Incentivizing Exploration

PETER FRAZIER, Cornell University, Ithaca NY
DAVID KEMPE, University of Southern California, Los Angeles CA
JON KLEINBERG, Cornell University, Ithaca NY
ROBERT KLEINBERG, Cornell University, Ithaca NY

We study a Bayesian multi-armed bandit (MAB) setting in which a principal seeks to maximize the sum of expected time-discounted rewards obtained by pulling arms, when the arms are actually pulled by selfish and myopic individuals. Since such individuals pull the arm with highest expected posterior reward (i.e., they always exploit and never explore), the principal must incentivize them to explore by offering suitable payments. Among others, this setting models crowdsourced information discovery and funding agencies incentivizing scientists to perform high-risk, high-reward research.

We explore the tradeoff between the principal's total expected time-discounted incentive payments, and the total time-discounted rewards realized. Specifically, with a time-discount factor $\gamma \in (0, 1)$, let OPT denote the total expected time-discounted reward achievable by a principal who pulls arms directly in a MAB problem, without having to incentivize selfish agents. We call a pair $(\rho, b) \in [0, 1]^2$ consisting of a reward ρ and payment b *achievable* if for every MAB instance, using expected time-discounted payments of at most $b \cdot \mathrm{OPT}$, the principal can guarantee an expected time-discounted reward of at least $\rho \cdot \mathrm{OPT}$. Our main result is an essentially complete characterization of achievable (payment, reward) pairs: if $\sqrt{b} + \sqrt{1-\rho} > \sqrt{\gamma}$, then (ρ, b) is achievable, and if $\sqrt{b} + \sqrt{1-\rho} < \sqrt{\gamma}$, then (ρ, b) is not achievable.

In proving this characterization, we analyze so-called *time-expanded* policies, which in each step let the agents choose myopically with some probability p, and incentivize them to choose "optimally" with probability $1 - p$. The analysis of time-expanded policies leads to a question that may be of independent interest: If the same MAB instance (without selfish agents) is considered under two different time-discount rates $\gamma > \eta$, how small can the ratio of OPT_η to OPT_γ be? We give a complete answer to this question, showing that $\mathrm{OPT}_\eta \geq \frac{(1-\gamma)^2}{(1-\eta)^2} \cdot \mathrm{OPT}_\gamma$, and that this bound is tight.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

Additional Key Words and Phrases: multi-armed bandit problems; crowdsourcing; exploration; incentives

1. INTRODUCTION

An important recent theme in the development of on-line social systems is the potential of crowdsourced effort to solve large problems — defining tasks in which many people can each contribute a small amount of time to the overall goal. In some cases, such as jobs on Amazon Mechanical Turk and similar crowdsourced work platforms, the arrangement is based on a direct compensation scheme, in which a (low) rate is paid for each unit of work performed. But in many settings, and in a large range of emerging applications, one only has access to a crowd "in the wild," as they go about their everyday activities.

Many of these unstructured crowdwork settings have the following basic structure: the designer of the system wants the members of the crowd to "explore" a space of options and learn from their observations and reactions; but each member of the crowd

Authors' emails: pf98@cornell.edu, dkempe@usc.edu, {kleinber, rdk}@cs.cornell.edu.

EC'14, June 8–12, 2014, Stanford, CA, USA. Copyright © 2014 ACM 978-1-4503-2565-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/http://dx.doi.org/10.1145/2600057.2602897

individually wants to focus on the “good” options rather than directly helping in this exploration process. This trade-off comes up in different guises across a surprisingly wide range of domains.

- *Crowdsourced information discovery.* Social news readers or similar sites, which promote stories or pages of interest to their readers, typically rely on readers themselves to discover and share interesting stories, and to rate stories of which they are aware. While individuals may a priori prefer to read stories that have already been rated highly by many others, to discover new material they should be incentivized to explore stories with few reviews.
- *Product Ratings.* Customers of online retailers rely heavily on other customers’ reviews in their choice of products. The retailer would like to see the customers converge to the best product, but individuals who buy and review only one product have an incentive to be myopic about their purchase. The retailer can implement incentives directly with discounts.
- *Citizen science.* Scientific organizations such as Galaxy Zoo for astronomy and eBird for Ornithology try to synthesize the activities of a large group of enthusiasts to make scientific observations at a collective scale [Lintott et al. 2008; Sheldon et al. 2007]. This involves guiding the enthusiasts toward relatively unexplored parts of the domain (e.g., either in the night sky or a natural bird habitat), whereas each individual enthusiast would rather focus on the areas where the star-gazing or bird-watching appears to be the best [Xue et al. 2013].
- *Funding of research efforts.* Viewed in this context, government funding of scientific research has a similar structure. While funding agencies such as NSF, NIH, or DARPA, functioning as arms of the government, may be able to define overall research agendas, the actual research is carried out by individual research groups. Since most groups typically pursue short-term rewards and projects that are very likely to succeed, funding agencies may wish to use grants to incentivize high-risk high-reward research efforts with potential for large gains in overall welfare.

In all of these domains, there is a fundamental incentive problem: the designer’s goal is to carry out exploration (of the space of news stories, products, bird habitats, or scientific research questions) as efficiently as possible, but for reasons of scale, they cannot perform this exploration on their own. Rather, they must implement the exploration via a crowd composed of members who each derive their own, different utility from participating in the exploration. The designer’s goal, at a high level, is to incentivize exploration in a way that increases overall welfare as much as possible, while simultaneously minimizing the amount of incentive transfer made to the crowd.

Perhaps the most natural framework for modeling this incentive problem is via the multi-armed bandit (MAB) problem, where it leads to a basic variant of the question that has received surprisingly little study. In typical MAB settings, multiple actions with unknown payoff distributions present themselves to a decision maker, who strives to choose a sequence of actions to maximize his total payoff over time. This entails repeatedly deciding between taking a sure payoff versus exploring higher-risk/higher-reward actions which might have higher long-term benefits.

We consider a variant of the MAB in which the *principal* (the crowdsourcing website, retailer, or citizen science organization) cannot choose directly which actions to pursue, and instead must rely on a stream of selfish and myopic agents. Without incentives, these agents will act myopically, failing to explore actions that might lead to large long-term rewards. To steer their efforts in a direction more fruitful for overall welfare, the principal can offer monetary (or other) rewards if they choose particular actions (pages or products to rate, or regions of the night sky to observe). In choosing an incentive structure, the principal must trade rewards and knowledge gained through exploration, against payments made to incentivize this exploration.

Here we give an informal overview of this model, and present it formally in Section 2. The model is based on the standard time-discounted infinite time-horizon Bayesian MAB model [Robbins 1952], and includes a finite set of arms i. Each arm has an associated payoff distribution, which is unknown to the algorithm, and drawn independently from a known distribution over distributions.[1] When an algorithm plays arm i at time t, it receives a random reward $r_{t,i}$, and may also use the observed reward to update its posterior belief about the arm's reward distribution. An algorithm chooses an arm to play for each discrete time step $t = 0, 1, 2, \ldots$, making adaptive choices based on observed payoffs. Time is discounted exponentially with a time discount factor γ; thus, when playing arm i_t at time t, the algorithm obtains expected reward $\gamma^t \mathbb{E}\left[r_{t,i_t}\right]$, where the expectation is taken over all sources of randomness (including the history-dependent choice of i_t). The total time-discounted reward of an algorithm $\mathcal{A}$ for the principal is then $\sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[R_{t,i_t}\right]$.

When interacting with selfish and myopic agents, the algorithm can offer them payments for playing certain arms, which may also depend on the history of observed rewards. A selfish and myopic agent chooses the arm maximizing the sum of the offered arm-specific payment and the expected posterior reward of the arm itself. This models a scenario in which the agents obtain the full reward from their action, receive the payment, and then do not return. Note that we assume that both the principal and the agent receive the full reward of each arm that is pulled.

It is instructive to consider two extreme policies with respect to payments. At one extreme is the MYOPIC policy that never offers payments to agents. Then, the agents always myopically choose the arm maximizing the expected reward conditioned on the observed history. It is well known that such myopic exploitation-only policies can be far from optimal in terms of total time-discounted reward. At the other extreme is the PAY-WHAT-IT-TAKES policy that always offers a payment large enough to induce the agents to follow the optimal policy (which is the Gittins index policy [Gittins and Jones 1974]). Such a policy is obtained by offering, in each round, a payment for playing the arm that would be played by the incentive-free optimal policy; this payment equals the difference of conditional expectations between the arm maximizing expected posterior reward and the arm chosen by the optimal policy. This policy would achieve optimal expected rewards, but at a possibly exorbitant price.

