rewards. Table I compares our work with the most closely related publications in terms of their assumptions and the regret bounds they show.

III. PROBLEM FORMULATION

CMAB-PTA is a decision-making problem where the learner interacts with its environment through m base arms, indexed by the set $[m] := \{1, 2, \ldots, m\}$ sequentially over rounds indexed by $t \in [T]$. In this paper, we consider the model introduced in [20] and borrow the notation from [28]. In this model, the following events take place in order in each round t:

- The learner selects a subset of base arms, denoted by $S(t)$, which is called a super arm.
- $S(t)$ causes some other base arms to probabilistically trigger based on a stochastic triggering process, which results in a set of triggered base arms $S'(t)$ that contains $S(t)$.
- The learner obtains a reward that depends on $S'(t)$ and observes the outcomes of the base arms in $S'(t)$.

Next, we describe in detail the base arm outcomes, the super arms, the triggering process, the reward, the observation (feedback) model and the regret.

A. Base Arm Outcomes

In each round t, the environment draws a random outcome vector $\boldsymbol{X}(t) := (X_1(t), X_2(t), \ldots, X_m(t))$ from a probability distribution D on $[0,1]^m$ independent of the previous rounds, where $X_i(t)$ represents the outcome of base arm i. D is unknown by the learner, but it belongs to a class of distributions $\mathcal{D}$ which is known by the learner. We define the mean outcome (parameter) vector as $\boldsymbol{\mu} := (\mu_1, \mu_2, \ldots, \mu_m)$, where $\mu_i := \mathbb{E}_{\boldsymbol{X}\sim D}[X_i(t)]$, and use $\boldsymbol{\mu}_S$ to denote the projection of $\boldsymbol{\mu}$ on S for $S \subseteq [m]$.

Since CTS computes a posterior over $\boldsymbol{\mu}$, the following assumption is made to have an efficient and simple update of the posterior distribution.

Assumption 1: The outcomes of all base arms are mutually independent, i.e., $D = D_1 \times D_2 \times \cdots \times D_m$.

Note that this independence assumption holds in many applications, including the influence maximization problem with independent cascade influence propagation model [21].

B. Super Arms and the Triggering Process

The learner is allowed to select $S(t)$ from a subset of $2^{[m]}$ denoted by $\mathcal{I}$, which corresponds to the set of feasible super arms. Once $S(t)$ is selected, all base arms $i \in S(t)$ are immediately triggered. These arms can trigger other base arms that are not in $S(t)$, and those arms can further trigger other base arms, and so on. At the end, a random superset $S'(t)$ of $S(t)$ is formed that consists of all triggered base arms as a result of selecting $S(t)$. We have $S'(t) \sim D^{\text{trig}}(S(t), \boldsymbol{X}(t))$, where D^{trig} is the probabilistic triggering function that describes the triggering process. For instance, in the influence maximization problem, D^{trig} may correspond to the independent cascade influence propagation model defined over a given influence graph [21]. The triggering process can also be described by a set of triggering probabilities. For each $i \in [m]$ and $S \in \mathcal{I}$, $p_i^{D',S}$ denotes the probability that base arm i is triggered when super arm S is selected given that the arm outcome distribution is $D' \in \mathcal{D}$. For simplicity, we let $p_i^S = p_i^{D,S}$, where D is the true arm outcome distribution. Let $\tilde{S} := \{i \in [m] : p_i^S > 0\}$ be the set of all base arms that could potentially be triggered by super arm S, which is called the *triggering set* of S. We have that $S(t) \subseteq S'(t) \subseteq \tilde{S}(t) \subseteq [m]$. We define $p_i := \min_{S\in\mathcal{I}:i\in\tilde{S}} p_i^S$ as the minimum nonzero triggering probability of base arm i, and $p^* := \min_{i\in[m]} p_i$ as the minimum nonzero triggering probability.

Before moving on, we would like to point out that the entire triggering process could have been represented by writing $S'(t) \sim \bar{D}^{\text{trig}}(S(t))$, where any possible dependence of the process on the outcome distribution D would have been hidden inside $\bar{D}^{\text{trig}}$. Instead, we chose to break down the triggering process into two stages: $\boldsymbol{X}(t) \sim D$ and $S'(t) \sim D^{\text{trig}}(S(t), \boldsymbol{X}(t))$, where D and D^{trig} together are equivalent to $\bar{D}^{\text{trig}}$. This is motivated by the prior knowledge of the learner. Note that, while the learner fully knows D^{trig}, it does not know anything about D except the class of distributions $\mathcal{D}$ that it belongs to, resulting in only a partial knowledge about $\bar{D}^{\text{trig}}$.

C. Reward

At the end of round t, the learner receives a reward that depends on the set of triggered arms $S'(t)$ and the outcome vector $\boldsymbol{X}(t)$, which is denoted by $R(S'(t), \boldsymbol{X}(t))$. For simplicity of notation, we also use $R(t) = R(S'(t), \boldsymbol{X}(t))$ to denote the reward in round t. Note that whether a base arm is in the selected super arm or is triggered afterwards is not relevant in terms of the reward. We assume that the expected reward depends on the mean outcome vector in a specific way by making the following mild assumptions about the expected reward function. We note that these assumptions are standard in the CMAB literature [20], [28] and hold for the networking

applications given in Section IV. The first assumption states that the expected reward is only a function of $S(t)$ and $\boldsymbol{\mu}$.

Assumption 2: The expected reward of super arm $S \in \mathcal{I}$ only depends on S and the mean outcome vector $\boldsymbol{\mu}$, i.e., there exists a function r such that

$$\begin{aligned}\mathbb{E}[R(t)] &= \mathbb{E}_{S'(t)\sim D^{\text{trig}}(S(t),\boldsymbol{X}(t)),\boldsymbol{X}(t)\sim D}[R(S'(t),\boldsymbol{X}(t))] \\ &= r(S(t),\boldsymbol{\mu}) \ .\end{aligned}$$

In order to learn the best action, we require the estimate of the expected reward vector to converge to the true expected reward vector as the number of observations increases. This can be done when the expected reward varies smoothly with the mean outcome vector. Below, we state a form of continuity for the expected reward.

Assumption 3: (Lipschitz continuity) There exists a constant $B > 0$, such that for every super arm S and every pair of mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$, we have

$$|r(S,\boldsymbol{\mu}) - r(S,\boldsymbol{\mu'})| \leq B\|\boldsymbol{\mu}_{\tilde{S}} - \boldsymbol{\mu'}_{\tilde{S}}\|_1$$

where $\|\cdot\|_1$ denotes the l_1 norm.

In addition to Lipschitz continuity, we also consider the *triggering probability modulated* (TPM) Lipschitz continuity introduced in [22]. This is a stricter assumption than the regular Lipschitz continuity (one implies the other) but leads to tighter regret bounds in terms of the triggering probabilities. All of the networking applications considered in Section IV still satisfy the TPM Lipschitz continuity.

Assumption 4: (Triggering probability modulated Lipschitz continuity) There exists a constant $B' > 0$, such that for every super arm S and every pair of outcome distributions D and D' with mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$ respectively, we have

$$|r(S,\boldsymbol{\mu}) - r(S,\boldsymbol{\mu'})| \leq B' \sum_{i\in\tilde{S}} p_i^{D,S}|\mu_i - \mu_i'| \ .$$

Finally, we require a monotonicity assumption in order to facilitate the UCB-based analysis that some of our results rely on, namely Theorems 2 and 3. Again, all of the networking applications considered in Section IV satisfy the following monotonicity assumption.

Assumption 5: For every super arm S and every pair of mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$, we have $r(S,\boldsymbol{\mu}) \leq r(S,\boldsymbol{\mu'})$ if $\mu_i \leq \mu_i'$ for all $i \in [m]$.

D. Observation Model

We consider the semi-bandit feedback model, where at the end of round t, the learner observes the individual outcomes of the triggered arms, denoted by $Q(S'(t),\boldsymbol{X}(t)) := \{(i, X_i(t)) : i \in S'(t)\}$. Again, for simplicity of notation, we also use $Q(t) = Q(S'(t),\boldsymbol{X}(t))$ to denote the observation at the end of round t. Based on this, the only information available to the learner when choosing the super arm to select in round $t+1$ is its observation history, given as $\mathcal{F}_t := \{(S(\tau), Q(\tau)) : \tau \in [t]\}$.

In short, the tuple $([m], \mathcal{I}, D, D^{\text{trig}}, R)$ constitutes a CMAB-PTA problem instance. Among the elements of this tuple only D is unknown to the learner.

