Networked Restless Multi-Armed Bandits for Mobile Interventions

Han-Ching Ou[1]*, Christoph Siebenbrunner[1]*, Jackson Killian[1], Meredith B Brooks[1], David Kempe[2], Yevgeniy Vorobeychik[3], Milind Tambe[1]
[1]Harvard University, [2]University of Southern California, [3]Washington University in St. Louis
{hou@g.,csiebenbrunner@seas.,jkillian@g.,Meredith_Brooks@hms.}harvard.edu,
dkempe@usc.edu,yvorobeychik@wustl.edu,milind_tambe@harvard.edu

ABSTRACT

Motivated by a broad class of mobile intervention problems, we propose and study restless multi-armed bandits (RMABs) with network effects. In our model, arms are partially recharging and connected through a graph, so that pulling one arm also improves the state of neighboring arms, significantly extending the previously studied setting of fully recharging bandits with no network effects. In mobile interventions, network effects may arise due to regular population movements (such as commuting between home and work). We show that network effects in RMABs induce strong reward coupling that is not accounted for by existing solution methods. We propose a new solution approach for networked RMABs, exploiting concavity properties which arise under natural assumptions on the structure of intervention effects. We provide sufficient conditions for optimality of our approach in idealized settings and demonstrate that it empirically outperforms state-of-the art baselines in three mobile intervention domains using real-world graphs.

KEYWORDS

Restless Bandits, Commuting Networks, Scheduling

ACM Reference Format:
Han-Ching Ou[1]*, Christoph Siebenbrunner[1]*, Jackson Killian[1], Meredith B Brooks[1], David Kempe[2], Yevgeniy Vorobeychik[3], Milind Tambe[1], [1]Harvard University, [2]University of Southern California, [3]Washington University in St. Louis . 2022. Networked Restless Multi-Armed Bandits for Mobile Interventions. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

1 INTRODUCTION

Mobile interventions are a model for providing services in which agents are sent to different locations where they provide various forms of interventions locally. Of particular importance are mobile health clinics (MHCs), a model of healthcare delivery in which mobile units deliver health services directly to target communities. MHCs are successful in reaching vulnerable populations; they overcome typical barriers to health services access, such as limited transportation, finances, insurance, or legal status [37]. A wide variety of MHC services—such as primary care, prevention screenings, disease management, and treatment support—have been very successful. Their success is based on their flexibility in meeting the changing needs of target communities, and providing these services at discounted rates or free of charge. Compared to other healthcare service models, MHCs have been observed to provide cost savings and cost-effectiveness [37]. Another important application of mobile interventions is in food pantry services, which cater to communities experiencing food insecurity by dispatching food trucks.

Restless multi-armed bandits (RMABs) have become a widely adopted mathematical model for studying various types of intervention services [11, 16, 25, 27, 29, 31, 44]. RMABs are a model for sequential planning problems: in each round, a planner has to select k out of m arms to pull. Arms transition randomly between states, but the transition probabilities differ based on whether an arm was pulled or not. The arms dispense rewards depending on their state. In our motivating applications, arms represent locations, k may represent the budget (e.g., number of available MHC units), and rewards are the number of people positively affected by an intervention. In this paper, we extend existing RMAB models for interventions by considering network effects. Such network effects often arise due to individual commuting behavior: when an MHC visits one location, it provides interventions not only to people who reside there, but also to others who have traveled to this location (e.g., as a part of their routine work-related commuting). On the flip side, the same MHC may *miss* people who have traveled to a different location. Visiting one location may thus deliver an intervention to residents of multiple locations, giving rise to network effects. To the best of our knowledge, we are the first to consider RMAB models with network effects.

Network effects lead to significant new challenges in the formal model. Common solution approaches for RMABs treat each arm as a Markov Decision Process (MDP) and exploit the fact that these MDPs are coupled only through the joint budget constraint. This weak coupling forms the basis for solutions based on index values, which are computed separately for each of the m arms. Policies that select the k arms with the highest indices can be shown to be asymptotically optimal for several domains [21, 23, 28]. We show that the aforementioned network effects induce a stronger coupling between arms, making these solution approaches significantly less effective. The main contributions of our work are (1) we present a class of RMAB models with network effects suitable for modeling mobile intervention domains, (2) we present a solution approach for this class of problems and provide sufficient conditions for the optimality of our approach, and (3) we show empirically that our solution delivers superior performance compared to existing approaches across multiple domains.

*The first two authors have equal contributions.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

2 RELATED WORK

In the most general setting, the RMAB problem is known to be PSPACE-hard to solve optimally [34]. However, by exploiting the problem structure of certain restricted classes of RMABs, efficient algorithms have been derived, sometimes with performance guarantees. The most popular of these is the Whittle index policy [42] which is asymptotically optimal for *indexable* bandits [41] and fast to compute if a closed form can be derived for the index. Many works are dedicated to proving the indexability of different RMAB subclasses and deriving closed-form or efficient approximations of the Whittle index [4, 19, 22, 31]. Others have provided sufficient conditions for indexability [32] or developed expensive methods for computing policies with tighter reward bounds [3, 10]. However, all of these methods rely on the idea that the only factor coupling the arms are one or more budget constraints which we refer to as the *weakly coupled property*. Thus, previous RMAB methods will not be applicable for our work as the network effect strongly couples the states, actions, transitions, and rewards of neighboring arms.

In terms of applications, RMAB models have been widely used for scheduling problems, such as machine maintenance and repair [2, 19, 40]. In these works, machines in factories are modeled as arms, and the goal is to find the optimal schedule to visit factories to maintain the machines. Other examples include anti-poaching patrol planning ([35] propose a RMAB framework in which arms are poaching targets, and playing an arm corresponds to a patrol) or recommendation systems (e.g., for music streaming [45, 46]). Such problems also motivated the recharging bandit model [24]. In this model, each arm's reward is determined by a function of the time elapsed since the arm was last pulled. Implicitly, this resets the arm's reward to time 0 whenever the arm is pulled. When these functions are increasing and concave for each arm, [24] develop a concave program to solve the optimal frequency of pulling each arm; the program's value upper-bounds the value of an optimal schedule. Scheduling the arm then becomes a pinwheel scheduling problem [20], and [24] use a rounding scheme to approximate the scheduling of arm pulls, while obeying the frequency restriction. We extend this setting by allowing the arms' rewards to be only *partially* reset when the arm is selected, as well as by considering network effects.

In the public health domain, this paper's focus, [31] proposed collapsing bandits to improve medication adherence through interventions on patients. [27] and [8] proposed RMABs for scheduling cancer screenings and hepatitis treatments, respectively. In [16], the closest RMAB application to ours, the authors model the resource allocation problem of delivering school-based asthma care for children. The most important difference between our work and theirs is that we consider network effects in the RMAB model.

Related Work in Network Planning Sequential resource allocation problems on networks constitute another active area of research. Previous works have considered the non-restless setting, in which arms remain static when they are not pulled, such as influence maximization [15, 38], or have studied the network effect on state transitions [17, 33] instead of on actions. To the best of our knowledge, ours is the first work to study RMABs with interventions that have network effects.