The goal of the present work is to characterize the tradeoff between the algorithm's payments and the total reward that can be achieved. Let OPT denote the expected total time-discounted reward of the optimum algorithm which does not have to interact with selfish agents. We call a point $(\rho, b) \in [0,1]^2$ *achievable* at discount rate γ if for every MAB instance with non-negative rewards, there exists an adaptive policy satisfying the following: (1) Its expected time-discounted reward is at least $\rho \cdot \mathrm{OPT}$. (2) Its expected time-discounted payment to agents is at most $b \cdot \mathrm{OPT}$. The main technical result of the paper is the following essentially complete characterization of achievable points.

THEOREM 1.1. *Let (ρ, b) be a point in $[0,1]^2$.*

(*1*) *If $\sqrt{b} + \sqrt{1-\rho} < \sqrt{\gamma}$, then (ρ, b) is not achievable.*
(*2*) *If $\sqrt{b} + \sqrt{1-\rho} > \sqrt{\gamma}$, then (ρ, b) is achievable.*

Figure 1 illustrates the achievable points.

[1] In fact, we consider a more general model, in which each arm follows an independent Markov chain. But for concreteness, in this section, we will describe the results for the simpler model.

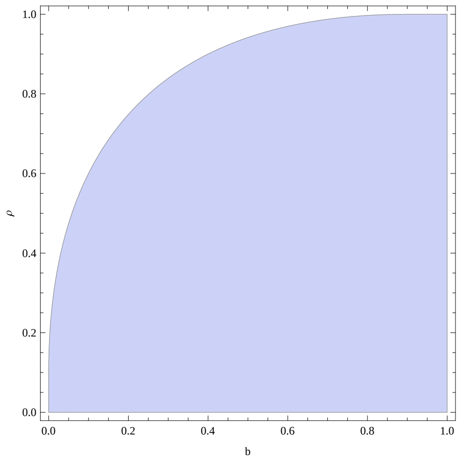

Fig. 1. The achievable region of reward-payment tradeoffs with $\gamma = 0.9$ (shown shaded)

The proof of Theorem 1.1 proceeds as follows: We first consider a Lagrangian relaxation of the problem of maximizing the total expected reward subject to a constraint on the total expected payment. In the Lagrangian relaxation, the budget constraint is removed, but the objective function is modified to incorporate a term that penalizes the expected (time-discounted) payments made to agents. The resulting optimization problem can be approached with a policy that independently randomizes between the MYOPIC and PAY-WHAT-IT-TAKES policy in each round. By randomizing between these two policies with appropriate probabilities, we can ensure that the cost of the payments offered when playing PAY-WHAT-IT-TAKES is exactly canceled by the surplus payoff received when playing MYOPIC. We leverage this cancellation to show that the expected reward of our randomized policy equals the reward of the *optimum* policy for the same MAB instance, but with a steeper time discount $\eta < \gamma$. We then prove the following general theorem, comparing the rewards of policies on the same MAB instance, with different discounts.

THEOREM 1.2. *Consider a fixed MAB instance (without selfish agents) with two different time discount factors $\eta < \gamma$, and let $\mathrm{OPT}_\eta, \mathrm{OPT}_\gamma$ be the optimum time-discounted rewards achievable under these discounts. Then, $\mathrm{OPT}_\eta \geq \frac{(1-\gamma)^2}{(1-\eta)^2} \cdot \mathrm{OPT}_\gamma$, and this bound is tight.*

Theorem 1.2 may be of independent interest, since it characterizes the maximum rate at which an optimal policy may lose rewards as they are discounted more steeply; in particular, it shows that the reward decreases gracefully with the discount factor. Theorem 1.2 is proved by analyzing an algorithm which follows the optimum policy at rate γ, but randomly selects a set of arms to "censor," so that pulling these arms is counted as a no-op instead.

As a final step in the analysis, we show that by appropriately randomizing between two different policies, each of which randomizes between the myopic and an optimal policy in each step, we can come arbitrarily close to a factor $1 - (\sqrt{\gamma} - \sqrt{b})^2$ of the optimum policy while heeding the budget constraint.

Related Work

The Bayesian MAB problem was introduced by Robbins [1952], and studied by a great number of authors as a model for the tradeoff between exploration and exploitation. A major breakthrough came with the work of Gittins [Gittins and Jones 1974], who gave a tractable method for computing the optimal policy. The MAB problem has also been studied extensively in the stochastic non-Bayesian [Lai and Robbins 1985] and adversarial [Auer et al. 1995, 2003] settings. For an overview, see the survey article [Mahajan and Teneketzis 2007] or the monograph [Gittins et al. 2011]. Almost all of the previous work on multi-armed bandits focuses on the single-agent problem, in which the principal directly controls the arms pulled without needing incentive payments.

A similar motivation to ours was considered in [Kremer et al. 2013]. Their goal is also to incentivize selfish agents who arrive one by one to explore different options. The difference between our mechanism design problem and the one in [Kremer et al. 2013] is as follows: in the model of [Kremer et al. 2013], only the principal observes the outcomes of prior agents' actions, whereas the other agents are not privy to this information. The principal does not offer the agents rewards; instead, he can strategically decide which outcomes to reveal in order to incentivize exploration. In this sense, the model of [Kremer et al. 2013] applies naturally to recommendation systems such as traffic-based driving recommendations. The precise models are sufficiently different that a technical comparison of results is not useful.

The goal of learning in MAB settings with a budget has been pursued in a series of recent papers (e.g., [Goel et al. 2009; Guha and Munagala 2007, 2009, 2013]). Guha and Munagala [2007] introduced the problem and gave an LP-based approximation; Goel et al. [2009] gave a much simpler index-based algorithm for this problem. In these papers, the costs are not dependent on the observed history, so the techniques do not appear to carry over to our setting, although our work and the work of Guha and Munagala [2007, 2009, 2013] share the theme of using Lagrangian relaxation to reduce a budget-constrained learning problem to an unconstrained problem with a mixed-sign objective accounting for a linear combination of payoffs and costs.

Learning with selfish agents has been considered in several papers, e.g., [Bolton and Harris 1999; Keller et al. 2005]. Traditionally, the focus is on the information sharing aspect: agents can strategically choose whether to pull arms, but learn from all other agents' pulls as well, which leads to an analysis of the extent of free-riding.

A different strategic MAB setting was considered by Bergemann and Välimäki [Bergemann and Välimäki 1996] and work extending their model. Here, the arms themselves can set strategic prices for being pulled: this models a setting in which firms can set prices for products that users want to try. The strategic considerations that arise are sufficiently different from those in our work for the technical results to be incomparable.

Several recent papers (e.g., [Abraham et al. 2013; Badanidiyuru et al. 2012, 2013; Ho et al. 2013; Singla and Krause 2013]) have drawn a connection between MAB models and crowd-sourcing of tasks. Contrary to our model, these papers typically treat the *agents* as arms in the MAB instance, i.e., the algorithm's goal is to learn the quality of the agents' work from observing them while providing enough incentives for them to work. Thus, despite a general interest in the same problem domain, the specific technical approaches are very different.

2. DEFINITIONS AND PRELIMINARIES

We consider a collection of k arms, and time indexed by $t = 0, 1, \ldots$. Each arm has associated with it an independent Markov chain (with possibly infinitely many states). Let $S_{t,i}$ denote the state of the Markov chain for arm i at time t. $S_{t,i}$ is perfectly observable by the principal and all selfish agents. In the standard MAB setting, an algorithm (or *policy*) decides, for each time t, which of k arms to pull. At time t, if arm i is pulled, it generates a reward whose distribution depends only on i's current state $S_{t,i}$; arm i then advances to a new state $S_{t+1,i}$ according to its Markov chain's known transition kernel. If arm i is not pulled, then $S_{t+1,i} = S_{t,i}$.

Let $r_{t,i}$ denote the random reward that would have been generated by pulling arm i at time t. We assume that the reward sequence generated by this Markov chain is non-negative, and that the Markov chain has an optimal policy for each discount factor; this issue is discussed more below. Also, we consider only the case that the reward sequence forms a Martingale, in the following sense:

$$\mathbb{E}\left[\mathbb{E}\left[r_{t+1,i} \mid S_{t+1,i}\right] \mid S_{t,i}\right] \; = \; \mathbb{E}\left[r_{t,i} \mid S_{t,i}\right].$$

An important and well-studied special case captures all of the motivating examples discussed in Section 1, and is the model we outlined there as well. Each arm i has associated with it some latent random variable θ_i, drawn independently at random from a known Bayesian prior probability distribution over a space Θ_i. This θ_i determines the payoff distribution $\Gamma_i = \Gamma_i(\theta_i)$ for arm i. Conditioned on θ_i, the rewards from arm i are i.i.d. from Γ_i. This scenario is a special case of the Markovian framework, as follows: the state space of the Markov chain for arm i is the set of all finite sequences of rewards that can be generated from arm i's distribution. The reward distribution conditioned on $S_{t,i}$ is then the Bayesian posterior distribution of arm i's payoff conditioned on the observed history of rewards. The law of iterated conditional expectations can be used to show that the reward sequence obtained from such a Markov chain is a Martingale.