E. Regret

In order to evaluate the performance of the learner, we define the set of optimal super arms given an m-dimensional parameter vector $\boldsymbol{\theta}$ as $\text{OPT}(\boldsymbol{\theta}) := \text{argmax}_{S\in\mathcal{I}} r(S,\boldsymbol{\theta})$. We use $\text{OPT} := \text{OPT}(\boldsymbol{\mu})$ to denote the set of optimal super arms given the true mean outcome vector $\boldsymbol{\mu}$. Based on this, we let S^* to represent a specific super arm in $\text{argmin}_{S\in\text{OPT}}|\tilde{S}|$, which is the set of super arms that have triggering sets with minimum cardinality among all optimal super arms. We also let $k^* := |S^*|$ and $\tilde{k}^* := |\tilde{S}^*|$.

Next, we define the suboptimality gap due to selecting super arm $S \in \mathcal{I}$ as $\Delta_S := r(S^*,\boldsymbol{\mu}) - r(S,\boldsymbol{\mu})$, the maximum suboptimality gap as $\Delta_{\max} := \max_{S\in\mathcal{I}}\Delta_S$, and the minimum suboptimality gap of base arm i as $\Delta_i := \min_{S\in\mathcal{I}-\text{OPT}:i\in\tilde{S}}\Delta_S$.[1] The goal of the learner is to minimize the (expected) regret over the time horizon T, given by

$$\begin{aligned}\text{Reg}(T) &:= \mathbb{E}\left[\sum_{t=1}^{T}(r(S^*,\boldsymbol{\mu}) - r(S(t),\boldsymbol{\mu}))\middle|\boldsymbol{\mu}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\Delta_{S(t)}\middle|\boldsymbol{\mu}\right] \ . \end{aligned} \tag{1}$$

In addition to the expected regret, we also consider the *Bayesian regret*, given by

$$\begin{aligned}\text{BayReg}(T) &:= \mathbb{E}\left[\sum_{t=1}^{T}(r(S^*,\boldsymbol{\mu}) - r(S(t),\boldsymbol{\mu}))\right] \\ &= \mathbb{E}_{\boldsymbol{\mu}}[\text{Reg}(T)]\end{aligned}$$

where the true mean outcome vector $\boldsymbol{\mu}$ is viewed as a random variable. For simplicity, we will assume that $\boldsymbol{\mu}$ has a uniform prior. However, this can easily be extended to any other Dirichlet prior simply by modifying the initial values of a_i's and b_i's in Algorithm 1, which determine the initial prior over the base arm outcomes. It is important to note here that asymptotic bounds on the Bayesian regret are essentially asymptotic (gap-free) bounds on the regret [12]. Formally, if $\text{BayReg}(T) \in O(f(T))$ for some non-negative function $f(T)$, then $\text{Reg}(T) \in O_P(f(T))$, that is there exists $T_0 > 0$ such that for all $\epsilon > 0$ there exists $M > 0$ such that $\mathbb{P}(\text{Reg}(T)/f(T) \geq M) \leq \varepsilon$ for all $T > T_0$.

IV. Networking Applications

Here, we introduce three networking applications of CMAB-PTA: cascading bandits, probabilistic maximum coverage bandits, and influence maximization bandits. Numerical experiments given in Section VIII explore specific cases of all these problems that are generated either synthetically or from real-world data.

A. Cascading Bandits

1) Disjunctive Form for Search Engine Optimization: In the disjunctive form of the cascading bandit problem [4], a search engine outputs a list of K web pages for each of its W users among a set of V web pages. Then, the users examine

[1] If there is no such super arm S, let $\Delta_i = \infty$.

their respective lists, and click on the first page that they find attractive. If all pages fail to attract them, they do not click on any page. The goal of the search engine is to maximize the number of clicks.

This problem can be modeled as an instance of CMAB-PTA as follows. The base arms are page-user pairs (i,j), where $i \in [V]$ and $j \in [W]$. User j finds page i attractive independent of other users and other pages with probability $p_{i,j}$. The super arms are W-many lists of K-tuples, where each K-tuple represents the list of pages shown to a user. Given a super arm S, let $S(k,j)$ denote the kth page that is selected for user j. Then, the triggering probabilities can be written as

$$p^S_{(i,j)} = \begin{cases} 1 & \text{if } i = S(1,j) \\ \prod_{k'=1}^{k-1}(1-p_{S(k',j),j}) & \text{if } \exists k \neq 1 : i = S(k,j) \\ 0 & \text{otherwise} \end{cases},$$

that is we observe feedback for a top selection immediately, and observe feedback for the other selections only if all previous selections fail to attract the user. The expected reward of playing super arm S can be written as

$$r(S,\boldsymbol{p}) = \sum_{j=1}^{W}\left(1 - \prod_{k=1}^{K}(1-p_{S(k,j),j})\right)$$

for which Assumptions 3 and 4 hold when $B=1$ and $B'=1$ respectively.

2) Conjunctive Form for Network Routing Reliability: One can also consider the conjunctive analogue of the problem, where the goal of the search engine is to—somewhat peculiarly—maximize the number of users with lists that do not contain any unattractive page, and when examining their lists, users provide feedback by reporting the first unattractive page. Formally,

$$p^S_{(i,j)} = \begin{cases} 1 & \text{if } i = S(1,j) \\ \prod_{k'=1}^{k-1} p_{S(k',j),j} & \text{if } \exists k \neq 1 : i = S(k,j) \\ 0 & \text{otherwise} \end{cases}$$

and

$$r(S,\boldsymbol{p}) = \sum_{j=1}^{W}\prod_{k=1}^{K} p_{S(k,j),j} \ .$$

This conjunctive form fits particularly well to the network reliability problem [7], where we are interested in finding the most reliable routing path in a communication network. We consider routing paths as super arms, $\mathcal{I}$ being the set of all possible routing paths. Each routing path $S \in \mathcal{I}$ consists of a variable number of ordered links that correspond to the base arms. We denote the index of kth link in routing path S as $S(k)$ and the length of the path as $|S|$. Each link $i \in [m]$ in a routing path can fail independently from all other links with probability $1-p_i$. Then, the probabilistic reliability of a routing path is defined as the probability of successful operation with no link in the path failing.

Since we can only observe whether a link has failed or not up to the first link that has failed, the triggering probability of link i when routing path S is selected can be written as

$$p^S_i = \begin{cases} 1 & \text{if } i = S(1) \\ \prod_{k'=1}^{k-1} p_{S(k')} & \text{if } \exists k \neq 1 : i = S(k) \\ 0 & \text{otherwise} \end{cases}$$

and the probabilistic reliability of routing path S—in other words, the expected reward—becomes

$$r(S,\boldsymbol{p}) = \prod_{k=1}^{|S|} p_{S(k)} \ .$$

B. Probabilistic Maximum Coverage Bandits

In the probabilistic maximum coverage problem, an online shopping site advertises K items that are selected from a catalog of V items to its W users. Each user inspects all of the items that are advertised and likes one of the attractive items. The users do not like any item if none of the items attract them. The goal of the shopping site is to maximize the number of likes. Analogous to cascading bandits, in this problem, base arms are item-user pairs (i,j), where $i \in [V]$ and $j \in [W]$. User j finds item i attractive independent of other users and other items with probability $p_{i,j}$. The super arms are the set of all pairs (i,j) such that item i is the element of a size-K subset of $[V]$.

This can also model the problem of allocating orthogonal channels to secondary users in a cognitive radio network [5]. Consider V as the number of orthogonal channels, W as the number of secondary users ($V > W$), and $p_{i,j}$ as the expected throughput that user j can obtain using channel i. We would like to maximize the expected sum throughput by allocating each user j a unique channel $c_j \in [V]$ so that $c_j = c_{j'}$ if and only if $j = j'$ for all $j, j' \in [W]$. Given one such allocation, the corresponding super arm would be the set $S = \{(c_j, j)\}_{j=1}^{W}$ and the expected reward of it can be written as $r(S,\boldsymbol{p}) = \sum_{(i,j)\in S} p_{i,j}$. Allocating orthogonal channels to secondary users can also be conceptualized as allocating tasks to workers in a mobile crowdsourcing platform [6], [43]. Then, $p_{i,j}$ would be the probability of worker j completing task i successfully and $r(S,\boldsymbol{p})$ would be the expected number of completed tasks.

In its classical form, this problem does not have any PTAs. In order to provide an example case with strictly positive triggering probabilities, we introduce the *word-of-mouth effect* as follows. Regardless of the shopping site's decisions, we assume that users inspect, i.e., they explicitly search or navigate to, unadvertised items independently with probability p^*.[2] This can happen if users hear about the items outside of the shopping site (e.g., from their friends or from another venue). Then, the triggering probabilities can be written as

$$p^S_{(i,j)} = \begin{cases} 1 & \text{if } (i,j) \in S \\ p^* & \text{otherwise} \end{cases}$$

[2]For simplicity we assume that p^* is the same for all items while it can be different in practice.

and the expected reward of super arm S can be written as

$$r(S,\boldsymbol{p}) = \sum_{j=1}^{W}\left(1-\prod_{i=1}^{V}(1-p_{(i,j)}^{S}p_{i,j})\right)$$

for which Assumptions 3 and 4 hold when $B=1$ and $B'=1$ respectively.