3 PROBLEM FORMULATION

General RMABs. RMABs are a generalization of the well-studied multi-armed bandit model with many real-world applications. There are m arms $V = \{1, 2, \ldots, m\}$; each arm $v \in V$ can be in one of several states $s_{v,t} \in \mathcal{S}$ at any time step $t \in \mathbb{N}$. At any time step, the decision maker can pull up to k arms. Each chosen arm v transitions in a Markovian fashion according to a transition matrix $\mathbf{P}^a$ and yields a reward $r_v(s_{v,t}) \geq 0$ that depends only on the state of the arm v at time t. In the restless setting, arms that are not chosen also transition, according to a different matrix $\mathbf{P}^p$. The elements $p^a_{s,s'}$ ($p^p_{s,s'}$) of the transition matrix capture the probability of transitioning from state s to s' when the arm is played (not played). Let $V_{a,t}$ denote the set of arms being played at time step t. The total reward of time step t can be expressed as $R_t = \sum_{v \in V_{a,t}} r_{v,t}(s_{v,t})$. Each arm can be described as a two-action Markov Decision Process (MDP) $(\mathcal{S}, \{0, 1\}, \mathcal{R}, \mathcal{P})$. An action of 1 denotes that the arm is played and 0 that the arm is not played. Given the m MDPs and their initial states, the goal of this work is to find a policy for playing a sequence of k arms per round to maximize the average reward $\bar{R} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} R_t$.[1]

Networked RMABs for mobile interventions. We consider a setting where each arm v corresponds to a location which has a population $n_v \in \mathbb{N}$. The state $s_v \in \mathcal{S} = \{0, \ldots, n_v\}$ of a location is the number of healthy individuals. Individuals can either be in a healthy or, more generally, "good" state G or in a "bad" state B. Pulling an arm means visiting a location with a mobile intervention service, thereby exposing individuals at the location to the intervention. We thus consider the transition matrices for individuals, depending on whether they receive an intervention ($\mathbf{P}^a_v$) or not ($\mathbf{P}^p_v$):

$$\mathbf{P}^a_v = \begin{array}{c} \\ G \\ B \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{bmatrix} 1 - p^a_{v,GB} & p^a_{v,GB} \\ p^a_{v,BG} & 1 - p^a_{v,BG} \end{bmatrix} \end{array}, \mathbf{P}^p_v = \begin{array}{c} \\ G \\ B \end{array} \begin{array}{c} \begin{array}{cc} G & B \end{array} \\ \begin{bmatrix} 1 - p^p_{v,GB} & p^p_{v,GB} \\ p^p_{v,BG} & 1 - p^p_{v,BG} \end{bmatrix} \end{array}. \tag{1}$$

The transition probabilities are the same for all individuals with the same home location. Below, we will consider travel by individuals, which may result in them being exposed to the intervention at a different location. We stress that even in that case, an individual with home location v will transition according to the matrix P_v. This is because the characteristics of one's neighborhood are an important factor for one's health [36], keeping in mind the intended application domains of the model. We assume that the transition probabilities and the initial states are known, but the transitions are not observed. This is because while population-level health data can be monitored, this rarely happens in real time. We omit subscripts when they are clear from the context.

In order to account for network effects from commuting (or more general travelling) behavior, we define a probability distribution for individuals over locations. Let $w_{u,v} \in [0, 1]$ denote the probability that an individual with home location v is actually present in location u at any given moment (or that an individual from location v receives the intervention if location u is visited; we assume that individuals are sampled uniformly). Individuals can only be

[1] Another frequently considered reward criterion is the discounted reward $\sum_{t=0}^{\infty} \beta^t R_t$ with $0 \leq \beta < 1$.

in one location at any given time, implying that $\sum_{u \in V} w_{u,v} = 1$. The matrix $\mathbf{W} \in [0,1]^{m \times m}$ with elements $w_{u,v}$ is the weighted adjacency matrix of the travelling network. Introducing the travelling network has two effects:

(1) Not all individuals from location v are exposed to an intervention that visits v. In expectation, only $n_v w_{v,v}$ individuals from location v will receive the intervention (transition according to P_v^a) due to a visit at location v. This property is an important extension of the recharging bandits model [24]; in that model, it is assumed that each intervention fully "resets" the arm, i.e., puts all individuals into the good state.
(2) Individuals from other locations receive the intervention when v is visited. In expectation, $\sum_{u \in V \setminus \{v\}} n_u w_{v,u}$ individuals from other locations receive the intervention at v.

The total number of individuals reached in any location thus depends on whether other locations are visited, and we define the vector $\mathbf{a}_t \in \{0,1\}^m$, with at most k elements equal to 1, to represent all actions taken in round t. The vector of expected fractions of the populations at each location v reached by an action vector $\mathbf{a}$ is given by $\hat{\mathbf{w}}(\mathbf{a}) = \mathbf{W} \cdot \mathbf{a}$. Letting $\hat{w}_v$ denote the v-th entry of $\hat{\mathbf{w}}$, we also define the weighted average transition probabilities for a location v as $\hat{\mathbf{P}}_v(\mathbf{a}) = \hat{w}_v(\mathbf{a}_t) \cdot \mathbf{P}_v^a + (1 - \hat{w}_v(\mathbf{a}_t)) \cdot \mathbf{P}_v^p$. Further let $\mathbf{s}_{v,t} = [s_{v,t}, n_v - s_{v,t}]$ be the total number of individuals in the good and bad state in location v at time t. By conditioning on the current state $\mathbf{s}_{v,t}$ and actions, we are able to obtain a closed form expression for the expected state in the next time step:

$$\begin{aligned}\mathbb{E}(\mathbf{s}_{v,t+1} \mid \mathbf{s}_{v,t}, \mathbf{a}_t, \ldots, \mathbf{a}_0) =& \mathbb{E}(\mathbf{s}_{v,t+1} \mid \mathbf{s}_{v,t}, \mathbf{a}_t) \\ =& \hat{w}_v \mathbf{s}_{v,t} \mathbf{P}_v^a + (1 - \hat{w}_v) \mathbf{s}_{v,t} \mathbf{P}_v^p \\ =& \mathbf{s}_{v,t} \hat{\mathbf{P}}_v(\mathbf{a}_t).\end{aligned}$$

However, the current state is unknown according to our assumptions. Hence we seek an expression for the expected future state that does not require knowledge of the current state. Consider the expected state at time t conditional only on the action history: $\mathbb{E}_t(\mathbf{s}_{v,t}) := \mathbb{E}(\mathbf{s}_{v,t} \mid \mathbf{a}_{t-1}, \ldots, \mathbf{a}_0)$. Using the law of total expectation, we obtain

$$\begin{aligned}\mathbb{E}_{t+1}(\mathbf{s}_{v,t+1}) =& \mathbb{E}(\mathbf{s}_{v,t+1} \mid \mathbf{a}_t, \ldots, \mathbf{a}_0) \\ =& \mathbb{E}(\mathbb{E}(\mathbf{s}_{v,t+1} \mid \mathbf{s}_{v,t}, \mathbf{a}_t) \mid \mathbf{a}_t, \ldots, \mathbf{a}_0) \\ =& \mathbb{E}(\mathbf{s}_{v,t} \hat{\mathbf{P}}_v(\mathbf{a}_t) \mid \mathbf{a}_t, \ldots, \mathbf{a}_0) \\ =& \mathbb{E}(\mathbf{s}_{v,t} \mid \mathbf{a}_{t-1}, \ldots, \mathbf{a}_0) \hat{\mathbf{P}}_v(\mathbf{a}_t),\end{aligned}$$

since $\mathbf{s}_{v,t}$ does not depend on $\mathbf{a}_t$ (only on previous actions). We thus obtain a recurrence relation for the expected state:

$$\mathbb{E}_{t+1}(\mathbf{s}_{v,t+1}) = \mathbb{E}_t(\mathbf{s}_{v,t}) \hat{\mathbf{P}}_v(\mathbf{a}_t). \tag{2}$$