The algorithm's choice will be based on the current state of all arms. Let $\boldsymbol{S}_t = (S_{t,i})_i$ be the vector containing the current state for each arm. An algorithm is then precisely a (distribution over) mappings from a time t and state $\boldsymbol{S}_t$ to arms i_t to pull next. We allow the mapping to be randomized, but to avoid excessive notation, we will typically describe an algorithm as a function $\mathcal{A} : (t, \boldsymbol{S}_t) \mapsto i_t$, and simply lump in the algorithm's randomness with the various other sources of randomness.

In the model with selfish agents studied in this work, the algorithm (which guides the principal's actions) cannot directly pull an arm, and instead must incentivize selfish agents to pull the arms. At each time $t = 0, 1, \ldots$, a selfish agent arrives and chooses an arm i_t to pull, based on the arm's expected reward and the incentive payments offered by this algorithm. When the arm is pulled, the principal and the current agent (but not other agents) are rewarded with $R_t = r_{t,i_t}$. While only the principal and the current agent actually earn the reward, the principal and *all* agents observe the current state of each arm's Markov chain; in the learning case, all agents observe each reward's value, from which each arm's current state is determined. Thus, all agents and the principal's algorithm have the same information $\boldsymbol{S}_t$ at any time t.

Based on the current state $\boldsymbol{S}_t$, at time t, the principal's algorithm determines payments $c_{t,i} \geq 0$ to offer the agent for playing the different arms i. The agent then chooses the arm i maximizing

$$\mathbb{E}\left[r_{t,i} \mid \boldsymbol{S}_t\right] + c_{t,i}. \tag{1}$$

Because the algorithm can evaluate Expression (1), we can assume w.l.o.g. that the algorithm chooses $c_{t,i} = 0$ for all but at most one arm. Since the algorithm's goal includes minimizing payments, we can describe the algorithm equivalently by specifying the (random) sequence $(i_t)_t$ of arms to pull. To achieve this sequence, the algorithm must offer a payment

$$c_t \; := \; c_{t,i_t} \; = \; \max_i \mathbb{E}\left[r_{t,i} \mid \boldsymbol{S}_t\right] - \mathbb{E}\left[r_{t,i_t} \mid \boldsymbol{S}_t\right], \tag{2}$$

which captures the minimum payment to give the desired arm the same expected reward in the eyes of the agent as the arm with highest posterior expected reward. Thus, in summary, even when interacting with selfish agents, an algorithm can be considered as a (possibly randomized) mapping $\mathcal{A} : (t, \boldsymbol{S}_t) \mapsto i_t$, with the understanding that this sequence determines payments made from the algorithm to the agents.

To indicate the expectation taken under algorithm $\mathcal{A}$, we use the notation $\mathbb{E}_{\mathcal{A}}[\cdot]$. When the algorithm is clear from the context, we simply write $\mathbb{E}[\cdot]$. The algorithm's payoffs and payments are discounted by a factor $\gamma \in (0, 1)$. Thus, the expected total discounted sum of payoffs of a given algorithm $\mathcal{A}$ is

$$R^{(\gamma)}(\mathcal{A}) \;=\; \mathbb{E}_{\mathcal{A}}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right], \tag{3}$$

and the expected total discounted payment is

$$C^{(\gamma)}(\mathcal{A}) \;=\; \mathbb{E}_{\mathcal{A}}\left[\sum_{t=0}^{\infty} \gamma^t c_t\right]. \tag{4}$$

We assume that for each discount factor $\gamma \in [0, 1]$, the MAB instance possesses an optimal policy, i.e., one that attains the supremum $\sup_{\mathcal{A}} R^{(\eta)}(\mathcal{A})$. We also assume that this supremum is finite. A sufficient condition for this assumption is that the state space of each Markov chain is countable, and the rewards $r_{t,i}$ are bounded above [Dynkin and Yushkevich 1979]. For other sufficient conditions, see [Dynkin and Yushkevich 1979].

The classical literature on Bayesian MABs [Gittins 1989] has considered the single-agent problem of finding a policy maximizing the expected discounted sum of payoffs, $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$, disregarding any incentive payments. An optimal policy for this problem was discovered in [Gittins and Jones 1974], and consists of computing at each time t the Gittins index associated with the current state of each arm's Markov chain, and pulling the arm for which this Gittins index is the highest, breaking ties arbitrarily.[2]

We let OPT_γ denote both this optimal Gittins index policy and the expected discounted sum of payoffs that it receives.

3. MAIN RESULT AND PROOF

Recall that our main goal is to characterize, for every γ, the *achievable* pairs (ρ, b) at discount rate γ, i.e., the pairs such that for each MAB instance with discount rate γ, an algorithm $\mathcal{A}$ using payments $C^{(\gamma)}(\mathcal{A}) \le b \cdot \mathrm{OPT}_\gamma$, can obtain total reward $R^{(\gamma)}(\mathcal{A}) \ge \rho \cdot \mathrm{OPT}_\gamma$. The characterization is given by Theorem 1.1, restated here for convenience:

Theorem 1.1 *Let (ρ, b) be a point in $[0, 1]^2$.*

(*1*) *If $\sqrt{b} + \sqrt{1-\rho} < \sqrt{\gamma}$, then (ρ, b) is not achievable.*
(*2*) *If $\sqrt{b} + \sqrt{1-\rho} > \sqrt{\gamma}$, then (ρ, b) is achievable.*

PROOF. To prove the first part, we present in Section 7 a family of instances we term *Diamonds in the Rough*. In those instances, one arm has known constant payoff. All remaining arms have payoffs of the following form: they are either a very large constant (with some small probability), or the constant 0. We call such arms *collapsing*

[2] While computing the Gittins index is computationally demanding, requiring, e.g., the solution of a sequence of dynamic programs [Whittle 1980], it can be computed using information about only a single arm, unlike the dynamic program for the full MAB MDP, which scales exponentially in the number of arms. Other algorithms for efficient computation of the Gittins index have been proposed in [Varaiya et al. 1985; Katehakis and Veinott 1987], and computation methods are surveyed in [Mahajan and Teneketzis 2007].

arms. The parameters are chosen such that myopic agents always choose the arm with constant payoff. With the right choice of parameters, the optimum policy is to explore collapsing arms until one with large constant payoff is found; subsequently, that arm is pulled forever. In Section 7, we show the following lemma, which implies the upper bound on the rewards:

LEMMA 3.1. *For any $b \leq \gamma$, there is a choice of arm payoffs and probabilities such that the optimum policy without selfish agents obtains payoff arbitrarily close to 1, yet any policy incurring a total expected payment of at most b obtains reward at most $1 - (\sqrt{\gamma} - \sqrt{b})^2$.*

For the second part, we fix γ and a MAB instance, and study the optimization problem of maximizing $R^{(\gamma)}(\mathcal{A})$, subject to the constraint that $C^{(\gamma)}(\mathcal{A}) \leq b$. We show a lower bound on $R^{(\gamma)}(\mathcal{A})$. We begin by studying the Lagrangian relaxation for some $\lambda \in [0, 1]$, i.e., maximizing

$$R^{(\gamma)}_{\lambda}(\mathcal{A}) \;=\; R^{(\gamma)}(\mathcal{A}) - \lambda \cdot C^{(\gamma)}(\mathcal{A}).$$

To achieve good bounds on the Lagrangian relaxation $R^{(\gamma)}_{\lambda}(\mathcal{A})$, we consider randomizations between *time-expanded versions* of policies. Specifically, let $\mathcal{A}$ be a stationary non-randomized policy. Informally, the time-expanded version of $\mathcal{A}$ with parameter $p \in [0, 1]$ is the following policy: at each time t, with probability p, offer no incentives (thus making agent t choose a myopically optimal arm), and with probability $1 - p$, offer the reward necessary to make the agent play the arm that $\mathcal{A}$ would choose based on only the "amount of information" $\mathcal{A}$ has acquired without the myopic pulls before time t. Make these choices independently for each time step.