C. Influence Maximization Bandits

In the influence maximization problem with the independent cascade model [21], the learner is given a directed graph denoted by $G=(V,E)$, where V is the set of nodes and E is the set of edges. The learner selects and triggers a set of nodes $S \subseteq V$ such that $|S|=K$, where K is one of the problem parameters. This is the first iteration of a diffusion process. In each subsequent iteration, a node i that was triggered in the previous iteration might trigger another node j that is not triggered yet if j is adjacent to one of its outgoing edges. This happens with probability $p_{i,j}$ independently from the states of all other nodes. The diffusion process ends when no new node triggers in an iteration. The goal of the learner is to maximize—through the initial decision of nodes—the number of triggered nodes at the end of the diffusion process.

The problem can be modeled as a CMAB problem with PTAs, where base arms are edges $(i,j) \in E$ and super arms are the set of all edges (i,j) such that $i \in S$.[3] Assumption 3 holds as proven in Lemma 6 in [20] and Assumption 4 holds as proven in Lemma 2 in [22].

V. Combinatorial Thompson Sampling

CTS is a Bayesian algorithm that selects super arms by sampling from posterior distributions of base arms. Its pseudocode is given in Algorithm 1. We assume that the learner has access to an exact computation oracle, which takes as input an m-dimensional parameter vector $\boldsymbol{\theta}$ and the problem structure $([m],\mathcal{I},D^{\text{trig}},R)$, and outputs a super arm, denoted by $\text{Oracle}(\boldsymbol{\theta})$ such that $\text{Oracle}(\boldsymbol{\theta}) \in \text{OPT}(\boldsymbol{\theta})$. CTS keeps a Beta posterior over the mean outcome of each base arm. At the beginning of round t, for each base arm i it draws a sample $\theta_i(t)$ from its posterior distribution. Then, it forms the parameter vector in round t as $\boldsymbol{\theta}(t) := (\theta_1(t),\ldots,\theta_m(t))$, gives it to the exact computational oracle, and selects the super arm $S(t) = \text{Oracle}(\boldsymbol{\theta}(t))$. At the end of the round, CTS updates the posterior distributions of the triggered base arms using the observation $Q(t)$.

Algorithm 1 Combinatorial Thompson Sampling (CTS)

For each base arm i, let $a_i=1$, $b_i=1$
for $t=1,2,\ldots$ **do**
  For each base arm i, draw a sample $\theta_i(t)$ from Beta distribution $\beta(a_i,b_i)$; let $\boldsymbol{\theta}(t) := (\theta_1(t),\ldots,\theta_m(t))$
  Select super arm $S(t) = \text{Oracle}(\boldsymbol{\theta}(t))$, get the observation $Q(t)$
  for all $(i,X_i) \in Q(t)$ **do**
    $Y_i \leftarrow 1$ with probability X_i, 0 with probability $1-X_i$
    $a_i \leftarrow a_i + Y_i$
    $b_i \leftarrow b_i + (1-Y_i)$
  end for
end for

A. Regret of CTS Under Lipchitz Continuity

Theorem 1: Under Assumptions 1, 2, and 3, for all D, the regret of CTS by round T is bounded as

$$\text{Reg}(T) \le \sum_{i=1}^{m}\max_{S\in\mathcal{I}-\text{OPT}:i\in\tilde{S}}\frac{16B^2|\tilde{S}|\log T}{(1-\rho)p_i(\Delta_S-2B(\tilde{k}^{*2}+2)\varepsilon)} + \left(3+\frac{\tilde{K}^2}{(1-\rho)p^*\varepsilon^2}+\frac{2\mathbb{I}\{p^*<1\}}{\rho^2p^*}\right)m\Delta_{\max} + \alpha\frac{8\tilde{k}^*}{p^*\varepsilon^2}\left(\frac{4}{\varepsilon^2}+1\right)^{\tilde{k}^*}\log\frac{\tilde{k}^*}{\varepsilon^2}\Delta_{\max}$$

for all $\rho \in (0,1)$, and for all $\varepsilon \in (0,1/\sqrt{e}]$ such that $\forall S \in \mathcal{I}-OPT, \Delta_S > 2B(\tilde{k}^{*2}+2)\varepsilon$, where B is the Lipschitz constant in Assumption 3, $\alpha > 0$ is a problem independent constant that is also independent of T, and $\tilde{K} := \max_{S\in\mathcal{I}}|\tilde{S}|$ is the maximum triggering set size among all super arms.

We compare the result in Theorem 1 with [20], which shows that the regret of CUCB is $O(\sum_{i\in[m]}\log T/(p_i\Delta_i))$ given an ℓ_∞ bounded smoothness condition on the expected reward function and a bounded smoothness function of $f(x)=\gamma x$. When ε is sufficiently small, the regret bound in Theorem 1 is asymptotically equivalent to the regret bound for CUCB (in terms of the dependence on T, p_i, and Δ_i for $i \in [m]$). For the case with $p^*=1$ (no probabilistic triggering), the regret bound in Theorem 1 matches with the regret bound in Theorem 1 in [28] (in terms of the dependence on T and Δ_i for $i \in [m]$).

As final remarks, it is shown in Theorem 3 in [22] that the $1/p_i$ factor that multiplies the $\log T$ term is unavoidable in general. Moreover, regarding the exponential term $(4/\varepsilon^2+1)^{\tilde{k}^*}$, it is shown in Theorem 3 in [28] that there is at least one instance of CMAB (hence, also an instance of CMAB-PTA) where the regret of CTS is $\Omega(2^{k^*})$. Intuitively, such an exponential term is unavoidable since for CTS to select an optimal super arm that can trigger $\tilde{k}^*$ base arms, all of the samples from those $\tilde{k}^*$ base arms should independently be close to their true means. The proof of Theorem 1 is given in the supplemental document. It can also be found in the conference version of the paper [1].

B. Bayesian Regret of CTS Under Lipchitz Continuity

Theorem 2: Under Assumptions 1, 2, 3, and 5, when averaged over D, the Bayesian regret of CTS by round T is bounded as

$$\text{BayReg}(T) \le 4mB\sqrt{\frac{T(2+6\log T)}{(1-\rho)}}\mathbb{E}_{\boldsymbol{\mu}}\left[\sqrt{\frac{1}{p^*}}\right] + 8m^2B\left(1+\frac{1}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{p^*}\right]\right)$$

for all $\rho \in (0,1)$, where B is the Lipschitz constant in Assumption 3.

[3] This is equivalent to defining the super arm as S itself.

As mentioned in Section III-E, the Bayesian regret bound in Theorem 2 can be interpreted as a gap-free regret bound for CTS that holds asymptotically.

C. Bayesian Regret of CTS Under the TPM Lipchitz Continuity

Theorem 3: Under Assumptions 1, 2, 4, and 5, when averaged over D, the Bayesian regret of CTS by round T is bounded as

$$\begin{aligned}\mathrm{BayReg}(T) \leq\ & 16mB'(1+\sqrt{2})\sqrt{(1+4\log T)T} \\ & + 4mB' + 8m^2B'\end{aligned}$$

where B' is the Lipschitz constant in Assumption 4.

We improve the Bayesian regret bound in Theorem 2 under the stricter TPM Lipchitz continuity assumption and obtain a regret bound that is completely-free of triggering probabilities. Similar to Theorem 2, the Bayesian regret bound in Theorem 3 can be interpreted as an asymptotic regret bound for CTS.

D. Regret of CTS for Strictly Positive Triggering Probabilities

We improve the regret bound in Theorem 1 when all triggering probabilities are strictly positive.

Theorem 4: Under Assumptions 1, 2, and 3, for all D such that $\forall i \in [m], S \in \mathcal{I}, p_i^{D,S} \geq p^* > 0$, the regret of CTS by round T is bounded as

$$\begin{aligned}\mathrm{Reg}(T) \leq \max\Bigg\{ & 16mB\sqrt{\frac{e}{(1-\rho)p^*}}, \\ & \max_{S\in\mathcal{I}-\mathrm{OPT}}\Bigg\{\frac{128mB^2|\tilde{S}|}{(1-\rho)p^*(\Delta_S - 2B(\tilde{k}^{*2}+2)\varepsilon)} \\ & \times\ \log\frac{4B|\tilde{S}|}{(1-\rho)p^*(\Delta_S - 2B(\tilde{k}^{*2}+2)\varepsilon)}\Bigg\}\Bigg\} \\ & + \left(5 + \frac{\tilde{K}^2}{(1-\rho)p^*\varepsilon^2} + \frac{2\mathbb{I}\{p^*<1\}}{\rho^2 p^*}\right) m\Delta_{\max} \\ & + \alpha\frac{8\tilde{k}^*}{p^*\varepsilon^2}\left(\frac{4}{\varepsilon^2}+1\right)^{\tilde{k}^*}\log\frac{\tilde{k}^*}{\varepsilon^2}\Delta_{\max}\end{aligned}$$

for all $\rho \in (0,1)$, and for all $\varepsilon \in (0, 1/\sqrt{e}]$ such that $\forall S \in \mathcal{I} - OPT$, $\Delta_S > 2B(\tilde{k}^{*2}+2)\varepsilon$, where B is the Lipschitz constant in Assumption 3, $\alpha > 0$ is a problem independent constant that is also independent of T, and $\tilde{K} := \max_{S\in\mathcal{I}}|\tilde{S}|$ is the maximum triggering set size among all super arms.