Eq. (2) allows us to compute the future expected state using only the current expectation and action vector. In order to fully describe the probability distribution of a single district, one would need $\binom{m}{k}$ matrices of size $(n_v + 1) \times (n_v + 1)$. Eq. (2) allows us to substantially reduce the complexity of the problem by focusing on the expected state. We write $\mathbb{E}_t(\mathbf{s}_{v,t}) = \mathbf{b}_{v,t}$ and use the recursion $\mathbf{b}_{v,t+1} = \mathbf{b}_{v,t} \hat{\mathbf{P}}_v(\mathbf{a}_t)$, where the initial state $\mathbf{b}_{v,0} = \mathbf{s}_{v,0}$ is known according to our assumptions.

The goal of the planner is to maximize the intervention benefit, taken as the sum of curing effects ($\text{cure}_v = p_{v,BG}^a - p_{v,BG}^p$) and prevention effects ($\text{prevention}_v = p_{v,GB}^p - p_{v,GB}^a$) for those individuals who received the intervention ($\text{cure}_v \hat{w}_v b_{v,t,2} + \text{prevention}_v \hat{w}_v b_{v,t,1}$, where $b_{v,t,1}$ and $b_{v,t,2}$ are the first and second element of $\mathbf{b}_{v,t}$, which are the expected total number of individuals in the good and bad state, respectively.), summed over locations and averaged over time steps. This criterion is chosen to align with the goals of applications such as MHCs which are to maximize the reach of a campaign [7], and to avoid underserving communities with a high probability of returning to the bad state, as could happen if only the total number of people in the good state were considered. Combining the curing and prevention effects, the reward per time step is given by: $R_t(\mathbf{a}_t) = \sum_{v \in V} \hat{w}_v(\mathbf{a}_t) \mathbf{s}_{v,t} (\mathbf{P}_v^a - \mathbf{P}_v^p) \cdot [1,0]^\top$. As discussed above, we focus on the expected reward and obtain:

$$\hat{R}_t := \mathbb{E}_t(R_t(\mathbf{a}_t)) = \sum_{v \in V} \hat{w}_v(\mathbf{a}_t) \mathbf{b}_{v,t} (\mathbf{P}_v^a - \mathbf{P}_v^p) \cdot [1,0]^\top. \tag{3}$$

We further make three assumptions that are natural in many relevant application domains; we combine assumptions made in prior work [31] (assumptions (1) and (2)) with input from health experts (assumption (3)).

(1) **The intervention is never bad for the individuals**: Health care interventions can help prevent disease or diagnose it early, reduce risk factors, and manage complications. Providing opportunities for increased access to quality services and interventions can reduce health disparities as well. Interventions provided via MHCs rarely result in negative impacts toward populations with little or no access to screening opportunities.
(2) **The individuals are more likely to stay in the good state than to change from the bad state to good**: In most applications, moving to the good state (curing of a disease or access to food) is unlikely to happen spontaneously.
(3) **The curing effect of the intervention is larger then the prevention effect**: MHCs mostly serve otherwise underserved communities. Those who attend MHCs are typically concerned about their health and may already be exhibiting symptoms of underlying disease. This makes curing interventions generally more useful/desired than preventive measures. In food pantry applications, the prevention effect is typically small.

These assumptions are formalized in Eq. (4), for all $v \in V$:

$$p_{v,GB}^p \geq p_{v,GB}^a \text{ and } p_{v,BG}^a \geq p_{v,BG}^p \tag{4a}$$

$$1 - p_{v,GB}^p > p_{v,BG}^p \text{ and } 1 - p_{v,GB}^a > p_{v,BG}^a \tag{4b}$$

$$p_{v,BG}^a - p_{v,BG}^p > p_{v,GB}^p - p_{v,GB}^a \tag{4c}$$

Next, we show that these assumptions entail two properties that will prove useful later in constructing effective algorithms for the networked RMAB problem. Specifically, consider a district v, and suppose that there are no interventions in adjacent districts. We can then define the reward gain of visiting v after τ_v time steps as $H_v^{\text{upper}}(\tau_v, \hat{w}_v) = (p_{v,GB}^p - p_{v,GB}^a) \hat{w}_v \hat{s}_{v,\tau_v} + (p_{v,BG}^a - p_{v,BG}^p) \hat{w}_v (n_{v,\tau_v}^a -$

$\hat{s}_{v,\tau_v}$) where $\hat{s}_{v,\tau_v}$ is the number of individuals in the good state at the time when the arm pull happens. This function has the following properties:

Theorem 1. *Under the assumptions in Eq.* (4)*, and assuming no interventions in neighboring districts, H_v^u is a monotone increasing concave function with respect to time τ_v elapsed since the last pull.*

Theorem 2. *Under the assumptions in Eq.* (4)*, and assuming no interventions in neighboring districts, H_v^u is a monotone increasing concave function with respect to the expected population share $\hat{w}_v$ exposed to the intervention.*

The proofs are deferred to the supplementary material. Theorem 1 tells us that adding an extra pull to the intervention schedule of an arm will always improve the reward. From Theorem 2, we know that it is always preferable to intervene on a larger proportion of the population of an arm. These results suggest that the periodic policy is still a reasonable choice under the networked setting. The periodic policy in the non-networked setting is motivated by the following consideration: suppose that instead of pulling *exactly k* arms, we require only that *on average*, k arms are pulled in each round. In this relaxed problem, a periodic policy with suitable periods is optimal if the reward function is concave [24].[2] Theorems 1 and 2 tell us that the reward function for the networked problem is still concave.

4 SOLUTION APPROACHES

As discussed previously, our problem shares significant similarities with the recharging bandits problem [24]. Both in the network-free and networked setting, a natural solution approach is to (1) determine the frequencies with which arms should be pulled, and then (2) sequence the pulls optimally. Importantly, the network effects affect both stages of the solution approach. As a result, simple optimal (or near-optimal) policies from the non-networked setting may be far from optimal when networks are considered.

The fact that network effects must be taken into account in determining arm pull frequencies is easy to see. Consider a star graph in which the central node has population 0, while the $m-1$ leaf nodes have population $n_v = n$, and — importantly — have probability 1 of commuting to the central node. Without considering the network/commuting effect, any policy would choose a non-central node in each round (because the central node has population 0), whereas picking the central node in each round is clearly optimal.

Perhaps more interestingly, network effects also impact which sets of arms should be pulled simultaneously, even keeping the arm pull frequencies constant (and having identical arms). This is illustrated in the following example.