A formal definition requires additional notation. Let $X_{i,t} = 1_{[i_t \neq i]}$ be an indicator random variable that is 1 if $\mathcal{A}$ pulled arm i in round t, and 0 otherwise. Let $\tau_{i,\ell} = \min\{t \mid \sum_{t' \leq t} X_{i,t'} \geq \ell\}$ be the time when $\mathcal{A}$ pulled arm i for the ℓ^{th} time. (If arm i is pulled fewer than ℓ times total, then $\tau_{i,\ell} = \infty$; we will ensure that we reference this definition only for values ℓ such that the arm has been pulled at least ℓ times.) For a vector $\boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_k)$, we define the ***$\boldsymbol{\ell}$-pull state $\boldsymbol{S}(\boldsymbol{\ell})$*** to be the state when each arm i has been pulled ℓ_i times. That is, $\boldsymbol{S}(\boldsymbol{\ell}) = (S_{\tau_{i,\ell_i},i})_i$.

Let $\mathcal{Z} = (Z_0, Z_1, Z_2, \ldots)$ be a sequence of independent Bernoulli$(1 - p)$ random variables, and $\mathcal{Z}_{0:t} = (Z_0, Z_1, \ldots, Z_t)$. When $Z_t = 1$, the time-expanded version of $\mathcal{A}$ will follow $\mathcal{A}$, and when $Z_t = 0$, it will have the agent behave myopically. Let $N^{\mathcal{Z}_{0:T}}_{i,T} = \sum_{t=0}^{T} X_{i,t} Z_t$ be the number of time steps t up to and including time T at which the algorithm pulled arm i and $Z_t = 1$, and $\boldsymbol{N}^{\mathcal{Z}_{0:T}}_{T} = (N^{\mathcal{Z}_{0:T}}_{i,T})_i$ the vector of all of these numbers of time steps. If $T = -1$, then $N^{\mathcal{Z}_{0:T}}_{i,T} = 0$. The ***$\mathcal{Z}$-induced state*** up to time T is $\hat{\boldsymbol{S}}_T(\mathcal{Z}_{0:T}) = \boldsymbol{S}(\boldsymbol{N}^{\mathcal{Z}_{0:T}}_{T})$. In words, the $\mathcal{Z}$-induced state is obtained as follows: for each arm i, count the number $N^{\mathcal{Z}_{0:T}}_{i,T}$ of times it was pulled when $Z_t = 1$. Then, keep the state that results from the *first* $N^{\mathcal{Z}_{0:T}}_{i,T}$ pulls of arm i, for each arm i, even if they occurred at times when $Z_t = 0$.

Now, we can define the time-expanded version $\text{TE}_{p,\mathcal{A}}$ of $\mathcal{A}$ with parameter p:

$$\text{TE}_{p,\mathcal{A}}(t) = \begin{cases} \mathcal{A}(\hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1})), & \text{if } Z_t = 1, \\ \operatorname{argmax}_i \mathbb{E}\left[r_{t,i} \mid \boldsymbol{S}_t\right], & \text{if } Z_t = 0. \end{cases}$$

Notice the following:

(1) The choice of whether to follow a myopic policy or the policy $\mathcal{A}$ is determined by Z_t.

(2) In evaluating the choice of the policy $\mathcal{A}$, the time-expanded algorithm does not consider everything it has learned. For each arm, it considers only the "amount" of information it could have learned at time steps that were not myopic pulls. However, the actual information may be obtained at time steps when myopic pulls were made.

(3) To determine the payments to offer when incentivizing the agents to play according to $\mathcal{A}$, the time-expanded policy *does* take all past outcomes into account, so that the payments are based on the same information used by the agents to evaluate the arms.

The main result about the time-expanded policy, which is proved in Section 4, is summarized in the following lemma:

LEMMA 3.2. *Given a parameter λ, define $p = \frac{\lambda}{\lambda+1}$, and $\eta = \frac{(1-p)\gamma}{1-p\gamma}$. Then,*

$$R_{\lambda}^{(\gamma)}(\boldsymbol{TE}_{p,\mathcal{A}}) \;=\; \frac{1-\eta}{1-\gamma} \cdot R^{(\eta)}(\mathcal{A}).$$

In words, the Lagrangian payoff of the time-expanded policy is equal to the payoff of $\mathcal{A}$ in the absence of selfish agents, but with a different time-discount factor η, and multiplied by a factor depending on the two discount factors, γ and η. In particular, we will later apply this lemma to optimal policies $\mathrm{OPT}\eta$ for suitable choices of η.

The next step is to relate the time-discounted payoffs of the optimal policies OPT_{γ}, OPT_{η}, for different time discounts γ, η. This is accomplished via Theorem 1.2; we restate it here and prove it in Section 5.

Theorem 1.2 *Consider a fixed MAB instance (without selfish agents), with two different time discount factors $\eta < \gamma$, and let $\mathrm{OPT}_{\eta}, \mathrm{OPT}_{\gamma}$ be the optimum time-discounted rewards achievable under these discounts. Then, $\mathrm{OPT}_{\eta} \geq \frac{(1-\gamma)^2}{(1-\eta)^2} \cdot \mathrm{OPT}_{\gamma}$, and this bound is tight.*

Finally, in Section 6, we use the previous results to establish the existence of a budget-bounded policy that performs well compared to the optimum; this policy is a randomization between the time expansions of two policies $\mathrm{OPT}_{\eta}, \mathrm{OPT}_{\eta'}$ for suitably chosen η, η'. This is summarized in the following lemma.

LEMMA 3.3. *For every $\epsilon > 0$ and every b, there is a $p \in [0,1]$ and a policy $\mathcal{A}$ satisfying*

$$\begin{aligned} &R^{(\gamma)}(\mathcal{A}) \geq \left(1 - p\gamma + \frac{p \cdot b}{1-p}\right) \cdot \mathrm{OPT}_{\gamma} - \epsilon, \\ &C^{(\gamma)}(\mathcal{A}) \leq b \cdot \mathrm{OPT}_{\gamma}. \end{aligned}$$

We bound the factor on the right-hand side of the reward bound as follows:

$$\begin{aligned} 1 - p\gamma + \frac{p \cdot b}{1-p} &= 1 - \gamma + \gamma(1-p) + \frac{b}{1-p} - b \\ &\geq 1 - \gamma + 2\sqrt{\gamma \cdot b} - b \\ &= 1 - (\sqrt{\gamma} - \sqrt{b})^2, \end{aligned}$$

where the inequality in the middle is justified by comparing the arithmetic and geometric means of the numbers $\gamma(1-p)$ and $\frac{b}{1-p}$.

When $\sqrt{b} + \sqrt{1-\rho} > \sqrt{\gamma}$, we have $\rho < 1 - (\sqrt{\gamma} - \sqrt{b})^2$, and it is possible to choose ϵ small enough that $R^{(\gamma)}(\mathcal{A}) \geq \rho \cdot \mathrm{OPT}_{\gamma}$. □

4. FROM INCENTIVES TO TIME-EXPANSION: PROOF OF LEMMA 3.2

In this section, we prove Lemma 3.2. Throughout, we choose $p = \frac{\lambda}{1+\lambda}$, so that $\lambda = \frac{p}{1-p}$. Let $\mathcal{A}$ be an arbitrary stationary non-randomized policy (without selfish agents). We begin by relating the expected Lagrangian payoff of $\mathrm{TE}_{p,\mathcal{A}}$ in some round t with the expected (non-Lagrangian) payoff it would obtain from choosing the non-myopic action in the same round.

LEMMA 4.1. *Fix a time t, a sequence of state vectors $\mathcal{S}_{0:t} = (\boldsymbol{S}_0, \ldots \boldsymbol{S}_t)$, and sequence of random Bernoulli variables $\mathcal{Z}_{0:t-1}$. Let $i^* = \mathcal{A}(\hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}))$ be the arm that $TE_{p,\mathcal{A}}$ would pull in round t if $Z_t = 1$. Then,*

$$\mathbb{E}_{TE_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right] \;=\; \mathbb{E}_{TE_{p,\mathcal{A}}}\left[r_{t,i^*} \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right]. \tag{5}$$

Proof. Consider the expected payoffs and incentive payments of the two actions that could be made by $\mathrm{TE}_{p,\mathcal{A}}$ at time t, depending on the coin flip Z_t. When $Z_t = 0$, the algorithm will choose an arm $i' \in \operatorname{argmax}_i \mathbb{E}\left[r_{t,i} \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right]$, whereas when $Z_t = 1$, the algorithm will choose the arm i^*. Because all calculations are conditioned on $\mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}$ and $\mathcal{A}$ is non-randomized, both i' and i^* are completely determined by these histories. Thus, the values of expectations we take do not depend upon the algorithm under which they are taken; we therefore drop this dependence in the notation.