Note that having all triggering probabilities be strictly positive makes the exploration aspect of the MAB problem trivial. No matter which actions the learner takes, all base arms provide occasional feedback. As a result of this, the upper bound for the expected regret becomes independent of the time horizon T. We compare the result of Theorem 4 with [30], which shows a similar bound for CTS in the exact same setting. While the bound in [30] is on order $O((1/p^*)^4)$ with respect to p^*, the bound in Theorem 4 is on order $O(1/p^*\log(1/p^*))$.

As a final remark, we observe that the regret bound in Theorem 4 does not match the lower bound on order $\Omega(\log(1/p^*))$ given in Theorem 1 in [25] proven for a special case of our setting, where rewards only depend on the selected arm. Assumptions 3 and 4, on the other hand, allow rewards to depend on all arms in the triggering set of the selected super arm either independent of or proportionally to their triggering probabilities. Considering how the reward model in [25] satisfies both Assumption 3 and Assumption 4 and how Assumption 4 is necesary to get rid of the $1/p^*$ terms in the previously discussed upper bounds, showing an upper bound on order $O(\log(1/p^*))$ instead of order $O(1/p^*\log(1/p^*))$ for the case with strictly positive triggering probabilities might only be possible under Assumption 4. The proof of Theorem 4 is given in the supplemental document.

VI. Proof of Theorem 2

We extend the proof technique used in [12] to CMAB-PTA. The technique relies on Fact 1, which establishes a relationship between Thompson sampling and upper confidence sequences commonly encountered in UCB-based analyses. According to Fact 1, the Bayesian regret is bounded by the difference between the true rewards and an upper confidence bound for the estimated rewards of the selected super arm and the optimal super arm. We show that these differences either shrink quickly as sample size increases (for the selected super arm) or are less than zero (for the optimal super arm) with overwhelming probability.

A. Preliminaries

All equalities and inequalities concerning random variables hold with probability 1. The complement of set $\mathcal{S}$ is denoted by $\neg\mathcal{S}$. The indicator function is given as $\mathbb{I}\{\cdot\}$. $M_i(t) := \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in \tilde{S}(\tau)\}$ denotes the number of times base arm i is tried to be triggered (i.e. it was in the triggering set of the selected super arm) until round t, $N_i(t) := \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in S'(\tau)\}$ denotes the number of times base arm i is triggered until round t, and $\hat{\mu}_i(t) := \sum_{\tau:\tau<t, i\in S'(\tau)} Y_i(\tau)/N_i(t)$ denotes the empirical mean outcome of base arm i at the start of round t, where $Y_i(t)$ is the Bernoulli random variable with mean $X_i(t)$ that is used for updating the posterior distribution that corresponds to base arm i in CTS.

Given a particular base arm $i \in [m]$, let τ_w^i be the round for which base arm i is in the triggering set $\tilde{S}(t)$ of the selected super arm $S(t)$ for the wth time and let $\tau_0^i = 0$. Note that we have $i \in \tilde{S}(\tau_{w+1}^i)$ and $M_i(\tau_{w+1}^i) = w$ for all $w \geq 0$. In order to decompose the regret, we make use of an upper confidence bound sequence $U(S,t) := r(S, \bar{\boldsymbol{\mu}}(t))$ for the reward of super arm $S \in \mathcal{I}$ at round t, where $\bar{\boldsymbol{\mu}}(t) = (\bar{\mu}_1(t), \ldots, \bar{\mu}_m(t))$ and

$$\bar{\mu}_i(t) = \hat{\mu}_i(t) + \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\ .$$

We also make use of the following events:

$$\begin{aligned}\mathcal{G}_i(t) &:= \left\{|\hat{\mu}_i(t) - \mu_i| > \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right\} \\ \mathcal{G}(t) &:= \{\exists i \in [m] : \mathcal{G}_i(t)\} \\ \mathcal{H}_i(t) &:= \{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\} \\ \mathcal{H}(t) &:= \{\exists i \in [m] : \mathcal{H}_i(t)\}\ .\end{aligned}$$

B. Facts and Lemmas

Fact 1: (Proposition 1 in [12]) For any upper confidence bound sequence $U(S,t)$,

$$\begin{aligned}\mathrm{BayReg}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(U(S(t),t) - r(S(t),\boldsymbol{\mu}))\right] \\ + \mathbb{E}\left[\sum_{t=1}^{T}(r(S^*,\boldsymbol{\mu}) - U(S^*,t))\right] .\end{aligned}$$

Proof: Since $\boldsymbol{\theta}(t)$ is sampled from the posterior distribution of $\boldsymbol{\mu}$ given observation history $\mathcal{F}_{t-1}$, $S(t) = \mathrm{Oracle}(\boldsymbol{\theta}(t))$ and $S^* = \mathrm{Oracle}(\boldsymbol{\mu})$ follow the same distribution when conditioned on $\mathcal{F}_{t-1}$. Together with the fact that $U(S,t)$ is a deterministic function when conditioned on $\mathcal{F}_{t-1}$, we have

$$\begin{aligned}&\mathbb{E}[r(S^*,\boldsymbol{\mu}) - r(S(t),\boldsymbol{\mu})] \\ &= \mathbb{E}[\mathbb{E}[r(S^*,\boldsymbol{\mu}) - r(S(t),\boldsymbol{\mu})|\mathcal{F}_{t-1}]] \\ &= \mathbb{E}[\mathbb{E}[r(S^*,\boldsymbol{\mu}) - U(S^*,t) + U(S^*,t) - r(S(t),\boldsymbol{\mu})|\mathcal{F}_{t-1}]] \\ &= \mathbb{E}[\mathbb{E}[r(S^*,\boldsymbol{\mu}) - U(S^*,t) + U(S(t),t) - r(S(t),\boldsymbol{\mu})|\mathcal{F}_{t-1}]] \\ &= \mathbb{E}[r(S^*,\boldsymbol{\mu}) - U(S^*,t)] + \mathbb{E}[U(S(t),t) - r(S(t),\boldsymbol{\mu})] .\end{aligned}$$

for all $t \in [T]$. □

Fact 2: (Lemma 1 in [12])

$$\mathbb{P}\left(\bigcup_{t=1}^{T}\left\{|\hat{\mu}_i(t) - \mu_i| > \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right\}\right) \le \frac{1}{T}$$

Fact 3: (Multiplicative Chernoff bound [20], [44]) Let $X_1,\ldots,X_n$ be Bernoulli random variables taking values in $\{0,1\}$ such that $\mathbb{E}[X_t|X_1,\ldots,X_{t-1}] \ge \mu$ for all $t \le n$, and $Y = X_1 + \cdots + X_n$. Then, for all $\delta \in (0,1)$,

$$\mathbb{P}(Y \le (1-\delta)\mu n) \le e^{-\frac{\delta^2\mu n}{2}} .$$

Lemma 1: We have

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{i \in \tilde{S}(t), N_i(t) \le (1-\rho)p_i M_i(t)\}\middle|\boldsymbol{\mu}\right] \le 1 + \frac{2}{\rho^2 p^*}$$

for all $i \in [m]$, $\boldsymbol{\mu} \in [0,1]^m$, and $\rho \in (0,1)$.