Example 1. *Consider the example shown in Fig. 1. We set $k = 2$ and $(p^p_{GB}, p^p_{BG}, p^a_{GB}, p^a_{BG}) = (p_{GB}, 0, p_{GB}, 1)$. All arms in Fig. 1 are identical. The optimal periodic policy is to select each arm every two rounds [24]. Such a policy can be achieved without any rounding by selecting exactly two arms in each round. However, different ways of choosing these two arms result in policies with different rewards. Specifically, we consider the following two policies: Policy NN: Select two non-neighboring locations in each round. Policy NB: select two neighboring locations in each round. We also consider two different network scenarios with different commuting probabilities. In scenario 1, $w_{u,v} = \frac{1}{2}$ for all $(u,v) \in E$ and $w_{v,v} = 0$ for all $v \in V$, i.e., all individuals commute to adjacent nodes. In scenario 2, $w_{u,v} = \frac{1}{4}$ for all $(u,v) \in E$ and $w_{v,v} = \frac{1}{2}$ for all $v \in V$, i.e., half of the individuals stay put. Table 1 summarizes the rewards of the two policies in the two scenarios: In scenario 1, the policy NN is the better policy for any p_{GB}, and the relative reward difference can be as large as $\frac{2}{3}$. In scenario 2, the policy NB becomes the better policy. For large p_{GB}, the relative reward difference approaches $\frac{13}{15}$. In particular, we see that the network effects must be taken into account in order to find the optimal way to coordinate the arm pulls of different arms.*

	Scenario 1	*Scenario 2*
Policy NN	$\frac{4p_{GB}-2p_{GB}^2}{1+p_{GB}-p_{GB}^2} \to 2$	$\frac{4p_{GB}}{2p_{GB}+1} \to \frac{4}{3}$
Policy NB	$\frac{4p_{GB}}{2p_{GB}+1} \to \frac{4}{3}$	$\frac{52p_{GB}-32p_{GB}^2}{13+16p_{GB}-16p_{GB}^2} \to \frac{20}{13}$

Table 1: Rewards of the two policies, and limits as $p_{GB} \to 1$, in the two scenarios.

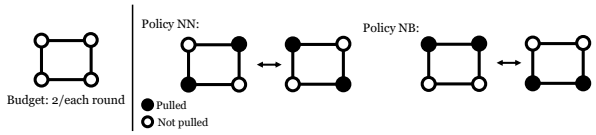


Figure 1: Example for how network combinatorial effects affect the reward of periodic policies.

Our proposed solution consists of two parts. In Section 4.1.1, we present an approach to obtain the optimal visiting period for each district. In Section 4.1.2, we illustrate our approach for synchronizing the arm pulls to optimize reward coupling.

4.1 Proposed approach

Despite the added model complexities compared to the non-networked Recharging Bandits model, our problem preserves similar concavity properties. In a similar vein as [24], we thus aim to provide periodic policies for the networked RMAB problem, i.e., policies that repeat after T time steps. This not only facilitates scheduling, but can also reinforce intervention benefits in MHC domains [43]. Exhaustively searching the action space of size $\binom{m}{k}^T$ is clearly impractical for reasonable problem sizes m. Fortunately, we can reduce the search space by exploiting the concavity we proved in Theorem 1.

4.1.1 Obtaining Visiting Periods. Let x_v be the fraction of times that arm v is chosen. When $1/x_v$ is integral, it can easily be shown that pulling the arm every $1/x_v$ rounds will maximize reward due to the concavity of the reward function [24]. Define the period of pulling $\tau_v = 1/x_v \in \{1, 2, 3, \ldots, T\}$, meaning that v is visited every τ_v time steps. Let T be the maximum period considered, which could be a month, a season, or a year, depending on the application. Our goal is to find the optimal time period for each arm, subject

[2]The constrained version is then a more difficult problem that involves solving a pinwheel problem, which is NP-hard.

to the sum of intervention frequencies being at most the budget $\sum_{v \in V} x_v \leq k$.

Suppose that a policy pulls arm v every τ_v time steps and follows some schedule $\pi : t \to \mathbf{a}_t$. We define $\mathbf{P}_v^*(\tau_v, \pi) = \prod_{t=0}^{\tau_v} \hat{\mathbf{P}}_v(\pi(t))$ as the transition matrix of the expected state vector right before the next arm pull. Note that the reward gained from pulling an arm v will depend on whether neighboring arms have recently been pulled, as this would imply that some share of v's population has already been exposed to the intervention. For a given τ_v, the reward gained from pulling v is minimized when all neighboring arms are visited in every round and maximized when no locations other than v are visited. We denote these two policies by π^ℓ and π^u, respectively. We can thus bound the average reward gained from pulling arm v every τ_v rounds (defined as $H_v(\tau_v)$) as:

$$\frac{1}{\tau_v} \bar{\mathbf{b}}_v^\ell \mathbf{P}_v^*(\tau_v, \pi^\ell) \mathbf{n}_{v,G} \leq H_v(\tau_v) \leq \frac{1}{\tau_v} \bar{\mathbf{b}}_v^u \mathbf{P}_v^*(\tau_v, \pi^u) \mathbf{n}_{v,G},$$

where $\bar{\mathbf{b}}_v^\ell$ ($\bar{\mathbf{b}}_v^u$) is the steady state of $\mathbf{P}_v^*(\tau_v, \pi^\ell)$ ($\mathbf{P}_v^*(\tau_v, \pi^u)$), which is also its eigenvector corresponding to its smallest eigenvalue. $\mathbf{P}_v^*(\tau_v, \pi)$ is the τ_v-step transition matrix of arm v given the policy of other arms π.

Given the upper bound $H_v^{\text{upper}}(\tau_v) = \frac{1}{\tau_v} \bar{\mathbf{b}}_\mathbf{v}^\mathbf{u} \mathbf{P}_v^*(\tau_v, \pi^u) \mathbf{n}_{v,G}$, we can construct the reward table for each arm v by calculating the upper bound of each possible τ_v. Finding the optimal period for each arm thus becomes an optimization problem

$$\max \sum_{v \in V} H_v^{\text{upper}}(\tau_v) \qquad \text{s.t.} \sum_{v \in V} x_v \leq k.$$

We explicitly write the optimization problem as a MILP with integer variables $x_{v,t} \in \{0, 1\}$ for all $v \in V, t \in \{1, 2, \ldots, T\}$. $x_{v,t} = 1$ denotes that location v has a period of t. In the MILP, we write $H_v^u(t) := \frac{1}{t} \bar{\mathbf{b}}_\mathbf{v}^\mathbf{u} \mathbf{P}_v^*(t, \pi^u) \mathbf{n}_{v,G}$ for all v and t.

$$\begin{array}{lll} \text{Maximize} & R & \\ \text{subject to} & \sum_v \sum_{t=1}^T \frac{x_{v,t}}{t} \leq k & \text{(budget)} \\ & \sum_{t=1}^T x_{v,t} \leq 1 \quad \text{for all } v & \text{(periods)} \\ & R \leq \sum_{v \in V} \sum_{t=1}^T x_{v,t} H_v^u(t) & \text{(reward)} \\ & x_{v,t} \in \{0, 1\} \quad \text{for all } v, t. & \end{array} \tag{5}$$

The MILP (5) has $O(|V|T)$ constraints. Its implementation can be found in the source code provided. The first constraint captures that the chosen periods/frequencies allow a fractional solution of at most k visits per time step. The second set of constraints captures that each location has only one period. The third constraint bounds the reward. From the MILP solution, for each v, the period τ_v can be obtained as the (at most one) t such that $x_{v,t} = 1$. If $x_{v,t} = 0$ for all t for a particular v, then the arm is never worth pulling and can be discarded from the candidate pool.