When $Z_t = 0$, the algorithm incurs no payment and obtains reward $x := \mathbb{E}\left[r_{t,i'} \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right]$. When $Z_t = 1$, the algorithm obtains reward $y := \mathbb{E}\left[r_{t,i^*} \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right]$, while paying the agent $x - y$. The expected Lagrangian payoff is

$$\begin{aligned}
&\mathbb{E}\left[R_t - \lambda c_t \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right] \\
&= (1-p) \cdot \mathbb{E}\left[R_t - \lambda c_t \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}, Z_t = 1\right] + p \cdot \mathbb{E}\left[R_t - \lambda c_t \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}, Z_t = 0\right] \\
&= (1-p)(y - \frac{p}{1-p}(x-y)) + px \\
&= (1-p)y - px + py + px \;=\; y. \quad \square
\end{aligned}$$

Next, we use Lemma 4.1 to relate the Lagrangian payoff of the time-expanded policy to the payoff of the original policy at a corresponding time.

LEMMA 4.2. *Assume that $p < 1$. Let $\zeta_{t-1} = \sum_{t' < t} Z_{t'}$ be the total number of non-myopic steps performed by the time-expanded algorithm prior to time t. Then, for any $0 \le n \le t$,*

$$\mathbb{E}_{TE_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \zeta_{t-1} = n\right] = \mathbb{E}_{\mathcal{A}}\left[R_n\right].$$

PROOF. We begin with Equation (5), and condition iteratively on two random variables: first, we condition on $\zeta_{t-1} = n$, and subject to this, we condition on the $\mathcal{Z}$-induced state $\hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1})$. We first apply this conditioning on the left side of Equation (5), and use the law of iterated conditional expectation, along with the fact that $\hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1})$ and ζ_{t-1} are measurable with respect to $\mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}$:

$$\begin{aligned}
&\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right] \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right] \\
&\qquad = \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right].
\end{aligned}$$

Performing the same operation on the right-hand side of Equation (5) gives us that

$$\begin{aligned}
&\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{t,i^*} \mid \mathcal{S}_{0:t}, \mathcal{Z}_{0:t-1}\right] \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right] \\
&\qquad = \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{t,i^*} \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right].
\end{aligned}$$

Let $\tau = \inf\{\tau \geq t \mid Z_\tau = 1\} \geq t \geq n$ be the (random) next time that $\mathrm{TE}_{p,\mathcal{A}}$ will choose an action according to $\mathcal{A}$. Recall that the sequence of conditional expected rewards $(\mathbb{E}\left[r_{u,i} \mid S_{u,i}\right])_{u=0,1,\ldots}$ is a Martingale. Furthermore, i^* is measurable with respect to $\hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1})$. Because $t \geq n$, we have that $\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{u,i^*} \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right]$ is identical for all $u \geq t$. Moreover, this conditional expectation does not change if we also condition on $\tau = u$, so

$$\begin{aligned}
&\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{\tau,i^*} \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right] \\
&= \sum_{u \geq t} \Pr[\tau = u] \cdot \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{\tau,i^*} \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n, \tau = u\right] \\
&= \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[r_{t,i^*} \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right],
\end{aligned}$$

meaning that the expected conditional reward is the same at the random time τ as at time t. Noting that $r_{\tau,i^*} = R_\tau$ and combining this with the expressions above implies that

$$\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right] \;=\; \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_\tau \mid \hat{\boldsymbol{S}}_{t-1}(\mathcal{Z}_{0:t-1}), \zeta_{t-1} = n\right].$$

Taking the conditional expectation with respect to $\zeta_{t-1} = n$ on both sides, and applying again the law of iterated conditional expectation, shows that

$$\mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \zeta_{t-1} = n\right] \;=\; \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_\tau \mid \zeta_{t-1} = n\right].$$

By definition of the time-expanded policy $\mathrm{TE}_{p,\mathcal{A}}$, the right-hand side is equal to $\mathbb{E}_{\mathcal{A}}\left[R_n\right]$, completing the proof. □

Proof of Lemma 3.2. We write the entire time-discounted Lagrangian reward as follows:

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t\right] \quad &= \quad \sum_{t=0}^{\infty}\sum_{n=0}^{\infty} \gamma^t \mathbb{E}_{\mathrm{TE}_{p,\mathcal{A}}}\left[R_t - \lambda c_t \mid \zeta_{t-1} = n\right] \cdot \Pr[\zeta_{t-1} = n] \\
&\overset{\text{Lemma 4.2}}{=} \quad \sum_{n=0}^{\infty} \mathbb{E}_{\mathcal{A}}\left[R_n\right] \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \Pr[\zeta_{t-1} = n] \\
&= \quad \sum_{n=0}^{\infty} \mathbb{E}_{\mathcal{A}}\left[R_n\right] \cdot \sum_{t=n}^{\infty} \gamma^t \cdot \binom{t}{n} \cdot (1-p)^n \cdot p^{t-n} \\
&= \quad \sum_{n=0}^{\infty} \mathbb{E}_{\mathcal{A}}\left[R_n\right] \cdot \gamma^n (1-p)^n \cdot \sum_{s=0}^{\infty} \binom{n+s}{n} \cdot (\gamma p)^s \\
&= \quad \sum_{n=0}^{\infty} \mathbb{E}_{\mathcal{A}}\left[R_n\right] \cdot \gamma^n \cdot (1-p)^n \cdot (1-\gamma p)^{-(n+1)} \\
&= \quad \sum_{n=0}^{\infty} \mathbb{E}_{\mathcal{A}}\left[R_n\right] \cdot \tfrac{1-\eta}{1-\gamma} \cdot \eta^n.
\end{aligned}$$

In the penultimate step, we used the identity $\sum_{s=0}^{\infty} \binom{n+s}{n} x^s = (1-x)^{-(n+1)}$, which is valid for $n \geq 0$ and $|x| < 1$. In the last step, we used the definition that $\eta = \frac{(1-p)\gamma}{1-\gamma p}$. □

5. THE EFFECTS OF DIFFERENT TIME DISCOUNTS: PROOF OF THEOREM 1.2

In this section, we will prove Theorem 1.2, i.e., that $\mathrm{OPT}_\eta \geq \frac{(1-\gamma)^2}{(1-\eta)^2} \cdot \mathrm{OPT}_\gamma$. An example showing that the bound is tight is given in Section 7.

Proof of Theorem 1.2. Since it seems difficult to directly analyze OPT_η, we instead define a suboptimal policy RC for which we can establish the same inequality; this naturally implies the inequality for OPT_η. The policy RC is the following *random censorship policy*. At time zero, an independent coin is tossed for each arm i that marks i as *censored* with probability $1-q$ and *uncensored* with probability q, where $q = \frac{1-\gamma}{1-\eta}$. (The significance of this choice of q will become apparent later.) RC simulates the policy OPT_γ, but treats the censored arms differently from the uncensored ones. When a censored arm i is pulled in the simulation, RC simulates one state transition of the Markov chain for arm i but does not pull any arm in reality. When an uncensored arm i is pulled in the simulation, arm i is pulled in reality. (There is a corner case in which all arms are censored. In that case, the RC policy pulls arm 1 forever.)

To obtain a lower bound on the expected time-discounted payoff of RC with discount factor η, we will analyze each arm separately. Define the following random variables:

$$\begin{aligned} Y_i &= 1_{[\text{arm } i \text{ is uncensored}]} \\ \hat{R}_{t,i} &= \text{reward from } t^{\text{th}} \text{ pull of arm } i \text{ in the simulation} \\ M_{t,i,j} &= \text{number of times arm } j \text{ is pulled before } t^{\text{th}} \text{ simulated pull of arm } i. \end{aligned}$$

If arm i is pulled strictly fewer than t times in the simulation, then $\hat{R}_{t,i} = 0$, and $M_{t,i,j}$ is equal to the total number of times j is pulled throughout the simulation (which could be ∞). The discounted reward received by RC when it pulls arm i for the t^{th} time in the simulation is

$$Y_i \cdot \hat{R}_{t,i} \cdot \eta^{t-1+\sum_{j\neq i} Y_j M_{t,i,j}}. \tag{6}$$

When arm i is censored, Expression (6) is correct because it equals zero, due to the fact that $Y_i = 0$. When arm i is uncensored, $Y_i = 1$ and $\hat{R}_{t,i}$ is the reward received from the t^{th} pull of the arm. The time at which the t^{th} pull of arm i takes place (when i is uncensored) is $t-1+\sum_{j\neq i} Y_j M_{t,i,j}$, which explains the final term in Expression (6).

The random variables $\{Y_i\}$ are mutually independent. Because the random variables $\hat{R}_{t,i}$ and $M_{t,i,j}$ depend only on the rewards observed in the simulation, but not on which arms are censored, they are mutually independent of the variables Y_i.