Proof:

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{i \in \tilde{S}(t), N_i(t) \le (1-\rho)p_i M_i(t)\}\middle|\boldsymbol{\mu}\right] \\ &\le \mathbb{E}\left[\sum_{w=0}^{T}\sum_{t=\tau_w^i+1}^{\tau_{w+1}^i}\mathbb{I}\{i \in \tilde{S}(t), \right. \\ &\qquad\qquad \left. N_i(t) \le (1-\rho)p_i M_i(t)\}\middle|\boldsymbol{\mu}\right] \\ &\le \mathbb{E}\left[\sum_{w=0}^{T}\mathbb{I}\{N_i(\tau_{w+1}^i) \le (1-\rho)p_i M_i(\tau_{w+1}^i)\}\middle|\boldsymbol{\mu}\right] \\ &\le 1 + \sum_{w=1}^{T}\mathbb{P}(N_i(\tau_{w+1}^i) \le (1-\rho)p_i M_i(\tau_{w+1}^i)|\boldsymbol{\mu}) \\ &\le 1 + \sum_{w=1}^{T} e^{-\frac{\rho^2 p^* w}{2}} \\ &\le 1 + \frac{2}{\rho^2 p^*} \qquad (2)\end{aligned}$$

where (2) is due to Fact 3. □

C. Main Part of the Proof

We decompose the Bayesian regret as

$$\begin{aligned}&\mathrm{BayReg}(T) \\ &= \mathbb{E}\left[\sum_{t=1}^{T}(r(S(t),\bar{\boldsymbol{\mu}}(t)) - r(S(t),\boldsymbol{\mu}))\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}(r(S^*,\boldsymbol{\mu}) - r(S^*,\bar{\boldsymbol{\mu}}(t)))\right] \qquad (3) \\ &\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\neg\mathcal{G}(t),\neg\mathcal{H}(t)\}(r(S(t),\bar{\boldsymbol{\mu}}(t)) - r(S(t),\boldsymbol{\mu}))\right] \qquad (4) \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\neg\mathcal{G}(t),\neg\mathcal{H}(t)\}(r(S^*,\boldsymbol{\mu}) - r(S^*,\bar{\boldsymbol{\mu}}(t)))\right] \qquad (5) \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}(t)\vee\mathcal{H}(t)\}\right] \times 4mB , \qquad (6)\end{aligned}$$

where (3) is due to Fact 1, and (6) is obtained by observing

$$\begin{aligned}&|r(S,\boldsymbol{\mu}) - r(S,\bar{\boldsymbol{\mu}}(t))| \\ &\le B\sum_{i\in\tilde{S}}\left|\mu_i - \hat{\mu}_i(t) - \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right| \\ &\le B\sum_{i\in\tilde{S}}|\mu_i - \hat{\mu}_i(t)| \\ &\quad + B\sum_{i\in\tilde{S}}\min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \\ &\le 2mB\end{aligned}$$

for all $S \in \mathcal{I}$.

1) Bounding (4)*:* When $\neg\mathcal{G}(t)$ and $\neg\mathcal{H}(t)$ hold, we have

$$\begin{aligned}&r(S(t),\bar{\boldsymbol{\mu}}(t)) - r(S(t),\boldsymbol{\mu}) \\ &\le B\sum_{i\in\tilde{S}(t)}\left|\mu_i - \hat{\mu}_i(t) - \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right| \\ &\le B\sum_{i\in\tilde{S}(t)}|\mu_i - \hat{\mu}_i(t)| \\ &\quad + B\sum_{i\in\tilde{S}(t)}\min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \\ &\le 2B\sum_{i\in\tilde{S}(t)}\min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \qquad (7) \\ &\le 2B\sum_{i\in\tilde{S}(t)}\min\left\{1, \sqrt{\frac{2+6\log T}{(1-\rho)p_i M_i(t)}}\right\} , \qquad (8)\end{aligned}$$

where (7) is due to $\neg\mathcal{G}(t)$ and (8) is due to $\neg\mathcal{H}(t)$. Then,

$$\begin{aligned}(4) &\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\neg\mathcal{H}(t)\}2B\sum_{i\in\tilde{S}(t)}\sqrt{\frac{2+6\log T}{(1-\rho)p_i M_i(t)}}\right] \\ &\le \mathbb{E}\left[\sum_{i=1}^{m}\sum_{w=0}^{T}\sum_{t=\tau_w^i+1}^{\tau_{w+1}^i}\mathbb{I}\{i \in \tilde{S}(t), \neg\mathcal{H}(t)\}\right.\end{aligned}$$

$$\times 2B\sqrt{\frac{2+6\log T}{(1-\rho)p_i M_i(t)}}\Bigg]$$

$$\leq \mathbb{E}\left[\sum_{i=1}^{m}\sum_{w=0}^{T}\mathbb{I}\{\neg\mathcal{H}(\tau_{w+1}^i)\}2B\sqrt{\frac{2+6\log T}{(1-\rho)p_i M_i(\tau_{w+1}^i)}}\right]$$

$$\leq \mathbb{E}\left[\sum_{i=1}^{m}\sum_{w=1}^{T}2B\sqrt{\frac{2+6\log T}{(1-\rho)p_i M_i(\tau_{w+1}^i)}}\right] \quad (9)$$

$$\leq \sum_{i=1}^{m}\sum_{w=1}^{T}2B\sqrt{\frac{2+6\log T}{(1-\rho)w}}\mathbb{E}_{\boldsymbol{\mu}}\left[\sqrt{\frac{1}{p^*}}\right]$$

$$\leq 4mB\sqrt{\frac{T(2+6\log T)}{(1-\rho)}}\mathbb{E}_{\boldsymbol{\mu}}\left[\sqrt{\frac{1}{p^*}}\right] , \quad (10)$$

where (9) holds since $N_i(\tau_1^i) = M_i(\tau_1^i) = 0$ implies $\mathcal{H}(\tau_1^i)$ and (10) holds since $\sum_{n=1}^{N}\sqrt{1/n} \leq 2\sqrt{N}$.

2) Bounding (5)*:* When $\neg\mathcal{G}(t)$ holds, we have

$$\mu_i \leq \hat{\mu}_i(t) + \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} = \bar{\mu}(t)$$

for all $i \in [m]$. Then,

$$r(S^*,\boldsymbol{\mu}) - r(S^*,\bar{\boldsymbol{\mu}}(t)) \leq r(S^*,\bar{\boldsymbol{\mu}}(t)) - r(S^*,\bar{\boldsymbol{\mu}}(t)) \quad (11)$$
$$= 0 ,$$

where (11) is due to Assumption 5. Hence, (5) ≤ 0.

3) Bounding (6)*:* We have

$$(6) \leq 4mB\sum_{i=1}^{m}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}_i(t)\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{H}_i(t)\}\right]\right)$$
$$\leq 4mB\sum_{i=1}^{m}\left(T\mathbb{P}\left(\bigcup_{t=1}^{T}\{\mathcal{G}_i(t)\}\right) + \mathbb{E}_{\boldsymbol{\mu}}\left[\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{H}_i(t)\}\middle|\boldsymbol{\mu}\right]\right]\right)$$
$$\leq 8m^2B\left(1 + \frac{1}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}\left[\frac{1}{p^*}\right]\right) , \quad (12)$$

where (12) is due to Fact 2 and Lemma 1 respectively for the two terms.

VII. Proof of Theorem 3

In order to take advantage of Assumption 4, we use the concept of triggering probability groups from [22]. However, the rest of our analysis is quite different from [22] and mainly follows the same technique we have followed in Section VI when proving Theorem 2.

A. Preliminaries

In addition to the preliminaries in Section VI-A for the proof of Theorem 2, we make the following definitions. For $j \in \mathbb{Z}_+$, let $\mathcal{I}_{i,j} := \{S \in \mathcal{I} : 2^{-j} < p_i^S \leq 2\cdot 2^{-j}\}$ denote the jth *triggering probability group* of base arm i and let j_i^S denote the index of the triggering probability group of base arm i that super arm S belongs to, i.e., j_i^S is such that $S \in \mathcal{I}_{i,j_i^S}$. We use these definitions to introduce the following counters: $M_{i,j}(t) := \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in \tilde{S}(\tau), S(\tau) \in \mathcal{I}_{i,j}\}$ and $N_{i,j}(t) := \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in S'(\tau), S(\tau) \in \mathcal{I}_{i,j}\}$. By definition, $M_i(t) = \sum_{j=1}^{\infty}M_{i,j}(t)$ and $N_i(t) = \sum_{j=1}^{\infty}N_{i,j}(t)$.