The MILP can be adjusted to take fairness considerations into account as well. We list a few examples here; further details are discussed in the appendix:

- To achieve a minimum visiting frequency of $f_{\min}$, we can replace T with $T_{\min} = 1/f_{\min}$.
- To ensure that individuals from each node v have sufficient access to the intervention (either at v or a neighboring node), we can add the constraints $\sum_{u \in V} \sum_{t=0}^T \frac{w_{u,v} x_{u,t}}{t} \geq L$ for all v.
- To encourage the algorithm to increase the smallest node rewards, we can replace the reward with the alternative welfare function $R \leq \sum_{v \in V} \sum_{t=1}^T x_{v,t} (\frac{H_v^u(t)}{n_v})^\alpha / \alpha$ for $\alpha \leq 1$.

4.1.2 Finding optimal node sets to account for reward coupling. As illustrated in Example 1, the combinatorial effects of pulling arms in the networked RMAB problem induce reward coupling between the MDPs of the arms. In contrast to non-networked recharging bandits, the choice of which set of arms with equal optimal periods to pull in the same rounds thus matters in networked bandits. The potential loss in reward here stems from the fact that when two arms that are both neighboring arms of a third arm are intervened on in different time steps, they will deliver the intervention in part to the same individuals in the third arm.

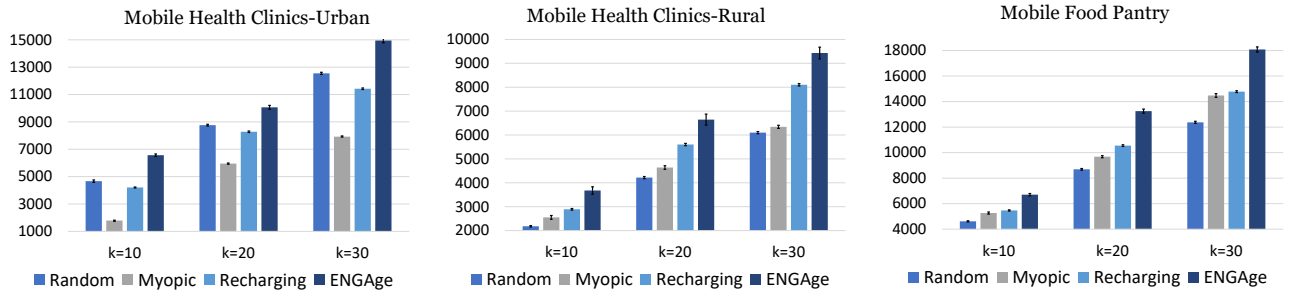
In any time step t, for any pair of arms that is pulled simultaneously, we seek to maximize the overlap between the shares of populations in the set of arms that are neighbors of both arms. For a pair of arms (v, v'), this intervention overlap can be computed as $\sum_{u \in \delta(v) \cap \delta(v')} w_{v,u} w_{v',u}$. If (and only if) the optimal periods τ_v and $\tau_{v'}$ are coprime to each other, this intervention overlap is independent of when the arms are intervened on. (As an example, two arms with periods 2 and 3 will be pulled together every six rounds, regardless of when the policy starts pulling each arm.) If the periods τ_u and τ_v have a common factor, on the other hand, they can never be pulled together if they are out of sync. (Arms with periods 2 and 4 will never be pulled together if their sequences start one time step apart.) We would thus be losing out on the reward gains from pulling the arms together every $\text{lcm}(\tau_u, \tau_v)$ rounds. In order to minimize this loss, we construct an undirected graph $\bar{G}(V, \bar{E})$ with the following edge weights:

$$\bar{w}_{v,v'}(\tau_v, \tau_{v'}) = \begin{cases} \sum_{u \in \delta(v) \cap \delta(v')} \frac{w_{v,u} w_{v',u}}{\text{lcm}(\tau_v, \tau_{v'})} & \text{if } \gcd(\tau_v, \tau_{v'}) > 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The weight of the cut between the selected and unselected arms on $\bar{G}$ equals the average reward loss due to the intervention overlap. We can thus select the arm set to pull by minimizing the cut between the selected node set (of size k) and the unselected node set. Graph partition problems with node cardinality constraints are generally NP-hard [39]. We use a heuristic based on spectral graph partitioning, by considering the k nodes with the largest or smallest value in the eigenvector corresponding to the second-smallest eigenvalue of $\bar{L}$ (also known as the Fiedler vector), where $\bar{L}$ denotes the Laplacian of the graph $\bar{G}$. The ENGAGE (Efficient Network Geography Aware scheduling) Algorithm (Algorithm 1) outputs an intervention policy based on this approach.

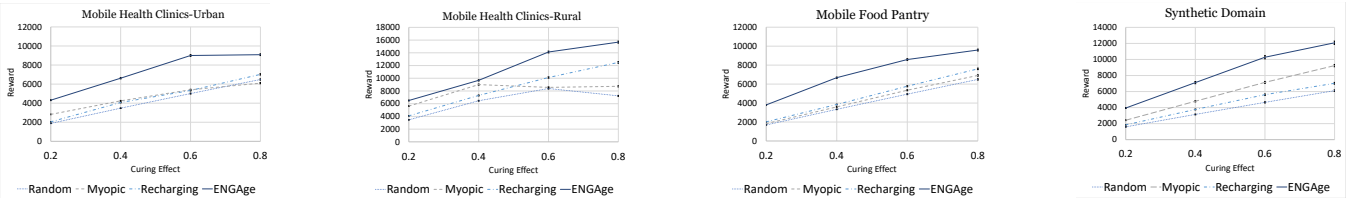
4.2 Analysis

We start by analyzing the complexity of the solution approach described above. The concave MILP (5) can be solved efficiently using time $O(|V|T \log(|V|T))$, by sorting the set of slopes of segments, corresponding to the different $H_v^u(t)$. Details are given in [24]. In our implementation, we instead use an off-the-shelf MILP solver. While its worst-case running time is larger, as our experiments show, it runs very efficiently in practice. Calculating the Laplacian $\bar{L}$ requires finding common neighbours ($O(\hat{d}|E|)$ by [6], where $\hat{d}$ is



(a) MHCs in urban area (b) MHCs in rural area (c) Mobile Food Pantry

Figure 2: Average reward in three different domains under different budget constraints.