Conditioning on the random variables $\{\hat{R}_{s,i}, M_{s,i,j} \mid 1 \leq s \leq t, 1 \leq i,j \leq k\}$, we obtain the following conditional expectation:

$$\begin{aligned} &\mathbb{E}\left[Y_i \cdot \hat{R}_{t,i} \cdot \eta^{t-1+\sum_{j\neq i} Y_j M_{t,i,j}} \mid \hat{R}_{s,i}, M_{s,i,j}, 1 \leq s \leq t, 1 \leq i,j \leq k\right] \\ &\qquad\qquad = q \cdot \hat{R}_{t,i} \cdot \eta^{t-1} \cdot \prod_{j\neq i} (1-q+q\eta^{M_{t,i,j}})). \end{aligned} \tag{7}$$

Next, we claim that our choice $q = \frac{1-\gamma}{1-\eta}$ ensures that $1-q+q\eta^s \geq \gamma^s$ for any integer $s \geq 0$. This is clear when $s=0$; for $s \geq 1$, observe the following calculation:

$$\begin{aligned} 1-q+q\eta^s \;&=\; 1-\left(\tfrac{1-\gamma}{1-\eta}\right)(1-\eta^s) = 1-(1-\gamma)(1+\eta+\cdots+\eta^{s-1}) \\ &\geq 1-(1-\gamma)(1+\gamma+\cdots+\gamma^{s-1}) \;=\; 1-(1-\gamma^s) \;=\; \gamma^s. \end{aligned}$$

Substituting this bound into the right-hand side of Equation (7) yields

$$\begin{aligned}&\mathbb{E}\left[Y_i \cdot \hat{R}_{t,i} \cdot \eta^{t-1+\sum_{j\neq i} Y_j M_{t,i,j}} \mid \hat{R}_{s,i}, M_{s,i,j}, 1 \le s \le t, 1 \le i,j \le k\right] \\ &\qquad\qquad \ge\ q \cdot \hat{R}_{t,i} \cdot \eta^{t-1} \cdot \prod_{j\neq i} \gamma^{M_{t,i,j}} \ =\ q\eta^{t-1} \cdot \left(\hat{R}_{t,i}\gamma^{\sum_{j\neq i} M_{t,i,j}}\right).\end{aligned} \tag{8}$$

Let $a_{t,i} = \mathbb{E}\left[\hat{R}_{t,i}\gamma^{\sum_{j\neq i} M_{t,i,j}}\right]$. We have derived that the (unconditional) expected amount that arm i contributes to the time-discounted reward of the policy RC is bounded below by

$$q \cdot \sum_{t=1}^{\infty} a_{t,i}\eta^{t-1}. \tag{9}$$

When the policy OPT_γ is applied with discount rate γ, the discounted reward received from the t^{th} pull of arm i is $\hat{R}_{t,i}\gamma^{t-1+\sum_{j\neq i} M_{t,i,j}}$. Consequently, the unconditional expected contribution of arm i to the discounted reward of OPT_γ is

$$\sum_{t=1}^{\infty} a_{t,i}\gamma^{t-1}. \tag{10}$$

We next prove that the numbers $(a_{t,i})_{t=0,1,\ldots}$ form a non-increasing sequence. In the following calculation, we use that $M_{t+1,i,j} \ge M_{t,i,j}$ for every t, i, j, and that the rewards from arm i form a Martingale sequence. We repeatedly apply the law of total probability:

$$\begin{aligned}a_{t+1,i} &= \mathbb{E}\left[\hat{R}_{t+1,i}\gamma^{\sum_{j\neq i} M_{t+1,i,j}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\hat{R}_{t+1,i}\gamma^{\sum_{j\neq i} M_{t+1,i,j}} \mid S_{t+1,i}\right] \mid S_{t,i}, \hat{R}_{t,i}, \{M_{t,i,j} \mid 1 \le j \le k\}\right]\right] \\ &\le \mathbb{E}\left[\mathbb{E}\left[\gamma^{\sum_{j\neq i} M_{t,i,j}} \cdot \mathbb{E}\left[\hat{R}_{t+1,i} \mid S_{t+1,i}\right] \mid S_{t,i}, \hat{R}_{t,i}, \{M_{t,i,j} \mid 1 \le j \le k\}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\gamma^{\sum_{j\neq i} M_{t,i,j}} \cdot \hat{R}_{t,i} \mid S_{t,i}, \hat{R}_{t,i}, \{M_{t,i,j} \mid 1 \le j \le k\}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\gamma^{\sum_{j\neq i} M_{t,i,j}} \cdot \hat{R}_{t,i} \mid \hat{R}_{t,i}, \{M_{t,i,j} \mid 1 \le j \le k\}\right]\right] \\ &= \mathbb{E}\left[\hat{R}_{t,i}\gamma^{\sum_{j\neq i} M_{t,i,j}}\right] \ =\ a_{t,i}.\end{aligned}$$

Because $a_{t,i} - a_{t+1,i} \ge 0$ for all t, we can now bound

$$(1-\gamma)\sum_{t=1}^{\infty} a_{t,i}\gamma^{t-1} = a_{1,i} - \sum_{t=1}^{\infty} \gamma^t \cdot (a_{t,i} - a_{t+1,i}) \le a_{1,i} - \sum_{t=1}^{\infty} \eta^t \cdot (a_{t,i} - a_{t+1,i}) = (1-\eta)\sum_{t=1}^{\infty} a_{t,i}\eta^{t-1}.$$

Combining this inequality with Equations (9) and (10), and summing over all arms i, we obtain that the expected time-discounted payoff of RC is at least

$$q \cdot \sum_i \sum_{t=1}^{\infty} a_{t,i}\eta^{t-1} \ \ge\ q \cdot \tfrac{1-\gamma}{1-\eta} \cdot \sum_i \sum_{t=1}^{\infty} a_{t,i}\gamma^{t-1} \ =\ \left(\tfrac{1-\gamma}{1-\eta}\right)^2 \cdot \mathrm{OPT}_\gamma,$$

where we used the definition of q in the last step. This completes the proof. □

6. EXISTENCE OF GOOD POLICIES: PROOF OF LEMMA 3.3

In this section, we prove Lemma 3.3, which relies on Theorem 1.2 and Lemma 3.2. Given a constraint on the fractional expected cost b, this lemma shows the existence of an algorithm which randomizes between two time-expanded policies, satisfying the payment constraint, and does well with respect to the optimal policy.

Proof of Lemma 3.3. For any probability $p \in [0,1]$, let $\eta(p) = \frac{(1-p)\gamma}{1-p\gamma}$, $\lambda(p) = \frac{p}{1-p}$, and define $b(p) = C^{(\gamma)}(\mathbf{TE}_{p,\mathrm{OPT}_{\eta(p)}})/\mathrm{OPT}_\gamma$ to be the fraction of the optimal reward that the time-expanded version of $\mathrm{OPT}_{\eta(p)}$ pays the agents at discount γ.

Let $b_{\max} = \sup_{p\in[0,1]} b(p)$. If $b \geq b_{\max}$, then we choose $p = 0$ and $\mathcal{A} = \mathrm{OPT}_\gamma$. We satisfy the claimed constraint on $R^{(\gamma)}(\mathcal{A}) = \mathrm{OPT}_\gamma$ because $1 - p\gamma + \frac{p\cdot b}{1-p} = 1$ at $p = 0$. We satisfy the claimed constraint on $C^{(\gamma)}(\mathcal{A}) = b(0) \cdot \mathrm{OPT}_\gamma$ because $b(0) \leq b_{\max} \leq b$. Thus, it is sufficient to consider the case $b < b_{\max}$ from now on.

For any $p \in [0,1]$, combining Theorem 1.2 with Lemma 3.2, we obtain that the time-expanded version of $\mathrm{OPT}_{\eta(p)}$ satisfies the following guarantee for the value of the Lagrangian relaxation:

$$R^{(\gamma)}_\lambda(\mathbf{TE}_{p,\mathrm{OPT}_{\eta(p)}}) \;\geq\; (1-p\gamma)\cdot R^{(\gamma)}(\mathrm{OPT}_{\eta(p)}). \tag{11}$$

We now use the result for the Lagrangian relaxation to obtain a bound on the constrained optimization problem. Equation (11), applied with $\lambda = \lambda(p)$, implies that

$$R^{(\gamma)}(\mathbf{TE}_{p,\mathrm{OPT}_{\eta(p)}}) \;\geq\; \left(1 - p\gamma + \frac{p\cdot b(p)}{1-p}\right)\cdot \mathrm{OPT}_\gamma. \tag{12}$$

Thus, if there exists a p with $b(p) = b$, then we can simply set $\mathcal{A} = \mathrm{OPT}_{\eta(p)}$, and obtain the desired inequality even for $\epsilon = 0$. Notice that for $b = 0$, we can choose $p = 1$ (i.e., the myopic policy), which gives $b(p) = 0 = b$, so that we have already established the existence of $\mathcal{A}$ for $b = 0$.