Given a particular base arm $i \in [m]$, let $\eta_w^{i,j}$ be the round for which base arm i is in the triggering set $\tilde{S}(t)$ of the selected super arm $S(t)$ and $S(t) \in \mathcal{I}_{i,j}$ for the wth time and let $\eta_0^{i,j} = 0$. Note that we have $i \in \tilde{S}(\eta_{w+1}^{i,j})$, $M_{i,j}(\eta_{w+1}^{i,j}) = w$, and $S(\eta_{w+1}^{i,j}) \in \mathcal{I}_{i,j}$ for all $w \geq 0$. We also make the following change to event $\mathcal{H}_i(t)$:

$$\mathcal{H}_i(t) := \left\{\max\{N_i(t), 8\log T\} \leq \frac{1}{2}\cdot 2^{-j_i^{S(t)}}M_{i,j_i^{S(t)}}(t)\right\} .$$

B. Facts and Lemmas

Lemma 2: Fix $i \in [m]$, $t \in [T]$ and $j \in \mathbb{Z}_+$. When CTS is run, we have

$$\mathbb{P}\left(\max\{N_i(t), 8\log T\} \leq \frac{1}{2}\cdot 2^{-j}M_{i,j}(t)\right) \leq \frac{1}{T} .$$

Proof:

$$\mathbb{P}\left(\max\{N_i(t), 8\log T\} \leq \frac{1}{2}\cdot 2^{-j}M_{i,j}(t)\right)$$
$$\leq \mathbb{P}\left(N_i(t) \leq \frac{1}{2}\cdot 2^{-j}M_{i,j}(t)\middle| 8\log T \leq \frac{1}{2}\cdot 2^{-j}M_{i,j}(t)\right)$$
$$\leq \sum_{w=0}^{T-1}\mathbb{I}\left(8\log T \leq \frac{1}{2}\cdot 2^{-j}w\right) \times\mathbb{P}\left(N_{i,j}(t) \leq \frac{1}{2}\cdot 2^{-j}w\middle| M_{i,j}(t) = w\right)$$
$$\leq \sum_{w=0}^{T-1}\mathbb{I}\left(8\log T \leq \frac{1}{2}\cdot 2^{-j}w\right)e^{-\frac{2^{-j}w}{8}} \quad (13)$$
$$\leq \sum_{w=0}^{T-1}e^{-2\log T} \quad (14)$$
$$\leq \frac{1}{T}$$

where (13) holds due to Fact 3 and (14) holds since $8\log T \leq 1/2\cdot 2^{-j}w$ implies that $e^{-2^{-j}w/8} \leq e^{-2\log T}$. □

C. Main Part of the Proof

We decompose the Bayesian regret the same way as we did in Section VI-C. Note that (6) still holds since

$$|r(S,\boldsymbol{\mu}) - r(S,\bar{\boldsymbol{\mu}}(t))|$$
$$\leq B'\sum_{i\in\tilde{S}}p_i^S\left|\mu_i - \hat{\mu}_i(t) - \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right|$$
$$\leq B'\sum_{i\in\tilde{S}}p_i^S|\mu_i - \hat{\mu}_i(t)| + B'\sum_{i\in\tilde{S}}p_i^S\min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}$$
$$\leq 2mB'$$

for all $S \in \mathcal{I}$.

1) Bounding (4)*:* When $\neg\mathcal{H}(t)$ holds, one of the following must be the case:

$$\begin{aligned} 8\log T &\geq \frac{1}{2}\cdot 2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t) \\ &\Longrightarrow 1 \leq \sqrt{\frac{16\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}} \\ &\leq \sqrt{\frac{4+16\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}}, \\ N_i(t) &\geq \frac{1}{2}\cdot 2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t) \\ &\Longrightarrow \sqrt{\frac{2+6\log T}{N_i(t)}} \leq \sqrt{\frac{4+12\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}} \\ &\leq \sqrt{\frac{4+16\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}}. \end{aligned}$$

Combining the two result together, we obtain

$$\min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \leq \sqrt{\frac{4+16\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}}. \quad (15)$$

When $\neg\mathcal{G}(t)$ also holds, we have

$$\begin{aligned} &r(S(t), \bar{\boldsymbol{\mu}}(t)) - r(S(t), \boldsymbol{\mu}) \\ &\leq B' \sum_{i\in\tilde{S}(t)} p_i^{S(t)} \left|\mu_i - \bar{\mu}_i(t) - \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\}\right| \\ &\leq B' \sum_{i\in\tilde{S}(t)} p_i^{S(t)} |\mu_i - \bar{\mu}_i(t)| \\ &\quad + B' \sum_{i\in\tilde{S}(t)} p_i^{S(t)} \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \\ &\leq 2B' \sum_{i\in\tilde{S}(t)} p_i^{S(t)} \min\left\{1, \sqrt{\frac{2+6\log T}{N_i(t)}}\right\} \quad (16) \\ &\leq 2B' \sum_{i\in\tilde{S}(t)} p_i^{S(t)} \min\left\{1, \sqrt{\frac{4+16\log T}{2^{-j_i^{S(t)}} M_{i,j_i^{S(t)}}(t)}}\right\} \quad (17) \\ &= 4B' \sum_{i\in\tilde{S}(t)} \min\left\{2^{-j_i^{S(t)}}, \sqrt{\frac{(4+16\log T)2^{-j_i^{S(t)}}}{M_{i,j_i^{S(t)}}(t)}}\right\}, \quad (18) \end{aligned}$$

where (16) is due to $\neg\mathcal{G}(t)$, (17) is due to (15), and (18) holds since $p_i^{S(t)} \leq 2\cdot 2^{-j_i^{S(t)}}$. Then,

$$\begin{aligned} &(4) \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} 4B' \sum_{i\in\tilde{S}(t)} \min\left\{2^{-j_i^{S(t)}}, \sqrt{\frac{(4+16\log T)2^{-j_i^{S(t)}}}{M_{i,j_i^{S(t)}}(t)}}\right\}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{\infty}\sum_{w=0}^{T}\sum_{t=\eta_w^{i,j}+1}^{\eta_{w+1}^{i,j}} \mathbb{I}\{i\in\tilde{S}(t), S(t)\in\mathcal{I}_{i,j}\}\right. \\ &\quad \left.\times 4B' \min\left\{2^{-j}, \sqrt{\frac{(4+16\log T)2^{-j}}{M_{i,j}(t)}}\right\}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{\infty}\sum_{w=0}^{T} 4B' \min\left\{2^{-j}, \sqrt{\frac{(4+16\log T)2^{-j}}{M_{i,j}(\eta_{w+1}^{i,j})}}\right\}\right] \\ &\leq \sum_{i=1}^{m}\sum_{j=1}^{\infty}\left(4B'\cdot 2^{-j} + \sum_{w=1}^{T} 4B'\sqrt{\frac{(4+16\log T)2^{-j}}{w}}\right) \\ &\leq 4mB' + 16mB'\sum_{j=1}^{\infty}\sqrt{(1+4\log T)T\cdot 2^{-j}} \quad (19) \\ &\leq 4mB' + 16mB'(1+\sqrt{2})\sqrt{(1+4\log T)T}, \end{aligned}$$

where (19) holds since $\sum_{n=1}^{N}\sqrt{1/n} \leq 2\sqrt{N}$.

2) Bounding (5)*:* We bound (5) the same way we did in Section VI-C.2.

3) Bounding (6)*:* We have

$$\begin{aligned} &(6) \\ &\leq 4mB' \sum_{i=1}^{m}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}_i(t)\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{H}_i(t)\}\right]\right) \\ &\leq 4mB' \sum_{i=1}^{m}\left(T\mathbb{P}\left(\bigcup_{t=1}^{T}\{\mathcal{G}_i(t)\}\right) + \sum_{t=1}^{T}\mathbb{P}(\mathcal{H}_i(t))\right) \\ &\leq 4mB' \sum_{i=1}^{m}\left(1 + \sum_{t=1}^{T}\sum_{j=1}^{\infty}\mathbb{P}(\mathcal{H}_i(t)|j_i^{S(t)}=j)\mathbb{P}(j_i^{S(t)}=j)\right) \quad (20) \\ &\leq 4mB' \sum_{i=1}^{m}\left(1 + \sum_{t=1}^{T}\sum_{j=1}^{\infty}\frac{1}{T}\mathbb{P}(j_i^{S(t)}=j)\right) \quad (21) \\ &\leq 8m^2B', \end{aligned}$$

where (20) is due to Fact 2 and (21) is due to Lemma 2.

VIII. Numerical Results

In this section, we compare CTS with other state-of-the-art CMAB algorithms in three different applications: cascading bandits, probabilistic maximum coverage bandits, and influence maximization bandits introduced in Section IV. We compare the performance of CTS with CUCB in [20] in all settings. For the first two problems, we assume that all algorithms have access to an exact computation oracle that computes the estimated optimal super arm in each round. On the other hand, for the third problem, we assume that all algorithms use an approximation oracle. For cascading bandits only, we also compare CTS with algorithms specifically designed for this setting: CascadeKL-UCB in [4] and TS-Cascade in [26]. The former uses the principle of optimism under the face of uncertainty to compute Kullback-Leibler divergence based UCBs while the latter uses Thompson sampling with Gaussian posterior over the base arms.