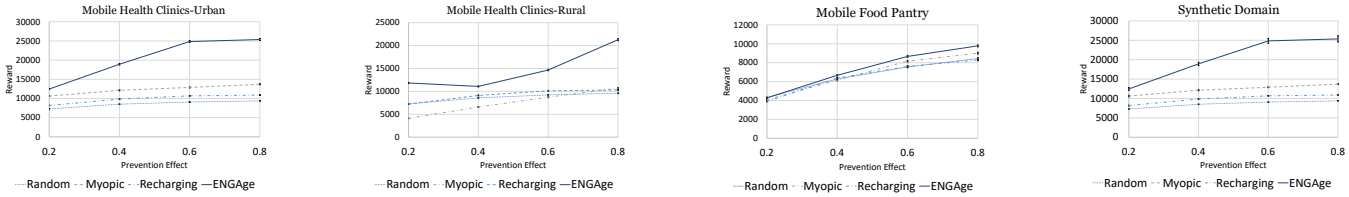
(a) Reward for different curing rates in urban MHC domain

(b) Reward for different curing rates in rural MHC domain

(c) Reward for different curing rates in food pantry domain

(d) Reward for different curing rates in synthetic domain

Figure 3: Average reward in three different domains under different curing effects ($P^a_{GB} - P^p_{GB}$) .



(a) Reward for prevention rates in urban MHC domain

(b) Reward for prevention rates in rural MHC domain

(c) Reward for prevention rates in food pantry domain

(d) Reward for prevention rates in synthetic domain

Figure 4: Average reward in three different domains under different prevention effects ($P^p_{BG} - P^a_{BG}$) .

the maximum degree in $\bar{G}$) and then computing their gcd ($O(\log T)$ using the Euclidean algorithm). The overall cost of computing the Laplacian is thus $O(\hat{d}|E| \log T)$. Again, our actual implementation is less efficient in terms of worst-case complexity, but runs fast in practice nonetheless. Finding a Fiedler vector takes time $O(\bar{d}|V|)$ using Lanczos' algorithm [26], where $\bar{d}$ is the average degree of $\bar{G}$. The rest of the planning takes time $O(|V|T)$. Thus, the total time complexity of our algorithm is $O(|V|T \log(|V|T) + \hat{d}|E| \log T)$.

In Section 5, we experimentally evaluate the performance of our algorithm on various graphs from real-world domains. We now turn to analyzing sufficient conditions that guarantee optimality for various cases that we will discuss below.

First consider the case of homogeneous nodes and edge weights, i.e., all nodes have the same populations and transition probabilities between states, and all edges have the same commute probabilities. If we replace the eigenvector-based heuristic in ENGAge with an oracle that optimally solves the min-cut problem with cardinality constraints, then ENGAge outputs the optimal policy for arbitrary graphs of N nodes whenever $k|N$. This is because in this case, the cut on the constructed graph measures the exact reward loss of the schedule. Solving the min-cut problem optimally will then lead to the optimal scheduling.

Next, consider the special case in which the graph $\bar{G}$ has γ connected components $C_1, \ldots, C_\gamma$, each of size $|C_i| = k$. Furthermore, we assume that all elements of the same component have the same optimal period; that is, if $u, v \in C_i$, then $\tau_u = \tau_v$. For $\gamma \geq 2$, note that $\bar{L}$ is positive semidefinite as $\bar{G}$ is undirected for arbitrary input graphs G by construction. The smallest eigenvalue 0 will have multiplicity γ in the Laplacian $\bar{L}$. Thus, $|\Lambda| = \gamma$, and it is known that each component C_i has a corresponding Fiedler vector supported entirely on C_i [30]. Hence, in each iteration, Algorithm 1 will select exactly all members of one component. As there are

Algorithm 1 ENGAGE

$V_{\text{candidate}} \leftarrow V$ and $V_{\text{wait}} \leftarrow \emptyset$.
Compute periods τ_v using the MILP (5).
Construct the new graph $\bar{G}(V, \bar{E})$ according to Eq. (6) and compute its Laplacian $\bar{L}$.
Find the set Λ of Fiedler vectors of $\bar{L}$ (more than one in case of eigenvalue multiplicity).
for $t = 1, \ldots, T$ **do**
  for $v \in V_{\text{wait}}$ **do**
    $\text{Timer}(v) \leftarrow \text{Timer}(v) - 1$.
    if $\text{Timer}(v) = 0$ **then**
      Move v from V_{wait} to $V_{\text{candidate}}$.
  $V_a(t) \leftarrow \emptyset$.
  for all $\eta \in \Lambda$ **do**
    Find the sets of nodes with k-th largest and smallest elements in η: Specifically, let $\eta_{(k)}$ denote the k-th largest entry of η, set $\bar{V} \leftarrow \{v \in V_{\text{candidate}} \mid \eta_v \leq \eta_{(k)}\}$ and $\underline{V} = \{v \in V_{\text{candidate}} \mid \eta_v \geq \eta_{(m-k+1)}\}$.
    If $|\bar{V}| > k$ or $|\underline{V}| > k$, reduce the set size to k by arbitrarily removing tied nodes at the cutoff threshold.
    Update $V_a(t)$ to the set S that minimizes the cut: $V_a(t) \leftarrow \text{argmin}_{S \in \{\bar{V}, \underline{V}, V_a(t)\}} c(S)$. Here, $c(S)$ denotes the cut capacity of the node set S in $\bar{G}$ (and is defined as ∞ for the empty set). Arbitrarily break ties.
    Move $V_a(t)$ from $V_{\text{candidate}}$ to V_{wait}, and set $\text{Timer}(v) \leftarrow \tau_v$ for these arms.
return $V_a(t)$ as arms to pull at time t for all times $t = 1, \ldots, T$.

no links between nodes in different components by definition, all members of a component will be fully intervened on. Our problem thus reduces to a pinwheel problem with γ arms and optimal periods τ_i for $i = 1, \ldots, \gamma$. Pinwheel problems are known to be NP-hard in general [13], but optimal solutions are known to exist in special cases where all periods are multiples of one another and $\sum_i^\gamma \tau_i \leq k$ [20]. The optimal solution in these cases can be obtained by a simple greedy policy (see [13]) which is realized by the sets V_{wait} of our algorithm. The latter condition is guaranteed by the setup of ENGAGE; hence, our proposed approach will output an optimal schedule in those cases. For $\gamma = 1$, the same conclusion follows trivially, because the algorithm can visit all locations in each time step.

Based on the above analysis, ENGAGE will output the optimal policy in the following settings, among others: (1) Complete graphs with equal edge weights, identical nodes, and $k|N$. (2) Graphs with multiple connected components, each of size k, with equal edge weights and identical nodes. (3) Rings with edge weights $1/2$, identical nodes, and $k = N/2$. (4) d-dimensional Hypercubes with edge weights $1/d$, identical nodes, and $k = N/d$. (5) Bipartite or multipartite graphs with partitions of size k, identical node degrees, and edge weights summing to 1 for all nodes. (6) Strongly d-regular graphs with equal edge weights and identical nodes. These are illustrative examples of graphs where our algorithm is guaranteed to perform optimally. In the next section, we will empirically show that it outperforms existing methods in more general settings, including real-world graphs.

Table 2: Properties of the network data sets.