It remains to consider the case $b \in (0, b_{\max})$ with $b(p) \neq b$ for all $p \in [0,1]$. Consider two sets, $S^+ = \{p \in [0,1] \mid b(p) > b\}$ and $S^- = \{p \in [0,1] \mid b(p) < b\}$. Because there are no p with $b(p) = b$, S^+ and S^- form a partition of $[0,1]$, and because $b \in (0, b_{\max})$, both sets are non-empty.

Because S^- and S^+ are complements of each other, and the complement of a closed set is open, S^- and S^+ cannot both be closed. If S^- is not closed, then $\mathrm{cl}(S^-)\setminus S^-$ is non-empty, and contains some point p; by definition, $p \in S^+$. Similarly, if S^+ is not closed then $\mathrm{cl}(S^+) \setminus S^+$ contains a point p, which is also in S^-. Without loss of generality, assume that there is a $p \in S^+ \cap \mathrm{cl}(S^-)$; the proof of the other case is similar. Let $(p_n)_n$ a sequence of values in S^- such that $\lim_{n\to\infty} p_n = p$. Applying Inequality (12) to p and p_n gives us that

$$R^{(\gamma)}(\mathbf{TE}_{p_n,\mathrm{OPT}_{\eta(p_n)}}) \geq \left(1 - p_n\cdot\gamma + \frac{p_n\cdot b(p_n)}{1-p_n}\right)\cdot \mathrm{OPT}_\gamma,$$
$$R^{(\gamma)}(\mathbf{TE}_{p,\mathrm{OPT}_{\eta(p)}}) \geq \left(1 - p\cdot\gamma + \frac{p\cdot b(p)}{1-p}\right)\cdot \mathrm{OPT}_\gamma.$$

Define $\alpha_n^+ = \frac{b-b(p_n)}{b(p)-b(p_n)}$ and $\alpha_n^- = \frac{b(p)-b}{b(p)-b(p_n)}$, so that $\alpha_n^- + \alpha_n^+ = 1$. Consider the algorithm $\mathcal{A}_n$ which initially flips a biased coin, and runs $\mathbf{TE}_{p_n,\mathrm{OPT}_{\eta(p_n)}}$ with probability α_n^- and $\mathbf{TE}_{p,\mathrm{OPT}_{\eta(p)}}$ with probability α_n^+.

By linearity of expectation, the expected total time-discounted payment of $\mathcal{A}_n$ is exactly $\alpha_n^- \cdot b(p_n) \cdot \mathrm{OPT}_\gamma + \alpha_n^+ \cdot b(p) \cdot \mathrm{OPT}_\gamma = b \cdot \mathrm{OPT}_\gamma$, and its expected total time-discounted reward (divided by OPT_γ for legibility) is at least

$$\alpha_n^- \cdot \left(1 - p_n \cdot \gamma + \frac{p_n \cdot b(p_n)}{1 - p_n}\right) + \alpha_n^+ \cdot \left(1 - p \cdot \gamma + \frac{p \cdot b(p)}{1-p}\right). \tag{13}$$

Let $\delta_n = |p_n - p|$. Then, Equation (13) is bounded below by

$$\alpha_n^- \cdot \left(1 - (p + \delta_n) \cdot \gamma + \frac{(p - \delta_n) \cdot b(p_n^-)}{1 - p + \delta_n}\right) + \alpha_n^+ \cdot \left(1 - (p + \delta_n) \cdot \gamma + \frac{(p - \delta_n) \cdot b(p_n^+)}{1 - p + \delta_n}\right).$$

Using that $\alpha_n^+ + \alpha_n^- = 1$ and $\alpha_n^- \cdot b(p_n^-) + \alpha_n^+ \cdot b(p_n^+) = b$ shows that this is equal to $1 - (p + \delta_n) \cdot \gamma + \frac{(p-\delta_n) \cdot b}{1-p+\delta_n}$, and we have the bound

$$R^{(\gamma)}(\mathcal{A}_n) \;\geq\; \left(1 - (p + \delta_n) \cdot \gamma + \frac{(p - \delta_n) \cdot b}{1 - p + \delta_n}\right) \cdot \mathrm{OPT}_\gamma.$$

In the limit as $n \to \infty$, this expression converges to $\left(1 - p \cdot \gamma + \frac{p \cdot b}{1-p}\right) \mathrm{OPT}_\gamma$. Thus, given $\epsilon > 0$, we can choose n large enough that $R^{(\gamma)}(\mathcal{A}_n) \geq \left(1 - p \cdot \gamma + \frac{p \cdot b}{1-p}\right) \cdot \mathrm{OPT}_\gamma - \epsilon$. □

7. A MATCHING LOWER BOUND: DIAMONDS IN THE ROUGH

In this section, we prove Lemma 3.1. For any $b \leq \gamma$, we exhibit a MAB instance of type "Diamonds in the Rough" with OPT_γ arbitrarily close to 1 on which any policy incurring a total expected payment of at most b obtains reward at most $1 - (\sqrt{\gamma} - \sqrt{b})^2$.

Given b and γ, define $\lambda = \sqrt{\frac{\gamma}{b}} - 1$ and $\psi = \frac{1}{1+\lambda}\gamma(1-\gamma)$. There is one arm with known constant reward $\psi + (1-\gamma)^2$. In addition, there is a practically infinite supply[3] of *collapsing arms*: Each such arm has a constant payoff; this constant is $M(1-\gamma)^2$ (the *good state*) with probability $1/M$, and 0 otherwise. The constant payoff is revealed (or, by the principle of deferred decisions, determined) the first time the arm is pulled. We treat M as infinite here; it is easy to see that the same bound is obtained formally by taking the limit of upper bounds using instances with $M \to \infty$.

The expected payoff when pulling one of the collapsing arms is $(1-\gamma)^2$; hence, unless a collapsing arm has already been collapsed to the good state (in which case both the optimal algorithm and a myopic agent will pull the good arm), the incentive payment required to get an agent to pull a collapsing arm is ψ.

Because $\psi \leq \gamma(1-\gamma)$, the policy that always pulls the deterministic arm achieves a time-discounted payoff of $\frac{1}{1-\gamma}\left((1-\gamma)^2 + \psi\right) = 1 - \gamma + \frac{\psi}{1-\gamma} \leq 1$. On the other hand, the policy that always pulls an uncollapsed arm until one of them collapses to the good state, and then pulls that arm forever after, achieves a time-discounted payoff W that satisfies

$$W \;=\; \tfrac{1}{M} \cdot M(1-\gamma)^2 \cdot \frac{1}{1-\gamma} + \left(1 - \frac{1}{M}\right)\gamma W \;=\; 1 - \gamma + \gamma W,$$

because M is infinite while W is finite. This implies that $W = 1$; thus, the optimal policy in the absence of selfish agents is the second one, and its time-discounted payoff is 1 (or arbitrarily close to 1, when the number of arms is finite).

[3] This infinite number of arms serves to make the construction clearer. By making the number k of arms sufficiently large (but still finite), the optimum payoff can be arbitrarily close to 1. The infinite supply avoids unnecessary complications from corner cases in which the optimum policy runs out of arms to explore.

Remark 7.1. By removing the arm with constant reward, we obtain an instance consisting only of collapsing arms. Changing the discount rate to η while keeping the arms' good payoff at $M(1-\gamma)^2$, we see that the solution for W is now $W = \frac{(1-\gamma)^2}{(1-\eta)^2}$. This shows that the bound of Theorem 1.2 is tight.

Consider the Lagrangian payoff $R^{(\gamma)}(\mathcal{A}) - \lambda C^{(\gamma)}(\mathcal{A})$, and let $\mathcal{A}$ be a policy maximizing this quantity, denoting by V its time-discounted Lagrangian payoff. Notice that $\mathcal{A}$ must always play a collapsed arm in a good state if there is one. If no arm has collapsed to a good state, the situation is exactly the same as at the beginning: the only two alternatives are to play the deterministic arm (and make no incentive payment) or to play an uncollapsed arm and make an incentive payment of ψ. In the former case, the immediate reward is $(1-\gamma)^2 + \psi$, and the cumulative time-discounted reward starting at the next time step is γV. In the latter case, a payment of ψ is made (and weighted by λ); then, with probability $1/M$, the reward $M(1-\gamma)^2$ is obtained for all remaining time steps, for a total time-discounted reward of $M(1-\gamma)$. With probability $1 - 1/M$, no reward is obtained, and the time-discounted reward starting at the next time step is γV. Combining these cases, using that $M \to \infty$ and the definition of ψ, we can now derive the following equation for V:

$$\begin{aligned} V &= \max\left\{(1-\gamma)^2 + \psi + \gamma V,\ \tfrac{1}{M} \cdot M \cdot (1-\gamma) + \left(1 - \tfrac{1}{M}\right) \cdot \gamma V - \lambda\psi\right\} \\ &= \gamma V + \max\left\{(1-\gamma)^2 + \psi,\ 1 - \gamma - \lambda\psi\right\} \\ &= \gamma V + (1-\gamma)(1 - \frac{\lambda\gamma}{1+\lambda}); \end{aligned}$$

in the last step, we used that both expressions in the maximum evaluate to the same. Solving for V gives that $V = 1 - \frac{\lambda\gamma}{1+\lambda}$. Thus, while the optimum policy without selfish agents achieves a time-discounted reward of 1, in the presence of selfish agents, no policy achieves a time-discounted Lagrangian reward more than $1 - \frac{\lambda\gamma}{1+\lambda}$.