A. Cascading Bandits

We consider the disjunctive case with $V = 100$, $W = 20$ and $K = 5$, and generate $p_{i,j}$s by sampling uniformly at

TABLE II
REGRETS OF CTS AND CUCB WITH THEIR STANDARD DEVIATIONS FOR VARIOUS PROBLEM INSTANCES

V	K	Δ	**CTS**	**CUCB**	**CascadeUCB1**	**CascadeKL-UCB**	**TS-Cascade**
16	2	0.15	155.4 ± 14.1	1284.1 ± 52.4	1300.6 ± 46.8	360.6 ± 23.4	381.1 ± 16.8
16	4	0.15	103.2 ± 9.0	998.9 ± 33.2	993.6 ± 32.8	267.3 ± 20.6	281.0 ± 11.8
16	8	0.15	52.1 ± 9.8	549.5 ± 16.8	546.4 ± 11.7	150.3 ± 15.6	137.9 ± 8.8
32	2	0.15	321.4 ± 18.9	2718.8 ± 61.2	2676.4 ± 59.4	749.2 ± 34.2	752.9 ± 49.9
32	4	0.15	252.2 ± 17.0	2227.0 ± 55.4	2232.1 ± 46.6	617.4 ± 39.9	612.3 ± 15.2
32	8	0.15	155.4 ± 25.7	1531.0 ± 21.9	1525.4 ± 30.0	420.6 ± 27.5	385.0 ± 16.3
16	2	0.075	276.9 ± 50.7	2057.6 ± 79.6	2065.4 ± 87.4	709.0 ± 60.4	688.3 ± 78.5
16	4	0.075	205.4 ± 25.7	1496.5 ± 65.2	1512.4 ± 87.0	546.3 ± 53.5	557.9 ± 45.0
16	8	0.075	113.1 ± 40.4	719.4 ± 53.7	717.5 ± 44.2	266.1 ± 32.4	273.8 ± 30.7

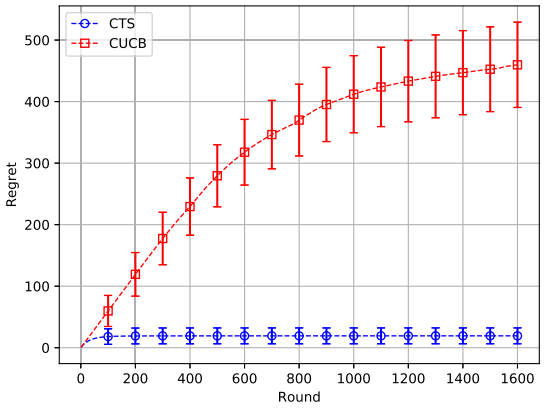

Fig. 1. Regrets of CTS and CUCB for the disjunctive cascading bandit problem.

random from $[0, 1]$. We run both CTS and CUCB for 1600 rounds, and report their regrets averaged over 1000 runs in Fig. 1, where error bars represent the standard deviation of the regret (multiplied by 10 for visibility). In this setting CTS significantly outperforms CUCB by achieving a final regret that is no more than 5% of the final regret of CUCB. Relatively bad performance of CUCB can be explained by excessive number of explorations due to the UCBs that stay high for a large number of rounds.

We also consider the same class of problems $B_{\mathrm{LB}}(V, K, p, \Delta)$ as in [4], where $W = 1$ and the probability that the user finds page j attractive is given as

$$p_{1,j} = \begin{cases} p & \text{if } j \leq K \\ p - \Delta & \text{otherwise .} \end{cases}$$

Similar to [4], we set $p = 0.2$ and vary other parameters, namely V, K, and Δ. We run both CTS and CUCB for 100000 rounds in all problem instances, and report their regrets averaged over 20 runs in Table II.

In addition to CUCB, we compare CTS against CascadeUCB1 and CascadeKL-UCB given in [4], and TS-Cascade given in [26] as well. Note that regrets of CUCB and CascadeUCB1 matches very closely as two algorithms are essentially the same when CUCB is applied to cascading bandits except for some minor differences in the initialization stage and how UCBs larger than 1 are handled. We observe that CTS outperforms all other algorithms in all problem instances by achieving a regret that is at most 44% of the regret of all other algorithms. For CTS, we also see that the regret increases as the number of pages (V) increases, it decreases as the number of recommended items (K) increases, and it increases as Δ decreases, which are very similar to the major observations that are made in [4].

B. Probabilistic Maximum Coverage Bandits

Our experimental setup for this case is based on MovieLens dataset [45] as in [30]. [4] The dataset contains 20 million movie ratings that are assigned between January 1995 and March 2015. Out of this, we only use the ones that are assigned between March 2014 and March 2015. In the experiments, the recommender chooses $K = 3$ movies out of $V = 30$ movies, which include 10 of the most rated movies, 10 of the least rated movies and 10 randomly selected movies from the dataset. These 30 movies are rated by $W = 57369$ users.

In total, there are 20 genres in the dataset. Each movie belongs to at least one genre. We take genre information into account to define attraction probabilities. For this, we create a 20-dimensional vector $\boldsymbol{g_i}$ for each movie $i \in [V]$, where $g_{ik} = 1$ if the movie belongs to genre k and 0 otherwise. Using these vectors, we calculate a genre preference vector $\boldsymbol{u_j}$ for each user $j \in [W]$ as

$$\boldsymbol{u_j} = \frac{\sum_{i \in \mathcal{V}_j} \boldsymbol{g_i}}{|\mathcal{V}_j|} + \boldsymbol{\epsilon_j}$$

where $\mathcal{V}_j$ is the set of movies that user j rated and $\boldsymbol{\epsilon_j}$ is a random vector such that $\epsilon_{jk} = |\chi_{jk}|$ for $\chi_{jk} \sim \mathcal{N}(0, 0.05)$. The noise $\boldsymbol{\epsilon_j}$ is introduced to model exploratory behavior of the user. Finally, defining $\hat{\boldsymbol{g}}_{\boldsymbol{i}} = \boldsymbol{g_i}/\|\boldsymbol{g_i}\|$ and $\hat{\boldsymbol{u}}_{\boldsymbol{j}} = \boldsymbol{u_j}/\|\boldsymbol{u_j}\|$ as the normalized versions of the vectors we have defined, the attraction probabilities are calculated as

$$p_{i,j} = 0.2 \times \frac{\langle \hat{\boldsymbol{g}}_{\boldsymbol{i}}, \hat{\boldsymbol{u}}_{\boldsymbol{j}} \rangle r_i}{\max_{i \in [V]} r_i}$$

where r_i is the average rating of movie i.

We run both CTS and CUCB for 1000 rounds, and report their regrets averaged over 10 runs in Fig. 2, where error bars represent standard deviation of the regret (multiplied by 100 for visibility). We consider two cases with $p^* = 0.01$ and $p^* = 0.05$. For both cases, CTS significantly outperforms CUCB by achieving a final regret that is no more than 9% of the final regret of CUCB.

[4]While the probabilistic maximum coverage problem is NP-hard, here we focus on a small-scale problem and use an exact computation oracle.

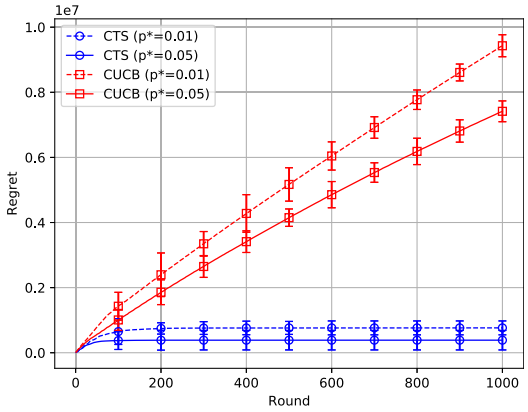

Fig. 2. Regrets of CTS and CUCB for the probabilistic maximum coverage bandit problem.

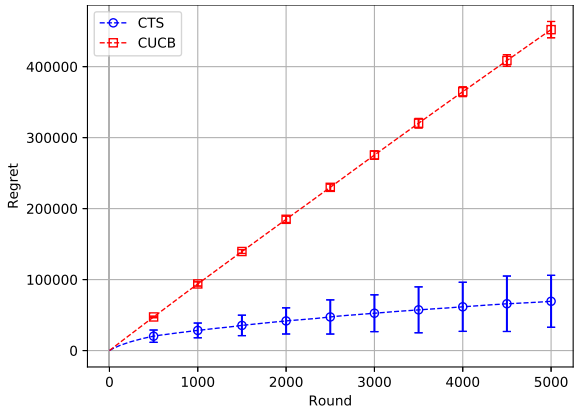

Fig. 3. Regrets of CTS and CUCB for the influence maximization bandit problem.

C. Influence Maximization Bandits

We consider a directed version of the Facebook network dataset [46] that consists of 15k edges and 3120 nodes. Since, the dataset does not contain influence probabilities, we artificially generate them by setting $p_{i,j} = 1/|\mathcal{V}_i|$ where $\mathcal{V}_i$ represents the set of outgoing neighbors of node i. We assume that in each round the learner selects a seed set of $K = 30$ nodes and this set forms the selected super arm. Moreover, we assume that the influence propagates—starting from the seed set—according to the independent cascade model [21], which is one of the most widely used influence propagation models. We adopt the edge-level feedback model in which the learner both observes the set of influenced nodes and the influence outcomes of the outgoing edges of these nodes.