Network	$\lvert V\rvert$	average degree	average degree centrality
Boston	431	2.92	0.005
Daniels County	631	2.53	0.008
Los Angeles	561	2.85	0.001

5 EXPERIMENTAL EVALUATION

We perform experiments comparing our algorithm to baselines in a variety of real-world application scenarios. We begin by describing the application domains and their properties:

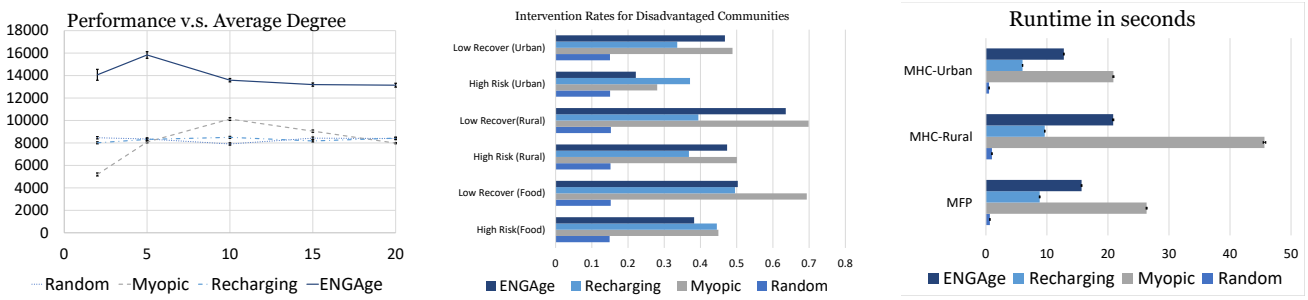
Mobile Health Clinics in urban areas: This domain setting is modeled on MHCs that are an important part of urban health care programs. Specifically, we consider a graph of the city of Boston (where such MHCs are used by non-profit organizations [14]), collected from [12]. The graph consists of 431 locations that are used as bandit arms. The populations n_v and transition probabilities $(p^p_{vGB}, p^p_{vBG}, p^a_{vBG}, p^a_{vGB})$ are generated from uniformly random distributions subject to the assumptions introduced in the problem formulation section[3].

Mobile Health Clinics in rural areas: In contrast to urban areas, rural areas are characterized by a larger number of less connected smaller communities, and may experience lower overall levels of access to health services. We model this domain using a graph of Daniels County, MT, with 631 locations, taken from [12]. Daniels County is considered one of the most rural counties in the US, as measured by the index of relative rurality [1]. We modify the previous setting to set a large portion of districts to have communities with relatively small population, to account for the characteristics described before.

Mobile Food Pantry: Due to a limited choice of means of transportation, residents of many socially disadvantaged neighborhoods can only access food within shorter distances; as a result, healthy food options are often limited. Mobile food pantries (MFPs) have become an important source of healthy food for these communities [5]. In the MFP scenario, the Los Angeles city graph with 561 locations collected from [12] is used, as food insecurity is an important issue in Los Angeles. In this scenario, it is assumed that there is no prevention effect ($p^a_{GB} = p^p_{GB}$), as the provided food needs to be fresh and will only be distributed to individuals in bad states.

We compare our algorithm to three baseline algorithms. RANDOM selects k locations uniformly at random in each time step. MYOPIC selects the locations with maximum reward in the current time step. RECHARGING is the rounding scheme scheduling provided in [24]. All experiments are conducted on a system with 6 cores, 2.60 GHz Intel CPU, and 16 GBs of RAM for 30 simulations over 100 time steps for each trial. All figures include approximate 95% confidence intervals as error bars. Figures 2a–2c show the average reward collected with different budgets of $k \in \{10, 20, 30\}$ arms,

[3]While we have access to real-world street graph data, we do not have access to population and commuting data at a matching granularity.



(a) Average Degree (b) Disadvantaged Communities (c) Runtime

Figure 5: (5a): average reward vs. graph average degree (5b): intervention rates for 15% most disadvantaged communities. (5c): average runtimes.

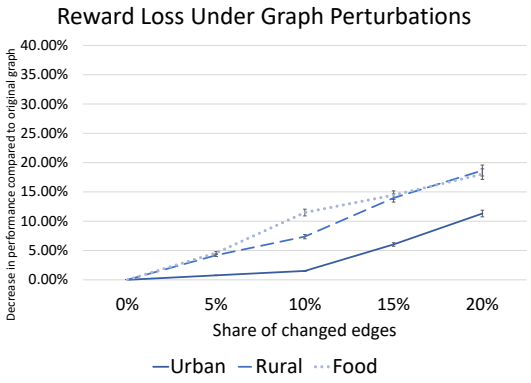


Figure 6: Sensitivity analysis.

for the three domains described above. Our algorithm consistently outperforms all baselines. RECHARGING mostly performs second-best, though in the urban MHC setting, it is slightly worse than RANDOM. Figure 5c shows the average runtime per simulation in seconds. Interestingly, MYOPIC is the slowest algorithm, because it has to compute the reward for each node in each round, while ENGAGE and RECHARGING use pre-computed period tables.

We further analyze the sensitivity of these results to several modeling parameters. Figure 5a shows the performance of the algorithms for different densities for a synthetic domain based on a spatial preferential attachment model [9, 18]. The results are non-monotonic for the ENGAGE algorithm. A possible explanation could be that there might exist a level of optimum connectivity, below which adding more links will increase the intervention benefit by spreading interventions more widely, and above which adding more links will cause too much overlap between the populations that are intervened on in different time steps. Figures 3 and 4 show that ENGAGE consistently outperforms the baselines across multiple values for cure and prevention rates in all domains.

We also analyze the impact of our algorithm on the most disadvantaged communities, i.e., those experiencing the highest risk of transitioning to the bad state, or which have small probability of recovering from the bad state. Figure 5b shows the average intervention frequencies for the 15% communities with the highest risk (p^p_{GB}) and lowest chance of recovery (p^p_{BG}). All algorithms except RANDOM intervene on the most disadvantaged communities disproportionately more often, showing that they are not discriminating against them. This is thanks to the design of the reward criterion that measures intervention benefit for individuals receiving the intervention.

Finally, we conduct a sensitivity analysis of the ENGAGE algorithm against graph perturbations. Figure 6 is constructed as follows: Starting with the real-world graphs from the three domains, we add perturbations by removing a given percentage of the edges, and adding back the same number of edges randomly. In the optimization, we then use the perturbed graph, while the original, unperturbed graph is used to compute the rewards. Overall, we observe that perturbing x% of edges generally reduces reward by less than x%. For example, with a graph perturbation of 15%, the performance reductions in the urban, rural and food settings are 6%, 13%, and 14%, respectively.

6 CONCLUSION

We present a networked RMAB model motivated by mobile interventions; our model captures network effects stemming from traveling behavior. Our model was built based on the input of domain experts in mobile health interventions. To the best of our knowledge, this is the first paper addressing the challenge of scheduling multiple interventions with network effects in the RMAB model. Network effects induce strong reward coupling between arms, substantially complicating the analysis of the RMAB. We propose the ENGAGE (Efficient Network Geography Aware scheduling) algorithm that takes reward coupling and network effects into account. We provide sufficient conditions for optimality and show that our algorithm outperforms several baselines empirically in three real-world domains and synthetic domains with varying properties.

ACKNOWLEDGMENTS

This work was supported by the Army Research Office (MURI W911NF1810208). J.A.K. was supported by an NSF Graduate Research Fellowship under Grant DGE1745303.