Now consider the constrained optimization problem instead of the Lagrangian. Let $\mathcal{A}$ be an optimal policy with $C^{(\gamma)}(\mathcal{A}) \leq b$. We just proved that $R^{(\gamma)}(\mathcal{A}) - \lambda C^{(\gamma)}(\mathcal{A}) \leq 1 - \frac{\lambda\gamma}{1+\lambda}$. Solving for $R^{(\gamma)}(\mathcal{A})$, and using the definition of λ gives us that

$$R^{(\gamma)}(\mathcal{A}) \ \leq\ \lambda b + 1 - \frac{\lambda\gamma}{1+\lambda} \ =\ \sqrt{\gamma b} - b + 1 - \gamma + \sqrt{\gamma b} \ =\ 1 - (\sqrt{\gamma} - \sqrt{b})^2.$$

This gives the desired upper bound on the time-discounted reward of any policy $\mathcal{A}$ subject to a constraint on the total time-discounted payment.

8. CONCLUSIONS

In this paper, we proposed the study of multi-armed bandit problems in which the arms are pulled by selfish and myopic agents. In order to incentivize the agents to explore arms other than the myopically optimal one, the principal must offer them payments. We studied the tradeoff between the total (time-discounted) payments made and the total (time-discounted) reward; our main result was a complete characterization of the region of (reward, payment) pairs achievable.

We believe that our model forms a natural and robust theoretical basis from which to analyze crowd-sourced information discovery, scientific agendas, and other social endeavors in which agents' myopic objectives stand in conflict with a principal's long-term agenda. However, several specific modeling choices can be altered and would give rise to different technical questions.

First, the bound on the total time-discounted payoffs is required to hold in expectation. Instead, one could require it to hold pointwise, i.e., the principal *never* pays the agents more than $b \cdot \mathrm{OPT}$. This requirement would necessitate using a non-Markovian policy in place of the time-expanded policy that we analyzed in this paper, thus making the problem potentially harder to solve. Nonetheless, it is a natural and interesting question for future research.

The principal's objective in some cases, e.g., when crowd-sourcing the design of an artifact such as a logo or a website, may not be the sum of payoffs, but rather the maximum. In this case, a finite time horizon may be a more suitable model, and the characterization question will take on a more discrete nature.

So far, we have also assumed that payoffs of arms are entirely uncorrelated. When exploring a possible design (or research) space, it is natural to assume that similar alternatives would yield similar payoffs. This notion of similarity among alternatives could be modeled, for instance, as a Lipschitz condition on payments [Kleinberg et al. 2008] or as a Gaussian process prior [Srinivas et al. 2012]. An interesting question would be whether such a model could be fruitfully combined with one involving selfish myopic agents performing the exploration.

ACKNOWLEDGMENTS

We would like to thank Shaddin Dughmi, Li Han, Shane Henderson, and Lian Jian for useful discussions. Peter Frazier acknowledges support from NSF CAREER CMMI-1254298, NSF IIS-1247696, AFOSR YIP FA9550-11-1-0083, and AFOSR FA9550-12-1-0200. David Kempe was in part supported by NSF CAREER 0545855, a Sloan Fellowship, and an Okawa Foundation Research Fellowship. Robert Kleinberg acknowledges support from NSF awards IIS-0905467 and AF-0910940, AFOSR grant FA9550-09-1-0100, a Microsoft Research New Faculty Fellowship, and a Google Research Grant.

REFERENCES

Ittai Abraham, Omar Alonso, Vasilis Kandylas, and Aleksandrs Slivkins. 2013. Adaptive Crowdsourcing Algorithms for the Bandit Survey Problem. In *Proc. 26th Conference on Learning Theory*. 882–910.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 1995. Gambling in a rigged casino: The adversarial multi-armed banditproblem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*. 322–331.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2003. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* 32, 1 (2003), 48–77.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Yaron Singer. 2012. Learning on a budget: Posted price mechanisms for online procurement. In *Proc. 14th ACM Conference on Electronic Commerce*. 128–145.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2013. Bandits with Knapsacks. In *Proc. 54th IEEE Symposium on Foundations of Computer Science (FOCS)*. 207–216. arxiv.org/abs/1305.2545.

Dirk Bergemann and Juuso Välimäki. 1996. Learning and Strategic Pricing. *Econometrica* 64, 5 (1996), 1124–1149.

Patrick Bolton and Christopher Harris. 1999. Strategic Experimentation. *Econometrica* 67, 2 (1999), 349–374.

Eugene B. Dynkin and Alexander A. Yushkevich. 1979. *Controlled Markov Processes*. Springer, New York.

John C. Gittins. 1989. *Multi-Armed Bandit Allocation Indices*. John Wiley and Sons, New York.

John C. Gittins, Kevin D. Glazebrook, and Richard Weber. 2011. *Multi-Armed Bandit Allocation Indices* (2nd ed.). John Wiley and Sons, New York.

John C. Gittins and David M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, J. Gani (Ed.). North-Holland, Amsterdam, 241–266.

Ashish Goel, Sanjeev Khanna, and Brad Null. 2009. The ratio index for budgeted learning, with applications. In *Proc. 20th ACM-SIAM Symp. on Discrete Algorithms*. 18–27.

Sudipto Guha and Kamesh Munagala. 2007. Approximation algorithms for budgeted learning problems. In *Proc. 38th ACM Symp. on Theory of Computing*. 104–113.

Sudipto Guha and Kamesh Munagala. 2009. Multi-armed bandits with metric switching costs. In *Proc. 36th International Colloquium on Automata, Languages and Programming (ICALP)*. Springer, 496–507.

Sudipto Guha and Kamesh Munagala. 2013. Approximation Algorithms for Bayesian Multi-Armed Bandit Problems. (2013). arxiv.org/abs/1306.3525.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. 2013. Adaptive Contract Design for Crowdsourcing. (2013). Working Paper.

Michael N. Katehakis and Arthur F. Veinott, Jr. 1987. The multi-armed bandit problem: decomposition and computation. *Math. Oper. Res.* 12, 2 (1987), 262–268.

Godfrey Keller, Sven Rady, and Martin Cripps. 2005. Strategic experimentation with exponential bandits. *Econometrica* 73, 1 (2005), 39–68.

Robert Kleinberg, Alexandrs Slivkins, and Eli Upfal. 2008. Multi-Armed Bandits in Metric Spaces. In *Proc. 39th ACM Symp. on Theory of Computing*. 681–690.

Ilan Kremer, Yishay Mansour, and Motty Perry. 2013. Implementing the "Wisdom of the Crowd". In *Proc. 15th ACM Conf. on Electronic Commerce*. 605–606.

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

Chris J. Lintott, Kevin Schawinski, Ane Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. 2008. Galaxy Zoo: morphologies derived from visual inspection of ga laxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (Sept. 2008), 1179–1189.

Aditya Mahajan and Demosthenos Teneketzis. 2007. Multi-armed Bandit Problems. In *Foundations and Applications of Sensor Management*, D. Cochran A. O. Hero III, D. A. Castanon and K. Kastella (Eds.). Springer-Verlag.

Hebert E. Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58 (1952), 527–535.

Daniel Sheldon, M. A. Saleh Elmohamed, and Dexter Kozen. 2007. Collective Inference on Markov Models for Modeling Bird Migration. In *NIPS*.

Adish Singla and Andreas Krause. 2013. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In *22nd Intl. World Wide Web Conference*. 1167–1178.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. 2012. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory* 58, 5 (2012), 3250–3265.

Pravin P. Varaiya, Jean C. Walrand, and Cagatay Buyukkoc. 1985. Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Automat. Control* 30, 5 (1985), 426–439.

Peter Whittle. 1980. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* 42, 2 (1980), 143–149.

Yexiang Xue, Bistra N. Dilkina, Theodoros Damoulas, Daniel Fink, Carla P. Gomes, and Steve Kelling. 2013. Improving Your Chances: Boosting Citizen Science Discovery. In *HCOMP*.