Since the problem itself is NP-hard, an exact computation oracle is computationally infeasible for the given graph size. Nevertheless, many computationally efficient approximation algorithms exist for the influence maximization problem (see e.g., CELF in [47], and TIM and TIM+ in [48]). Due to its computational efficiency and good performance in practice, we set the learner to use TIM+ as the approximation oracle. When given as input an influence graph with n nodes and m edges, the influence probabilities on these edges and parameters ε and ℓ, TIM+ is guaranteed to return an $\alpha = (1-1/e-\varepsilon)$-approximate solution with probability at least $\beta = 1 - 3n^{-\ell}$ and with time complexity $O((K+\ell)(n+m)\log n/\varepsilon^2)$. For all experiments, we set $\varepsilon = 0.1$ and $\ell = 1$. Since the learner uses an approximation oracle, instead of the regret given in (1) we consider the (α, β)-approximation regret as given in [20] in the remainder of this section.

We run both CTS and CUCB for 5000 rounds and report their regrets averaged over 10 runs in Fig. 3. Here, error bars represent standard deviation of the regret multiplied by 10 for visibility. Note that in these simulations, we consider the realized regret of the learner's actions instead of the expected regret as we do in the other experiments. This is once again due to the complexity of the problem and the difficulty in calculating expected regret. Again, it is observed that CTS significantly outperforms CUCB by achieving a final regret that is no more than 16% of the final regret of CUCB. Relatively bad performance of CUCB is due to the fact that the considered time horizon is not long enough for CUCB to efficiently explore all base arms. It is observed that the UCBs of many base arms remain above 1 even at the end of 5000 rounds. As an algorithm that is based on the principle of optimism in the face of uncertainty, CUCB's performance completely depends on the confidence sets it uses to calculate the UCB indices, and this example shows that these confidence sets are not tight enough to guarantee fast convergence.

IX. Conclusion

We analyzed the regret of CTS for CMAB-PTA and proved (i) an order optimal gap-dependent regret bound when the expected reward function is Lipschitz continuous without assuming monotonicity, (ii) a Bayesian regret bound equivalent to an asymptotic gap-free regret bound assuming monotonicity, (iii) a Bayesian regret bound that is independent of triggering probabilities under the triggering modulated Lipschitz continuity assumption, and (iv) an improved regret bound that is independent of the time horizon for the special case when the triggering probabilities are strictly positive.

References

[1] A. Hüyük and C. Tekin, "Analysis of Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1322–1330.

[2] T. N. Dinh, H. Zhang, D. T. Nguyen, and M. T. Thai, "Cost-effective viral marketing for time-critical campaigns in large-scale social networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 6, pp. 2001–2011, Dec. 2014.

[3] G. Tong, W. Wu, S. Tang, and D.-Z. Du, "Adaptive influence maximization in dynamic social networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 112–125, Feb. 2017.

[4] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 767–776.

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

[6] S. Klos nee Muller, C. Tekin, M. van der Schaar, and A. Klein, "Context-aware hierarchical online learning for performance maximization in mobile crowdsourcing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1334–1347, Jun. 2018.

[7] H.-W. Lee, E. Modiano, and K. Lee, "Diverse routing in networks with probabilistic failures," *IEEE/ACM Trans. Netw.*, vol. 18, no. 6, pp. 1895–1907, Dec. 2010.

[8] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 55, pp. 527–535, 1952.

[9] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[10] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, Dec. 1933.
[11] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. 25th Annu. Conf. Learn. Theory*, 2012, pp. 39.1–39.26.
[12] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Math. Operations Res.*, vol. 39, no. 4, pp. 1221–1243, Nov. 2014.
[13] R. Agrawal, "Sample mean based index policies by O(log n) regret for the multi-armed bandit problem," *Adv. Appl. Probab.*, vol. 27, no. 4, pp. 1054–1078, Dec. 1995.
[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.
[15] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2011, pp. 2249–2257.
[16] S. L. Scott, "A modern Bayesian look at the multi-armed bandit," *Appl. Stochastic Models Bus. Ind.*, vol. 26, no. 6, pp. 639–658, Nov. 2010.
[17] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, Sep. 2012.
[18] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, 2015, pp. 535–543.
[19] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 151–159.
[20] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *J. Mach. Learn. Res.*, vol. 17, no. 50, pp. 1746–1778, Jan. 2016.
[21] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 137–146.
[22] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1161–1171.
[23] F. Liu, S. Buccapatnam, and N. Shroff, "Information directed sampling for stochastic bandits with graph feedback," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[24] S. Li, W. Chen, Z. Wen, and K.-S. Leung, "Stochastic online learning with probabilistic graph feedback," 2019, *arXiv:1903.01083*. [Online]. Available: http://arxiv.org/abs/1903.01083
[25] R. Degenne, E. Garcelon, and V. Perchet, "Bandits with side observations: Bounded vs. logarithmic regret," in *Proc. 34th Conf. Uncertainty Artif. Intell.*, 2018, pp. 467–476.
[26] W. C. Cheung, V. Tan, and Z. Zhong, "A Thompson sampling algorithm for cascading bandits," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 438–447.
[27] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Programm.*, vol. 14, no. 1, pp. 265–294, 1978.
[28] S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5114–5122.
[29] N. Merlis and S. Mannor, "Batch-size independent regret bounds for the combinatorial multi-armed bandit problem," in *Proc. 32nd Annu. Conf. Learn. Theory*, 2019, pp. 2465–2489.
[30] A. O. Saritac and C. Tekin, "Combinatorial multi-armed bandit with probabilistically triggered arms: A case with bounded regret," 2017, *arXiv:1707.07443*. [Online]. Available: http://arxiv.org/abs/1707.07443
[31] B. Kveton, Z. Wen, A. Ashkan, H. Eydgahi, and B. Eriksson, "Matroid bandits: Fast combinatorial optimization with learning," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, 2014, pp. 420–429.
[32] M. S. Talebi and A. Proutiere, "An optimal algorithm for stochastic matroid bandit optimization," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, 2016, pp. 548–556.
[33] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Combinatorial cascading bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1450–1458.
[34] S. Li, B. Wang, S. Zhang, and W. Chen, "Contextual combinatorial cascading bandits," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, Pp. 1245–1253.
[35] L. Qin, S. Chen, and X. Zhu, "Contextual combinatorial bandit and its application on diversified online recommendation," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 461–469.
[36] A. O. Saritac, A. Karakurt, and C. Tekin, "Online contextual influence maximization with costly observations," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 5, no. 2, pp. 273–289, Jun. 2019.
[37] J.-Y. Audibert, S. Bubeck, and G. Lugosi, "Regret in online combinatorial optimization," *Math. Oper. Res.*, vol. 39, no. 1, pp. 31–45, Feb. 2014.
[38] R. Combes, M. S. Talebi, A. Proutiere, and L. Marc, "Combinatorial bandits revisited," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2116–2124.
[39] K. A. Sankararaman and A. Slivkins, "Combinatorial semi-bandits with knapsacks," in *Proc. 21st Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1760–1770.
[40] M. Agarwal and V. Aggarwal, "Regret bounds for stochastic combinatorial multi-armed bandits with linear space complexity," 2018, *arXiv:1811.11925*. [Online]. Available: http://arxiv.org/abs/1811.11925
[41] I. Rejwan and Y. Mansour, "Top-k combinatorial bandits with full-bandit feedback," 2019, *arXiv:1905.12624*. [Online]. Available: http://arxiv.org/abs/1905.12624
[42] T. Lin, B. Abrahao, R. Kleinberg, J. Lui, and W. Chen, "Combinatorial partial monitoring game with linear feedback and its applications," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 901–909.
[43] A. Nika, S. Elahi, and C. Tekin, "Contextual combinatorial volatile multi-armed bandit with adaptive discretization," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 1486–1496.
[44] M. Mitzenmacher and E. Upfal, *Probab. Computing: Randomized Algorithms Probabilistic Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[45] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, p. 19, Jan. 2016.
[46] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 539–547.
[47] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.
[48] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 75–86.

Alihan Hüyük received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2019. He is currently pursuing the Ph.D. degree in applied mathematics and theoretical physics with the University of Cambridge, U.K. His research interests include multi-armed bandit problems, imitation learning, and inverse reinforcement learning.

Cem Tekin (Senior Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011, and 2013, respectively. From February 2013 to January 2015, he was a Post-Doctoral Scholar with the University of California, Los Angeles, CA, USA. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara. His research interests include multi-armed bandit problems, reinforcement learning, multiagent systems, and cognitive communications. He received the Fred W. Ellersick Award for the best paper in MILCOM 2009 and the Distinguished Young Scientist (BAGEP) Award of the Science Academy Association of Turkey in 2019.