REFERENCES

[1] 2007. Measuring Rurality. http://www.incontext.indiana.edu/2007/january/2.asp. Accessed: 2021-05-24.

[2] Abderrahmane Abbou and Viliam Makis. 2019. Group maintenance: A restless bandits approach. *INFORMS Journal on Computing* 31, 4 (2019), 719–731.

[3] Daniel Adelman and Adam J Mersereau. 2008. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research* 56, 3 (2008), 712–727.

[4] Nima Akbarzadeh and Aditya Mahajan. 2019. Restless bandits with controlled restarts: Indexability and computation of Whittle index. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 7294–7300.

[5] Susan J Algert, Aditya Agrawal, and Douglas S Lewis. 2006. Disparities in access to fresh produce in low-income neighborhoods in Los Angeles. *American journal of preventive medicine* 30, 5 (2006), 365–370.

[6] Xiaojing An, Kasimir Gabert, James Fox, Oded Green, and David A Bader. 2019. Skip the Intersection: Quickly Counting Common Neighbors on Shared-Memory Systems. In *2019 IEEE HPEC*. IEEE, 1–7.

[7] John Auerbach. 2016. The 3 buckets of prevention. *Journal of public health management and practice: JPHMP* 22, 3 (2016), 215.

[8] Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research* 67, 3 (2019), 853–873.

[9] Marc Barthélemy. 2011. Spatial networks. *Physics Reports* 499, 1-3 (2011), 1–101.

[10] Dimitris Bertsimas and José Niño-Mora. 2000. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research* 48, 1 (2000), 80–90.

[11] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. 2021. Learn to Intervene: An Adaptive Learning Policy for Restless Bandits in Application to Preventive Healthcare. *2021 IJCAI* (2021).

[12] Geoff Boeing. 2017. U.S. Street Network. In *U.S. Street Network Shapefiles, Node Edge Lists, and GraphML Files*. Harvard Dataverse. https://doi.org/10.7910/DVN/CUWWYJ/JPXJTV

[13] Mee Yee Chan and Francis Chin. 1993. Schedulers for larger classes of pinwheel instances. *Algorithmica* 9, 5 (1993), 425–462.

[14] Haipeng Chen, Susobhan Ghosh, Gregory Fan, Nikhil Behari, Arpita Biswas, Mollie Williams, Nancy E Oriol, and Milind Tambe. 2022. Using Public Data to Predict Demand for Mobile Health Clinics. In *In 2022 IAAI*.

[15] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. 2016. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research* 17, 1 (2016), 1746–1778.

[16] Sarang Deo, Seyed Iravani, Tingting Jiang, Karen Smilowitz, and Stephen Samuelson. 2013. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research* 61, 6 (2013), 1277–1294.

[17] Mathilde Fekom, Nicolas Vayatis, and Argyris Kalogeratos. 2019. Sequential dynamic resource allocation for epidemic control. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 6338–6343.

[18] Luca Ferretti and Michele Cortelezzi. 2011. Preferential attachment in growing spatial networks. *Physical Review E* 84, 1 (2011), 016103.

[19] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride. 2006. Some indexable families of restless bandit problems. *Adv. Appl. Probab.* 38, 3 (2006), 643–672.

[20] Robert Holte, Aloysius Mok, Louis Rosier, Igor Tulchinsky, and Donald Varvel. 1989. The pinwheel: A real-time scheduling problem. In *Proceedings of the 22nd Hawaii International Conference of System Science*. 693–702.

[21] Junya Honda and Akimichi Takemura. 2010. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models.. In *COLT*. Citeseer, 67–79.

[22] Yu-Pin Hsu. 2018. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE ISIT*. IEEE, 2634–2638.

[23] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*. Springer, 199–213.

[24] Robert Kleinberg and Nicole Immorlica. 2018. Recharging bandits. In *2018 IEEE 59th FOCS*. IEEE, 309–319.

[25] U Dinesh Kumar and Haritha Saranga. 2010. Optimal selection of obsolescence mitigation strategies using a restless bandit model. *European Journal of Operational Research* 200, 1 (2010), 170–180.

[26] C. Lanczos. 1950. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Gov. Press Office Los Angeles, CA.

[27] Elliot Lee, Mariel S Lavieri, and Michael Volk. 2019. Optimal screening for hepatocellular carcinoma: A restless bandit model. *Manufacturing & Service Operations Management* 21, 1 (2019), 198–212.

[28] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. 2011. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*. JMLR Workshop and Conference Proceedings, 497–514.

[29] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 565–582.

[30] Anne Marsden. 2013. Eigenvalues of the laplacian and their relationship to the connectedness of a graph. *University of Chicago, REU* (2013).

[31] Aditya Mate, Jackson A Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. 2020. Collapsing Bandits and Their Application to Public Health Intervention.. In *2020 NeurIPS*.

[32] Jose Nino-Mora. 2001. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability* (2001), 76–98.

[33] Han-Ching Ou, Haipeng Chen, Shahin Jabbari, and Milind Tambe. 2021. Active Screening for Recurrent Diseases: A Reinforcement Learning Approach. In *In 2021 AAMAS*.

[34] Christos H Papadimitriou and John N Tsitsiklis. 1994. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*. IEEE, 318–322.

[35] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *In 2016 AAMAS*. 123–131.

[36] Catherine E Ross and John Mirowsky. 2001. Neighborhood disadvantage, disorder, and health. *Journal of health and social behavior* (2001), 258–276.

[37] WY Stephanie, Caterina Hill, Mariesa L Ricks, Jennifer Bennet, and Nancy E Oriol. 2017. The scope and impact of mobile health clinics in the United States: a literature review. *International journal for equity in health* 16, 1 (2017), 1–12.

[38] Sharan Vaswani, Laks Lakshmanan, Mark Schmidt, et al. 2015. Influence maximization with bandits. *arXiv preprint arXiv:1503.00024* (2015).

[39] Vijay V Vazirani. 2013. *Approximation algorithms*. Springer Science & Business Media.

[40] Hongzhou Wang. 2002. A survey of maintenance policies of deteriorating systems. *European journal of operational research* 139, 3 (2002), 469–489.

[41] R. R. Weber and G. Weiss. 1990. On an index policy for restless bandits. *J. Appl. Probab.* 27, 3 (1990), 637–648.

[42] Peter Whittle. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability* (1988), 287–298.

[43] Walter C Willett, Jeffrey P Koplan, Rachel Nugent, Courtenay Dusenbury, Pekka Puska, and Thomas A Gaziano. 2006. Prevention of chronic disease by means of diet and lifestyle changes. *Disease Control Priorities in Developing Countries. 2nd edition* (2006).

[44] Lily Xu, Elizabeth Bondi, Fei Fang, Andrew Perrault, Kai Wang, and Milind Tambe. 2021. Dual-Mandate Patrols: Multi-Armed Bandits for Green Security. In *In AAAI 2021*, Vol. 35. 14974–14982.

[45] Jinfeng Yi, Cho-Jui Hsieh, Kush Varshney, Lijun Zhang, and Yao Li. 2017. Scalable demand-aware recommendation. *arXiv preprint arXiv:1702.06347* (2017).

[46] Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. 2016. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD*. 2025–2034.